All of Us diversity and scale improve polygenic prediction contextually with greatest improvements for under-represented populations

Kristin Tsuo[1,2,3], Zhuozheng Shi[4], Tian Ge[2,5,6,7], Ravi Mandla[4], Kangcheng Hou[4], Yi Ding[4], Bogdan Pasaniuc[4,8,9,10,11], Ying Wang[1,2,3,†], Alicia R. Martin[1,2,3,†]

Affiliations

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.
[2] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[3] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[4]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[5] Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.
[6] Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA.
[7] Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA.
[8]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[9]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[10]Institute for Precision Health, University of California, Los Angeles, Los Angeles, CA 90095, USA.
[11]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA.

[†]These authors co-supervised the study.

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

Abstract

Recent studies have demonstrated that polygenic risk scores (PRS) trained on multi-ancestry data can improve prediction accuracy in groups historically underrepresented in genomic studies, but the availability of linked health and genetic data from large-scale diverse cohorts representative of a wide spectrum of human diversity remains limited. To address this need, the All of Us research program (AoU) generated whole-genome sequences of 245,388 individuals who collectively reflect the diversity of the USA. Leveraging this resource and another widely-used population-scale biobank, the UK Biobank (UKB) with a half million participants, we developed PRS trained on multi-ancestry and multi-biobank data with up to ~750,000 participants for 32 common, complex traits and diseases across a range of genetic architectures. We then compared effects of ancestry, PRS methodology, and genetic architecture on PRS accuracy across a held out subset of ancestrally diverse AoU participants. Due to the more heterogeneous study design of AoU, we found lower heritability on average compared to UKB (0.075 vs 0.165), which limited the maximal achievable PRS accuracy in AoU. Overall, we found that the increased diversity of AoU significantly improved PRS performance in some participants in AoU, especially underrepresented individuals, across multiple phenotypes. Notably, maximizing sample size by combining discovery data across AoU and UKB is not the optimal approach for predicting some phenotypes in African ancestry populations; rather, using data from only AoU for these traits resulted in the greatest accuracy. This was especially true for less polygenic traits with large ancestry-enriched effects, such as neutrophil count (R^2: 0.055 vs. 0.035 using AoU vs. cross-biobank meta-analysis, respectively, because of e.g. *DARC*). Lastly, we calculated individual-level PRS accuracies rather than grouping by continental ancestry, a critical step towards interpretability in precision medicine. Individualized PRS accuracy decays linearly as a function of ancestry divergence, but the slope was smaller using multi-ancestry GWAS compared to using European GWAS. Our results highlight the potential of biobanks with more balanced representations of human diversity to facilitate more accurate PRS for the individuals least represented in genomic studies.

Introduction

Population-scale biobanks with linked health records and genetic data have enabled an exponential increase in genome-wide association studies (GWAS), significantly expanding our understanding of the genetic basis of diseases[1,2]. Polygenic risk scores (PRS), which aggregate variant-disease associations discovered by GWAS, have been developed for many diseases and traits[3]. For some common, complex diseases, PRS have shown potential in aiding population risk stratification and screening, and their clinical implementation is on the horizon[4–7]. However, the vast majority of data used for GWAS still come from European ancestry (EUR) populations, resulting in the limited transferability of most PRS models to populations of other genetic ancestries [8]. This widely-recognized problem represents one of the most pressing challenges facing the clinical translation of PRS.

Several approaches can help mitigate this critical limitation. Statistical methods that leverage GWAS from multiple populations, including PRS-CSx and others, have been developed[9–12]. Benchmarking studies have evaluated these methods across traits of different genetic architectures using various study designs[13–15]. Complementing these empirical evaluations, theoretical studies have compared observed versus expected accuracies of PRS[13,16–18]. They find that while these methods can improve accuracy in some circumstances, the most direct path to increasing accuracy is through larger and more diverse study populations in GWAS.

Efforts like the Pan-UK Biobank (UKB) Project have maximized usage of current existing data resources by conducting GWAS for thousands of phenotypes using data from multiple ancestry groups, but its ancestral diversity is limited[19]. Other GWAS initiatives like the Global Biobank Meta-analysis Initiative (GBMI)[20] and disease- and trait-specific consortia, such as the Type 2 Diabetes Global Genomics Initiative[21] and the Genetic Investigation of ANthropometric Traits (GIANT)[22], focus on collecting ancestrally diverse data for meta-analysis. The Million Veterans Program is very large and diverse, and has recently conducted pan-trait and -ancestry GWAS, although access to summary statistics is more restricted[23].

Recent efforts have further expanded the availability of multi-ancestry genomic data. The All of Us Research Program (AoU), launched in 2018 by the National Institutes of Health of the United States, aims to gather health data from at least 1 million participants from diverse backgrounds. As of this study, it has released linked phenotypic and whole-genome sequencing data from 245,388 participants[24]. AoU is one of the largest and most accessible resources of populations traditionally underrepresented in biomedical research, with concerted efforts to capture ancestral diversity[24]. Given the ongoing efforts to increase the diversity of genomic studies, understanding how to best leverage multiple biobank resources to optimally predict complex traits with PRS will be a critical step towards their equitable applications.

Multi-ancestry PRS have been developed for a range of diseases and traits[13,25–27]. Recent studies have started utilizing the multi-ancestry data available in AoU[7,28–30]. However, the optimal approach for developing PRS from

multi-ancestry studies with large numbers of ancestrally diverse participants across population-scale biobanks remains unclear, especially across traits spanning a range of genetic architectures. Previous studies on optimal strategies for constructing multi-ancestry PRS have mostly used the UKB, which is not fully representative of the broader UK population and has limited ancestral diversity[15,31,32]. Additionally, studies investigating factors contributing to low PRS generalizability have largely focused on phenomena in population genetics, like the outsized impact of differences in allele frequencies and patterns of linkage disequilibrium (LD) on PRS accuracy[16,31]. Yet, there is also clear context-specificity to PRS accuracy that reflects factors like sex-specific heritability differences[33,34] and biobank-specific characteristics[35,36]. Our understanding of how differences between biobanks – for example, in ascertainment, data collection approaches, and sample recruiting strategy – impact polygenic prediction is still relatively limited. Some work on PRS development using multi-biobank data suggests that increases in sample size from combining heterogeneous biobanks can improve prediction performance for some diseases[25]. Furthermore, recent guidance on individualizing PRS performance evaluations have been based on single ancestry discovery cohorts[37], and understanding how this applies in multi-ancestry GWAS is an important outstanding question.

In this study, we developed PRS using multi-ancestry and multi-biobank data from AoU and UKB for dozens of commonly-studied diseases and quantitative traits with different genetic architectures. Specifically, we constructed PRS using single-ancestry GWAS from AoU, as well as multi-ancestry meta-analyses within and across AoU and UKB, to investigate the impacts of ancestry composition, sample size, trait genetic architecture, and biobank heterogeneity on PRS accuracy. Given the widespread adoption of UKB data, we also benchmark PRS performance with UKB. We illustrate nuance in optimal PRS strategies across phenotypes, particularly in underrepresented ancestry groups, providing guidelines and reference points for future PRS models developed in diverse genetic studies.

Results

Overview of study design

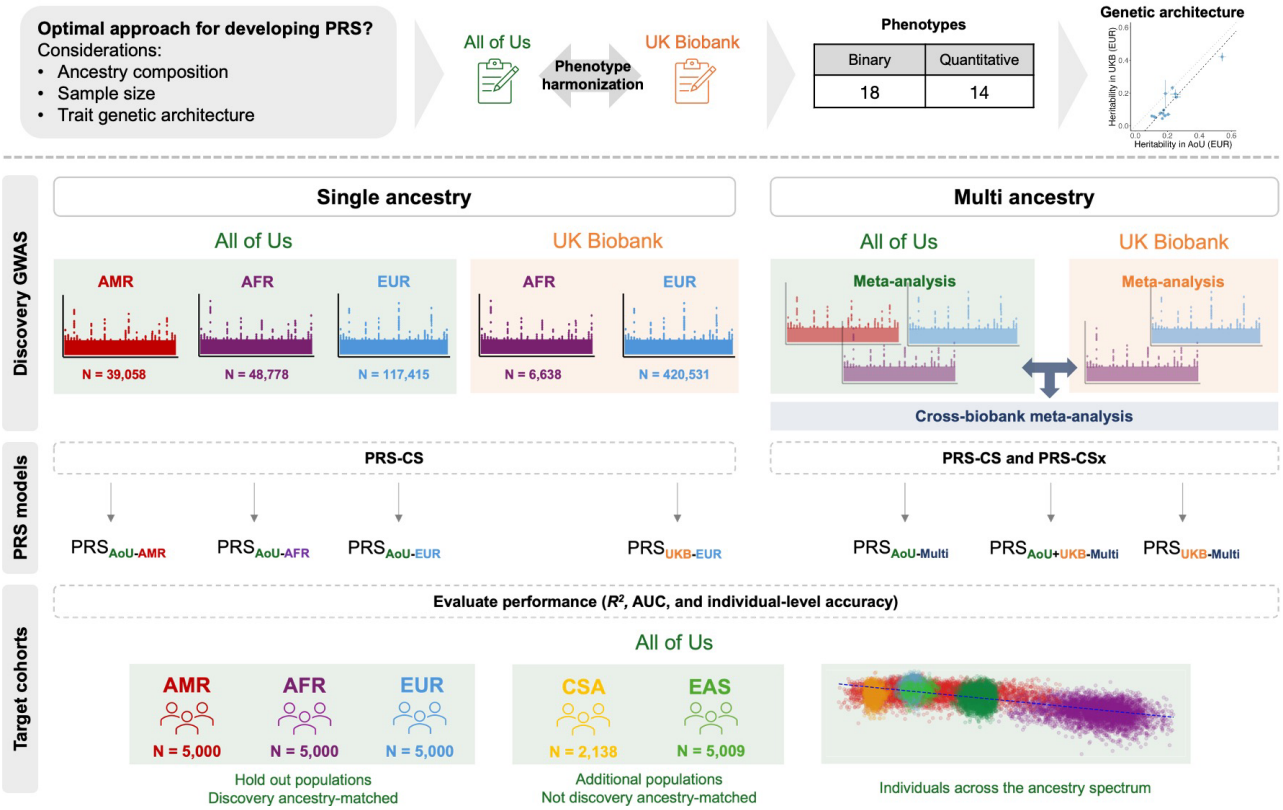

Figure 1. Study design for evaluating optimal PRS strategies that integrate ancestries and biobanks across multiple traits. Overview of workflow showing GWAS used for discovery data, methods for PRS construction, and cohorts used for PRS evaluation. AFR, African; AMR, admixed American; EAS, East Asian; MID, Middle Eastern; EUR, European; CSA, Central and South Asian.

Few frameworks have been developed for analyzing the wealth of phenotypic data available in AoU. We therefore adapted insights from previous UKB analyses. The Pan-UKB Project's quality control framework, which prioritizes phenotypes based on heritability estimates and other quality metrics, guided our phenotype selection[19]. From these prioritized phenotypes, we selected 14 quantitative and 18 binary phenotypes for our study based on data availability in AoU and other factors (**Methods**, **Supplementary Table 1**).

We assigned participants in AoU to genetically inferred ancestry groups based on principal component analysis (PCA) comparisons with population genetic reference panels (**Methods**). We trained a random forest model using labels from the Human Genome Diversity Panel (HGDP) and 1000 Genomes Project, which we use

throughout this study to refer to individuals with genetic ancestry most similar to those in the reference panels: EUR (European), AFR (African), AMR (Admixed American), CSA (Central/South Asian), and EAS (East Asian).

We conducted single-ancestry GWAS in AoU for all phenotypes using data from three groups with the largest sample sizes (N >10,000 including AFR, AMR, and EUR) (**Supplementary Table 1**). We combined GWAS across ancestries through inverse variance-weighted meta-analyses. For comparison, we included discovery GWAS from EUR and AFR populations in the UKB, excluding AMR due small sample size and unreliable genetic association results (**Figure 1**). Finally, we conducted cross-biobank, multi-ancestry meta-analyses.

To ensure consistency in phenotype definitions between AoU and UKB, we computed heritability estimates and genetic correlations across biobanks and population groups using LD score regression (LDSC) and Popcorn (**Methods, Supplementary Table 2**). We also compared effect sizes of genome-wide significant associations from biobank-specific GWAS (**Supplementary Fig. 1**) and raw phenotype distributions (**Supplementary Fig. 2**). Overall, our analyses indicated reasonable consistency between AoU and UKB phenotypes, although heritability estimates, which bound PRS accuracy, were significantly lower in AoU than UKB (sign test $p < 0.006$ for quantitative traits) (**Supplementary Fig. 6**)[17,38].

Using these GWAS and meta-analyses as training data, we constructed PRS using two Bayesian, genome-wide methods, PRS-continuous shrinkage (PRS-CS) and its multi-ancestry extension, PRS-CSx, as well as the classic pruning and thresholding method (P + T). We denoted PRS using the following nomenclature: $PRS_{[biobank]\text{-}[ancestry]}$, which indicates the GWAS data used to develop the PRS (e.g. $PRS_{AoU\text{-}AFR}$ refers to PRS from the GWAS of AFR individuals in AoU); $PRS_{[biobank]\text{-}Multi}$ was trained on the multi-ancestry meta-analyses from one or both biobanks (e.g. $PRS_{AoU+UKB\text{-}Multi}$ refers to PRS from the meta-analysis of GWAS from multiple ancestries in AoU and UKB). We assessed the performance of each PRS using incremental R^2 for quantitative traits and AUC for binary phenotypes in five ancestry groups with independent AoU target data (**Methods**). These included unrelated individuals from withheld EUR, AMR, and AFR groups (N=5,000 from each group), as well as CSA (N=2,138) and EAS (N=5,009).

Target ancestry-matched GWAS improve PRS performance for underrepresented ancestry groups

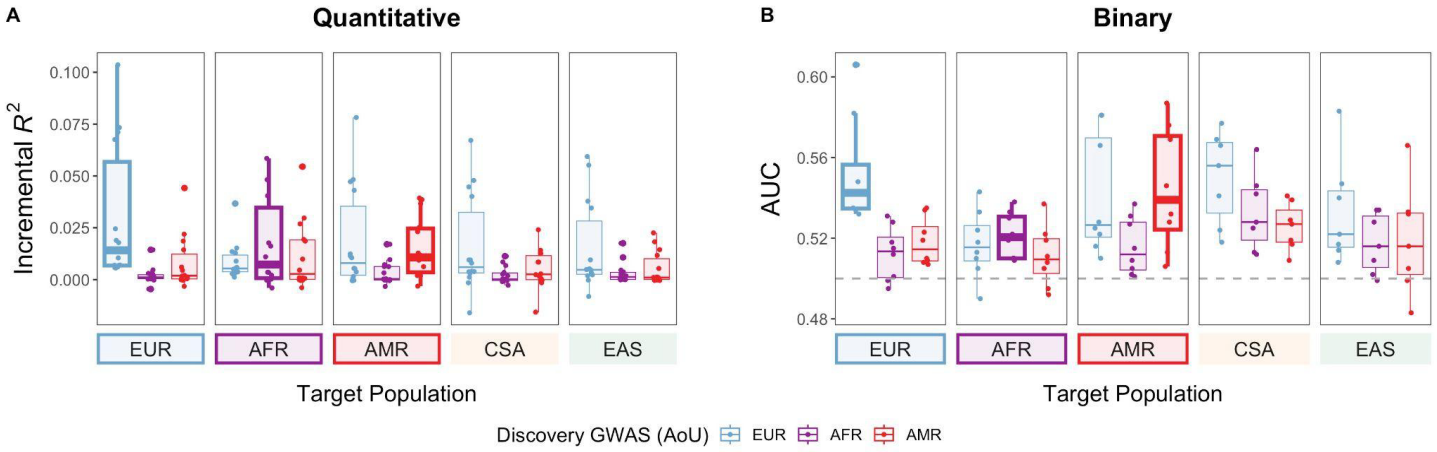

Figure 2. Single-ancestry discovery GWAS from AoU improve PRS performance for ancestry-matched target groups. Each point represents a phenotype, with PRS constructed from PRS-CS reported here. Target populations with ancestry-matched PRS are outlined.

Although the EUR group is still the largest single ancestry group in AoU, the sample sizes of the AFR and AMR groups in AoU are significantly larger compared to UKB (more than 7 and nearly 40 times larger, respectively). To determine if this increase in sample sizes improves PRS prediction accuracy in underrepresented ancestry groups, we first evaluated PRS constructed from single-ancestry GWAS in AoU. We focused on the results from PRS-CS in the following sections as PRS derived from PRS-CS outperformed or performed comparably to P+T (**Supplementary Tables 9 and 10**), consistent with previous findings[25]. As expected, in the EUR target group, $PRS_{AoU\text{-}EUR}$ significantly outperformed $PRS_{AoU\text{-}AFR}$ and $PRS_{AoU\text{-}AMR}$ across all quantitative traits (median R^2: 0.01 vs. 0.001 and 0.002, Wilcoxon rank sum exact test, p = 6.7e-06 and 5.3e-03, respectively) (**Fig. 2; Supplementary Table 3**). For the AFR and AMR groups, ancestry-matched discovery GWAS often performed best despite having much smaller sample sizes than EUR. $PRS_{AoU\text{-}AFR}$ achieved the highest median R^2 In the AFR target group across quantitative traits, a 1.4-fold increase compared to $PRS_{AoU\text{-}EUR}$ (median R^2: 0.007 vs. 0.003). Similarly, $PRS_{AoU\text{-}AMR}$ had highest accuracy in the AMR target group, with a 1.25-fold improvement over $PRS_{AoU\text{-}EUR}$ (median R^2: 0.01 vs. 0.008).

Despite larger sample sizes in the EUR GWAS, $PRS_{AoU\text{-}AFR}$ had greater accuracy than $PRS_{AoU\text{-}EUR}$ in the AFR target group for 8 out of the 14 quantitative traits; for 6 traits, $PRS_{AoU\text{-}AMR}$ had greater accuracy than $PRS_{AoU\text{-}EUR}$ in the AMR target group. This indicates that target ancestry-matched discovery GWAS can outperform larger-scale EUR-derived PRS in underrepresented ancestries with the sample sizes currently available in AoU. In the CSA and EAS target groups, $PRS_{AoU\text{-}EUR}$ generally performed best, but the median R^2 of $PRS_{AoU\text{-}EUR}$ in these

groups was lower than the the median R^2 of the corresponding ancestry-matched PRS in the AFR and AMR target groups, further highlighting the importance of ancestry matching between discovery and target groups.

Since the UKB has much larger sample sizes of EUR participants compared to AoU, we next investigated whether single-ancestry UKB training data improves prediction for underrepresented ancestry groups in AoU. Specifically, we evaluated $PRS_{UKB\text{-}EUR}$ in the AoU target populations. In the AMR target group, $PRS_{UKB\text{-}EUR}$ outperformed $PRS_{AoU\text{-}AMR}$ for all quantitative traits except neutrophil count, where $PRS_{AoU\text{-}AMR}$ showed a 2-fold improvement over $PRS_{UKB\text{-}EUR}$ (R^2: 0.02 vs. 0.01) (**Supplementary Fig. 3; Supplementary Table 3**). These results are expected given the low F_{ST} (0.02) between the AMR group in AoU and the EUR group in UKB, indicating relatively low genetic differentiation between these two groups (**Supplementary Fig. 4**). However, in the AFR target group, $PRS_{AoU\text{-}AFR}$ outperformed $PRS_{UKB\text{-}EUR}$ for 4 blood panel traits, and achieved comparable accuracy as $PRS_{UKB\text{-}EUR}$ for BMI and RBC count (BMI R^2: 0.16 vs. 0.17 and RBC count R^2: 0.11 vs. 0.13) (**Supplementary Fig. 3**). The >20-fold greater sample size of the EUR UKB vs. AFR AoU discovery groups (N=407,810 vs. N=18,044) did not result in significant PRS performance improvement for these traits. These results highlight the importance of training PRS on discovery cohorts that match the ancestry of target populations, particularly those with significant genetic differentiation from majority populations. Vast increases in EUR discovery sample sizes cannot compensate for the lack of training data from underrepresented groups.

We next investigated PRS performance for the binary phenotypes to compare with the well-powered quantitative traits. Due to overall smaller sample sizes (**Supplementary Table 1**), we limited evaluation of their PRS to diseases with at least 10,000 cases and larger heritability estimates (>0.03 in EUR), which included chronic ischaemic heart disease, chronic obstructive pulmonary disease (COPD), asthma, type 2 diabetes, lipid metabolism disorders, coronary atherosclerosis, esophagitis, and kidney stones. We observed similar patterns in PRS performance across these 8 disorders as we observed for the quantitative traits: the ancestry-matched PRS achieved the highest median AUC in each of the EUR, AFR, and AMR target groups (**Fig. 2**). Notably, in the AFR target group, the greatest improvements over $PRS_{AoU\text{-}EUR}$ were observed for asthma (AUC: 0.54 vs. 0.51) and lipid metabolism disorders (AUC: 0.53 vs. 0.51) (**Supplementary Table 4; Supplementary Fig. 5)**. $PRS_{AoU\text{-}AFR}$ had comparable AUC to $PRS_{UKB\text{-}EUR}$ for asthma (AUC: 0.54 vs. 0.53), despite the ~5-fold fewer cases of asthma among the AFR discovery group in AoU than in the EUR group in UKB (N=5,797 vs. N=31,030). For lipid metabolism disorders, $PRS_{AoU\text{-}AFR}$ had a 1.5% improvement over $PRS_{UKB\text{-}EUR}$ (AUC: 0.53 vs. 0.52).

Integrating multiple ancestries for discovery GWAS can improve PRS performance compared to single-ancestry GWAS

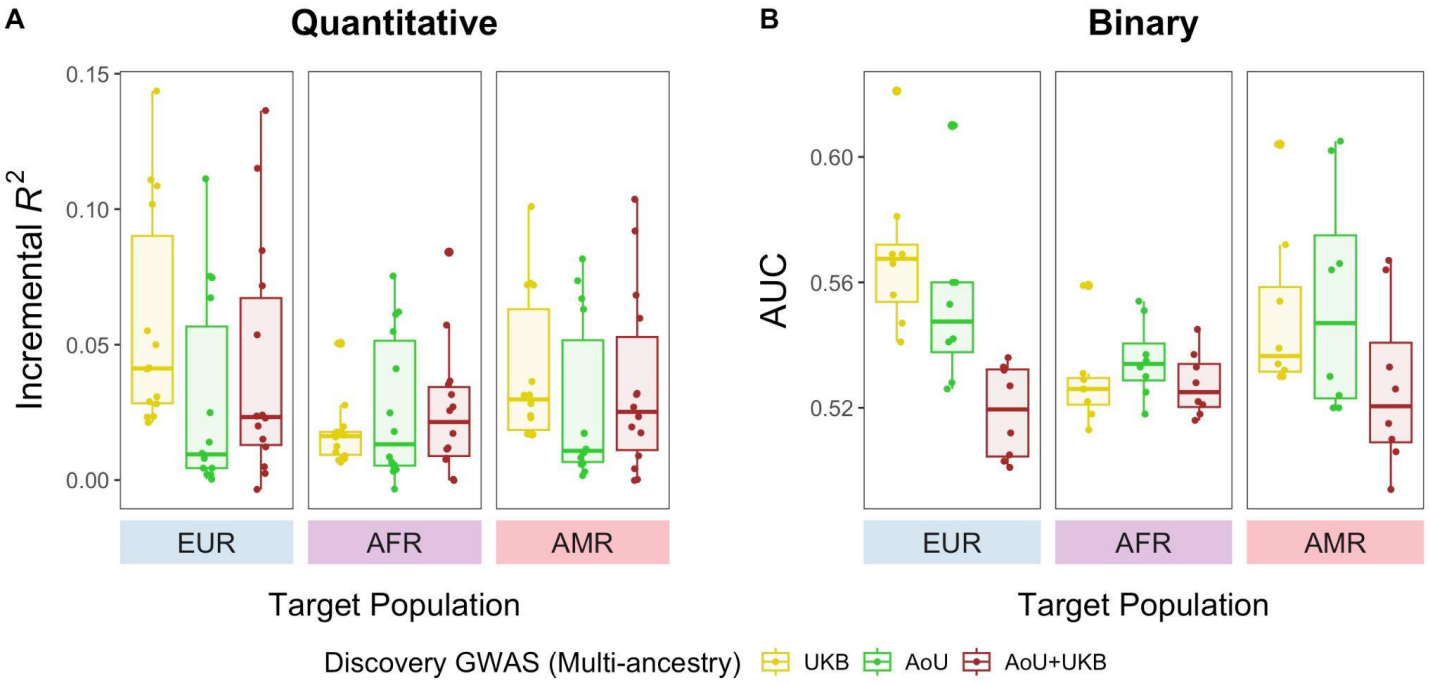

Figure 3. PRS derived from multi-ancestry meta-analyses show variable performance across target groups. Performance of PRS constructed from PRS-CS applied to UKB, AoU, and cross-biobank (AoU and UKB) multi-ancestry meta-analyses are reported here. Each point represents a phenotype.

Building on recommendations from previous studies[13,25], we next investigated how multi-ancestry meta-analyses affect PRS accuracy across quantitative and binary phenotypes. We first evaluated multi-ancestry meta-analyses from the UKB, and found that $PRS_{UKB\text{-}Multi}$ showed little to no improvement in PRS performance compared to $PRS_{UKB\text{-}EUR}$ across the target groups in AoU due to the vastly different sample sizes between EUR and AFR groups in the UKB (**Supplementary Table 3**).

We then evaluated the performance of PRS derived from the multi-ancestry AoU meta-analyses. Across the quantitative traits, $PRS_{AoU\text{-}Multi}$ had comparable accuracy to $PRS_{AoU\text{-}EUR}$ and $PRS_{AoU\text{-}AMR}$ in the EUR and AMR target groups, respectively (**Supplementary Table 3**). In the AFR target group, we observed an improvement of 0.6% in median R^2 compared to $PRS_{AoU\text{-}AFR}$, and $PRS_{AoU\text{-}Multi}$ outperformed $PRS_{AoU\text{-}AFR}$ for all quantitative traits. Accuracy gains from $PRS_{AoU\text{-}Multi}$ were especially large for some traits, including body mass index (BMI), mean corpuscular hemoglobin (MCH), mean corpuscular volume (MCV), neutrophil count, and white blood cell (WBC) count. Comparing the AoU and UKB meta-analyses, we found that in the EUR and AMR target groups, $PRS_{AoU\text{-}}$

$_{Multi}$ had lower performance across the traits compared to $PRS_{UKB\text{-}Multi}$ (**Fig. 3A**). The EUR group dominates the multi-ancestry UKB meta-analyses, and given that $PRS_{UKB\text{-}EUR}$ outperformed the target-ancestry matched PRS in these groups while $PRS_{AoU\text{-}Multi}$ did not, the difference in performance between $PRS_{AoU\text{-}Multi}$ and $PRS_{UKB\text{-}Multi}$ was expected. The low genetic differentiation, measured by F_{ST}, between the AMR in AoU and EUR in UKB, as well as between the EUR groups in both biobanks, further supports these results (**Supplementary Fig. 4**): not only is the EUR group in the UKB meta-analyses much larger than in the AoU meta-analyses, it is also genetically proximal to the AMR and EUR groups in AoU, thus contributing to the superior performance of $PRS_{UKB\text{-}Multi}$. Additionally, SNP-based heritability estimates (h^2), calculated using EUR GWAS from AoU and UKB, indicated systematically lower heritability in AoU than UKB (**Supplementary Fig. 6; Supplementary Table 5**). As PRS accuracy is bounded by h^2, this likely also contributed to the decreased performance of $PRS_{AoU\text{-}Multi}$ in the EUR and AMR groups.

To gauge the value of combining AoU and UKB for discovery, we next evaluated PRS derived from the cross-biobank multi-ancestry meta-analyses. In the AFR target group, $PRS_{AoU+UKB\text{-}Multi}$ offered some improvement in median R^2 compared to $PRS_{AoU\text{-}Multi}$ and $PRS_{UKB\text{-}Multi}$ (0.021 vs. 0.013 and 0.016, respectively) (**Fig. 3A**). However, that improvement depended on genetic architecture: prediction in more polygenic traits (**Supplementary Table 6**) such as BMI and DBP benefited from the increase in sample size in the cross-biobank meta-analyses; conversely, $PRS_{AoU\text{-}Multi}$ outperformed $PRS_{AoU+UKB\text{-}Multi}$ for less polygenic traits or those with large-effect ancestry-enriched variants, such as MCH and MCV.

$PRS_{AoU\text{-}Multi}$ had varying performance in the disease phenotypes as well (**Supplementary Table 4**). In the AFR target group, $PRS_{AoU\text{-}Multi}$ did not improve prediction performance in the diseases where $PRS_{AoU\text{-}AFR}$ outperformed $PRS_{AoU\text{-}EUR}$ (COPD, asthma, and lipid metabolism disorders). However, for ischaemic heart disease and coronary atherosclerosis, $PRS_{AoU\text{-}Multi}$ showed increased performance compared to $PRS_{AoU\text{-}AFR}$ (AUC: 0.55 vs. 0.51 for both diseases). In the AMR target group, $PRS_{AoU\text{-}Multi}$ marginally improved AUC compared to $PRS_{AoU\text{-}AMR}$ for T2D (AUC: 0.61 vs. 0.58) and COPD (AUC: 0.60 vs. 0.59). In both the AFR and AMR target groups, $PRS_{AoU+UKB\text{-}Multi}$ did not offer improved prediction compared to $PRS_{AoU\text{-}Multi}$ or any single-ancestry PRS_{AoU} across the diseases (**Fig. 3B**).

Finally, we compared the performances of multi-ancestry PRS developed using PRS-CS vs. PRS-CSx (**Supplementary Table 7; Supplementary Table 8**). In the AFR target group across the quantitative traits, PRS-CSx improved median R^2 by 0.008 over PRS-CS for $PRS_{AoU\text{-}Multi}$, with substantial improvements in alanine aminotransferase, BMI, MCH, MCV, and red blood cell (RBC) count. PRS-CSx did not significantly improve performance of $PRS_{AoU\text{-}Multi}$ in the EUR or AMR target groups. Across the binary phenotypes, applying PRS-CSx did not improve performance of $PRS_{AoU\text{-}Multi}$ in the EUR, AFR, and AMR target groups.

Optimal PRS differs across phenotypes and target ancestries

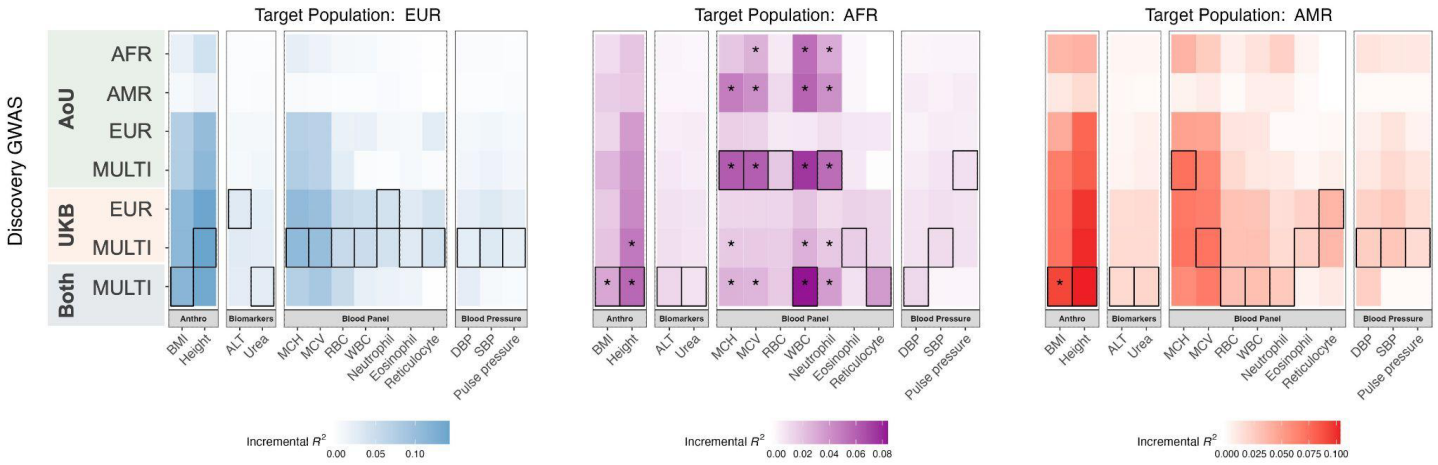

Figure 4. AoU discovery data improve PRS performance in AFR target group. Performance of all PRS models, denoted on y-axis, across quantitative traits, denoted on x-axis. PRS model with greatest R^2 per trait is outlined. Asterisk indicates significantly greater prediction accuracy than that of the PRS derived from the EUR UKB discovery group (Wald test, $p < 0.05$).

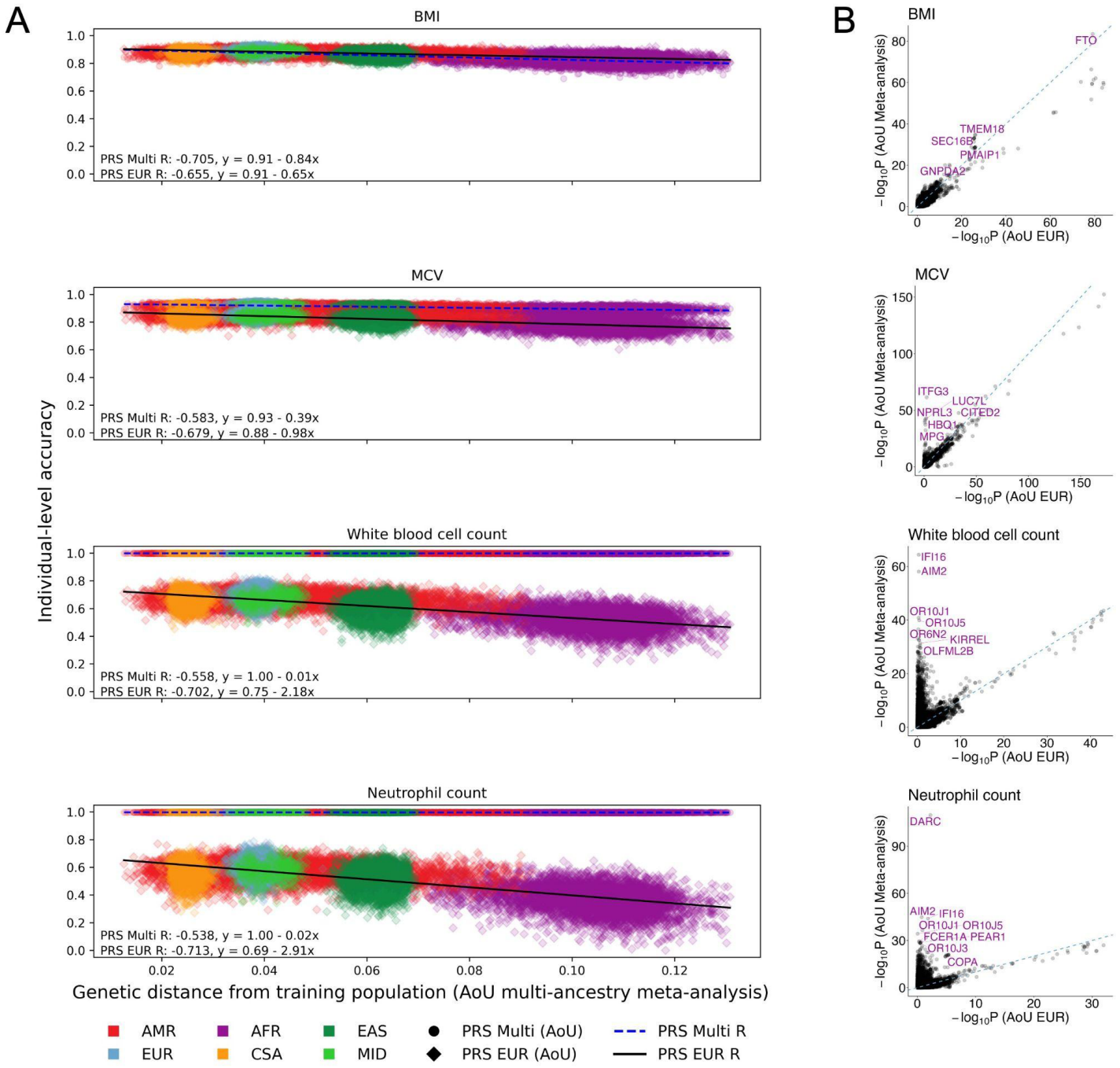

Figure 5. PRS derived from multi-ancestry meta-analyses for blood panel traits show improved accuracy on individual-level, driven by ancestry-enriched variants. A) Individual-level accuracy of PRS derived from AoU multi-ancestry meta-analyses and EUR GWAS across target individuals in AoU, represented by each point. The x-axis represents the genetic distance (GD) of each target individual from the combined discovery populations included in the AoU multi-ancestry meta-analyses. The y-axis shows the PRS accuracy, which was scaled to enable cross-trait comparisons of decay in accuracy as a function of GD; as a result, proportions of genetic liability explained by PRS for each individual are not represented here. *R* was calculated as the correlation between GD and

PRS accuracy from a two-sided Pearson correlation test. The colors represent genetic ancestry groups as inferred by PCA. B) Comparison of GWAS significance in AoU multi-ancestry meta-analyses and AoU EUR GWAS across blood panel traits. SNPs tested in both the AoU multi-ancestry meta-analyses and EUR GWAS are represented by each point. SNPs reaching genome-wide significance ($p < 5e-8$) in the AoU meta-analysis and AoU AFR GWAS for each phenotype are annotated. Dashed lines indicate y=x; x- and y-axis scales are specific to each phenotype and differ according to scale of significance in meta-analyses vs. EUR GWAS.

To identify the best-performing PRS, we compared all PRS models constructed from PRS-CS for each phenotype, focusing on the target groups with ancestry-matched PRS (**Fig. 4; Supplementary Fig. 7**). We tested for significant differences of prediction accuracy between each PRS and $PRS_{UKB\text{-}EUR}$, the best-powered single-ancestry PRS in this study (Wald test, p-value < 0.05 indicates significance). We found that in the EUR and AMR target groups, no PRS significantly improved prediction accuracy over $PRS_{UKB\text{-}EUR}$, except for the BMI $PRS_{AoU+UKB\text{-}Multi}$ in the AMR group (R^2: 0.09 vs. 0.07). However, in the AFR target group, we observed that for 6 out of the 14 quantitative traits, the accuracy of $PRS_{AoU+UKB\text{-}Multi}$ or $PRS_{AoU\text{-}Multi}$ was significantly higher than that of $PRS_{UKB\text{-}EUR}$, underlining the importance of using target ancestry-matched discovery data for populations with large genetic distances from EUR populations.

Improvements in PRS accuracy using data from AoU were largest for 4 quantitative traits in the AFR group: MCH, MCV, WBC count and neutrophil count. $PRS_{AoU\text{-}AFR}$ increased in accuracy over $PRS_{UKB\text{-}EUR}$ by almost 4-fold for MCH (R^2: 0.048 vs. 0.013) and neutrophil count (R^2: 0.041 vs. 0.010), and 3-fold for MCV (R^2: 0.040 vs. 0.013) and WBC count (R^2: 0.058 vs. 0.021). $PRS_{AoU\text{-}Multi}$ offered additional improvements in R^2 over $PRS_{AoU\text{-}AFR}$, although to a more modest degree of ~1.3-1.5 fold across these 4 traits.

Based on recent work proposing a shift from population- to individual-level metrics of PRS accuracy[37,39], we next examined individual-level PRS accuracy as a function of genetic distance (GD) using multi-ancestry AoU discovery data (**Methods**). We focused on the four blood panel traits for which $PRS_{AoU\text{-}Multi}$ performed best. For baseline comparison, we first computed individual PRS accuracy using the EUR GWAS from AoU. Across the blood panel traits, height, and BMI, PRS accuracy decreased with increasing GD from both the EUR and multi-ancestry discovery groups, consistent with previous findings[37] (**Supplementary Fig. 8, Fig. 5A**). Among the blood panel traits, we observed the largest decay in individual-level PRS accuracy in neutrophil count, WBC count, and MCV, described by more negative slopes and lower intercepts (slopes = -2.91, -2.18, and -0.98; intercepts = 0.69, 0.75, and 0.88) (**Fig. 5A**). In contrast, individual-level accuracy computed from the multi-ancestry AoU meta-analyses showed nearly no decay across the genetic ancestry spectrum for neutrophil and WBC count, and less decay for MCV (slopes = -0.02, -0.01, and -0.39; intercepts = 1.00, 1.00, and 0.93). However, for BMI, individual-level accuracy from the multi-ancestry meta-analysis showed greater decay than

the EUR GWAS (slopes = -0.84 vs. -0.65). For MCH and height, the linear decay in individual-level accuracy was still present using the multi-ancestry meta-analyses as discovery, but that decay was attenuated, as for WBC and neutrophil count (**Supplementary Fig. 9**).

Studies[31,40] have previously highlighted that greater diversity in the discovery data showed outsized improvements in PRS accuracy for certain blood panel traits, including MCV and WBC count, likely due to specific genetic loci that disproportionately explain population-specific risk and are more common in underrepresented ancestry groups. Indeed, we found that a few genome-wide significant loci from the AFR GWAS in AoU were highly significant in the AoU meta-analyses but not the EUR GWAS, including those closest to *DARC* associated with neutrophil count and *ITFG3* associated with MCH and MCV (**Fig. 5B**), likely driving the increased accuracy of $PRS_{AoU\text{-}Multi}$ in the AFR group and for individuals furthest in GD from the discovery data. These traits also had relatively lower polygenicity estimates, ranging from 0.011-0.014, compared to the other quantitative traits (**Supplementary Table 6**). Thus, population genetic factors and genetic architecture contribute to improved accuracy from AoU multi-ancestry training data on both the population- and individual-level.

Discussion

PRS are already being tested in clinical settings for a variety of diseases. For example, the eMERGE Network identified, validated, deployed, and returned PRS to patients for 10 clinical conditions, including heart disease, asthma, and type 1 and 2 diabetes[7]. This study ultimately spanned four years, highlighting the challenge of translating rapidly evolving GWAS findings into clinical practice. Given the remarkable polygenicity of common complex diseases, the rapid growth of GWAS, and where we are on the genomic discovery curve for most diseases, this lag time is particularly challenging[1]. Nimbleness is needed for PRS to be maximally effective in the clinic. However, studies have shown poor agreement between individuals at the extremes of the PRS distribution when using different GWAS with a best-case overlap of 60% of individuals above the 80th percentile[41]. Additionally, while it is widely recognized that PRS have different accuracies across ancestry groups mostly due to LD and allele frequency differences[8], PRS generalizability remains a critical challenge; large-scale datasets most commonly used for PRS development and evaluation are often skewed in representation.

The AoU Research Program offers a substantially more diverse resource of phenotypic and genomic data compared to other large-scale contemporary biobanks. This important step towards diversifying human genetic datasets raises new questions for PRS development, particularly for historically underrepresented groups. Our study investigated whether the sample sizes of diverse ancestry groups currently available in AoU are sufficient to increase PRS performance. We found that individuals in the AFR target group benefited most from AoU data, particularly from multi-ancestry meta-analyses. However, AoU discovery data did not significantly improve PRS accuracy in other ancestry groups compared to the largest EUR GWAS from UKB. Encouragingly, for some traits

with ancestry-enriched variants, AoU multi-ancestry meta-analyses substantially improved PRS accuracy for individuals furthest in GD from the training data.

Combining AoU and UKB GWAS in cross-biobank meta-analyses did not uniformly yield improved accuracy across the phenotypes and target groups, despite the increase in sample size. This highlights the complexity of developing optimal PRS, which is affected by complex interactions between sample size, ancestry matching of discovery and target cohorts, genetic architecture, and phenotype precision. Cross-biobank and cross-population genetic correlation estimates, for example, indicated greater alignment in phenotypes between the EUR groups in UKB and AoU, compared to the EUR and AFR groups in AoU. However, the overall lower h^2 estimated from AoU GWAS compared to UKB points to the greater heterogeneity of AoU, likely due to study design, recruitment strategies, and the diversity of hospital systems in the US. This heterogeneity between biobanks likely contributed to the comparatively decreased accuracy of PRS from the cross-biobank meta-analyses for some traits and ancestries. Understanding the impacts of inter- and intra-biobank heterogeneity on PRS accuracy will be important as AoU and other biobanks, like the Million Veteran Program[23], continue to grow in scale and diversity.

As the trajectory of PRS development advances towards clinical implementation, understanding the absolute risk conferred by PRS is crucial for translation. Although individualizing PRS metrics of accuracy is an important step towards translation, additional investigations into the calibration and interpretation of PRS will be needed. For example, integrating PRS into clinical models with other known risk factors that vary in frequency across healthcare systems is an important area for future investigation. Future work should also assess the effects of non-genetic risk factors, which differ across individuals and populations, on PRS accuracy as more clinical and environmental data becomes available in AoU and other diverse biobanks.

Methods

Datasets and quality control:

Pan-UK Biobank (Pan-UKB): The UK Biobank (UKB) is an extensively utilized cohort comprising approximately 500,000 participants from the United Kingdom, ranging in age from 40 to 69 years. Detailed documentation concerning this cohort has been previously reported[42]. In pursuit of harnessing the rich diversity present within the UKB beyond the customary European ancestry individuals, the Pan-UKB project (https://pan.ukbb.broadinstitute.org/) has undertaken a comprehensive multi-ancestry investigation. This project encompasses 7,228 distinct phenotypes across 6 continental ancestry groups, with a cumulative total of 16,131 GWAS. Rigorous quality control procedures were applied to scrutinize the phenotypic-level, individual-level, and variant-level data, with comprehensive details available in Karczewski et al.[19]

The All of Us Research Program (AoU): The All of Us Research Program, launched by National Institute Health in May 2018, represents a longitudinal cohort study with the goal of engaging at least 1 million participants encompassing diverse ancestral backgrounds. By leveraging comprehensive data collection including biospecimens, health questionnaires, electronic health records and physical measurements, AoU aims to advance precision medicine and enhance overall human health[43]. Participants, aged 18 years and older, are recruited from over 340 centers with informed consent. As of April 2023, a subset of around 250,000 participants has undergone whole genome sequencing (WGS). We assigned those individuals with WGS data into the nearest genetic ancestry based on principal components (PCs), resulting in 49,778 of African descent (AFR), 39,058 of American descent (AMR), 2,138 of Central and South Asian descent (CSA), 5,183 of East-Asian descent (EAS), 117,415 of European descent (EUR) and 432 of Middle Eastern descent (MID). The strategy was the same as described in the pan-UKB project[19]. Briefly, we projected all AoU individuals into the PC space using pre-estimated weights of 168,899 variants[20] from the Human Genome Diversity Panel (HGDP)[44] and 1000 Genomes Project[45]. For individuals with a probability > 50% from the random forest, we further refined initial ancestry assignments by pruning outliers within each continental assignment. We reran PCA within each assigned continental ancestry group and calculated total distances from population centroids across 10 PCs. Using these PC scores, we computed centroid distances across 3-5 centroids based on the heterogeneity within each group. We identified and removed ancestry outliers by plotting histograms of centroid distances and excluding individuals at the extreme high end.

Given the limited sample size within CSA, EAS and MID ancestral populations, we exclusively used them as independent test cohorts. For EUR, AMR and AFR populations, we split the data into separate training and test sets. Specifically, in each population, we randomly selected 5,000 individuals from unrelated samples as the withheld test dataset. We used the remaining individuals as the training dataset, which included related individuals to improve statistical power. To avoid relatedness between test and training dataset, we subsequently removed individuals in the training dataset that showed a kinship coefficient larger than 0.1 with any individual in the test dataset. The estimates of kinship coefficient were provided by AoU. We removed those individuals who did not pass AoU quality controls. Consequently, we used 43,926, 33,330 and 111,850 individuals as the training dataset for AFR, AMR and EUR, respectively. For the variant-level quality controls, we focused on only HapMap 3 variants and further removed those with minor allele frequency (MAF) lower than 0.01, genotype missing rates larger than 0.05 and hardy-weinberg equilibrium (HWE) *p*-value smaller than 1e-6.

Phenotypes:

UKB: For those 492 high quality phenotypes that passed different filters as described in Karczewski et al.[19], we calculated the variance explained by the top genome-wide significant loci as $\sum 2p(1-p)\beta^2$ where p is the MAF and β denotes the estimated per-allele effect sizes on the standardized phenotype. The top loci were defined using clumping in PLINK[46] based on ancestry-matched reference panels from UKB; more details can be

found in Karczewski et al.[19]. We identified a subset of 129 phenotypes, characterized by a greater variance explained in the multi-ancestry meta-analyzed GWAS in comparison to EUR-based GWAS. We focused on this subset of phenotypes, considering the potential to improve predictive accuracy in underrepresented populations by leveraging multi-ancestry discovery GWAS. Subsequently, those selected phenotypes were subject to further in-depth investigation in the AoU.

AoU: To enhance the quality and reliability of the phenotypic data available within the AoU, we curated and processed the phenotypes through a few steps. First, we checked whether there are matched phenotype descriptions in AoU based on data-field notes in the UKB showcase (https://biobank.ndph.ox.ac.uk/showcase/). Phenotypes derived from survey data were subsequently excluded from consideration. Following this filtering process, phenotypes with either matched or closely related descriptions in AoU were selected for further evaluation. We also added a few commonly studied quantitative traits (BMI, height, and eosinophil count), as well as three additional common diseases with high impact on public health (COPD, asthma, and coronary atherosclerosis). This resulted in 14 quantitative phenotypes, 7 ICD-10 codes and 11 PheCodes for all downstream analyses (**Supplementary Table 1**). The curation of raw phenotypic data encompassed a comprehensive analysis based on concept IDs, and the most recent measurements were sourced from diverse domains, such as conditions, lab and physical measurements, and surveys. For the PheCode curation, we employed the PheCode map v1.2 (https://phewascatalog.org/phecodes) to map ICD codes into corresponding phecodes. Notably, lab and physical measurements often exhibited variations in measurement units across individuals. To address this issue, the most frequent unit of measurement was adopted as a reference, and appropriate conversions were applied to standardize other units accordingly. In order to optimize the sample size available for analysis, individuals for whom the unit concept name was indicated as “empty”, "no matching concept," or "no value" were retained in the dataset. For quantitative phenotypes, individuals with values exceeding 5 standard deviations from the mean were systematically excluded from the dataset to ensure the robustness of subsequent analyses.

Genome-wide association studies (GWAS):

The Pan-UK Biobank Project, described in Karczewski et al.[19], has publicly released individual GWAS in each ancestry as well as meta-analyzed GWAS across ancestries. We utilized AFR and EUR GWAS, as well as the meta-analyzed GWAS across the AFR and EUR groups, from this resource.

The phenotypes within the AoU were processed using the same strategy described in Karczewski et al.[19], where the quantitative phenotypes were inverse-ranked normalized. We performed GWAS on the training datasets within AFR, AMR and EUR populations as described previously using the Regenie software[47]. Only the quantitative phenotypes with sample size larger than 5,000 and binary traits with case counts exceeding 100

were included for GWAS analysis. We included the follow covariates: age, sex, age^2, age*sex, age^2 * sex, and the first 10 PCs.

We then conducted meta-analyses of the AoU GWAS data with the UKB GWAS data, separately for EUR and AFR, as well as all ancestry groups combined. Meta-analyses across three ancestry groups within AoU were also performed. The meta-analyses were performed using the inverse-variance weighted approach in the METAL software[48]. Our analyses focused on common HapMap 3 variants only.

Genetic architecture estimates:

In this study, we investigated the impact of key parameters of genetic architecture on the performance of PRS. We assessed several trait-specific genetic architecture parameters, namely polygenicity (i.e. the proportion of SNPs with nonzero effects) and SNP-based heritability. To estimate polygenicity, we employed SBayesS, a summary statistics based method employing a Bayesian mixed linear model, with its default settings[49]. The input datasets for this analysis were the EUR GWAS from UKB. To estimate heritability, we conducted LD score regression analyses using LDSC[50] based on the AoU EUR GWAS, and obtained the LDSC estimates based on the UKB EUR GWAS from Karczewski et al.[19]. We used ancestry-matched reference panels from UKB for these analyses[19].

Genetic correlation estimates:

To estimate r_g between the EUR GWAS from AoU and UKB, we used the heritability Z-scores obtained from LDSC computations of heritability from AoU GWAS and as reported in Karczewski et al.[19] from UKB GWAS. To estimate cross-ancestry r_g between the EUR and AFR GWAS from AoU, and EUR and AMR GWAS from AoU, we used Popcorn[51] based on 1000 Genomes reference panels.

PRS construction and evaluation:

We constructed PRS using three different methods: the classic pruning and thresholding (P+T) method, and two Bayesian genome-wide methods, namely PRS-CS[52] and PRS-CSx[9]. P+T was performed using a LD r^2 threshold of 0.1 and a series of *p*-value thresholds (5e-8, 5e-07, 5e-06, 5e-05, 5e-04, 5e-03, 0.05, 0.1, 1). We used the auto model, which automatically estimates the global shrinkage parameter, implemented in PRS-CS and PRS-CSx. We used ancestry-specific AoU GWAS as inputs for the three methods. For P+T and PRS-CS, multi-ancestry meta-analyzed GWAS were additionally included. In order to comprehensively explore the advantages of incorporating AoU data, we constructed PRS using UKB GWAS data independently, as well as the meta-analyzed AoU and UKB GWAS data.

The LD reference panel used was dependent on the ancestry composition of the discovery GWAS. We used LD panels that matched the respective ancestral population for ancestry-specific GWAS. Since the multi-ancestry meta-analyzed GWAS primarily comprised European individuals, we used a European-based panel, as our previous studies demonstrated that it can adequately approximate the LD structure[13,25]. We used the pre-computed LD matrices obtained from Karczewski et al.[19] for P+T. Additionally, for PRS-CS and PRS-CSx, we employed the LD matrices provided by the software, which were computed from UKB data. We evaluated PRS performance in independent target datasets of AFR, AMR, EUR, EAS, and CSA ancestries within the AoU dataset. To evaluate the PRS performance for quantitative phenotypes, we estimated incremental R^2 by accounting for the covariates. Specifically, we compared two models: 1) the baseline model (phenotype ~ covariates) and 2) the full model including PRS (phenotype ~ PRS + covariates). Incremental R^2 represents the improvement in model accuracy with the inclusion of PRS. For binary phenotypes, we reported the Area Under the Receiver Operating Characteristic Curve (AUC) of PRS solely, Nagelkerker's R^2, and R^2 on the liability scale. In the latter case, we approximated the disease prevalence using the population prevalence. We calculated the corresponding 95% confidence intervals (CIs) of each estimate using 1,000 bootstrap iterations. For the P+T method, we adopted a two-step evaluation approach. First, we partitioned the target datasets evenly into a validation cohort and a test cohort. Next, we fine-tuned the *p*-value threshold using the validation cohort to optimize performance. Subsequently, we evaluated the PRS performance on the test cohort using the fine-tuned *p*-value threshold. This procedure ensured a robust evaluation of the PRS performance based on the optimal thresholds.

Estimates of population genetic differentiation:

To characterize the genetic distance between populations across the biobanks, we measured population genetic differentiation with Wright's fixation index, F_{st}, computed using the "wc" method in PLINK 2.0[46]. The analyses were performed using 168,899 pruned variants.

Individual PRS accuracy:

Posterior effect size calculation: We used the EUR GWAS and multi-ancestry meta-analysis from AoU as inputs for PRS-CS. Using the default setting of PRS-CS, which involves 1000 MCMC (Markov Chain Monte Carlo) iterations, 500 burn-in iterations, and a thinning factor of 5, we obtained an output of 100 sets of posterior effect estimates for each variant in an *M*x100 matrix, where *M* is the number of SNPs. This matches the output based on LDPred2 in Ding et al.[37]

PRS accuracy: We used the "--score" flag in PLINK 2.0 to compute 100 sets of PRS for the individuals in the AMR, AFR, EUR, CSA, and EAS target groups in AoU. The output matrix has shape $100 \times 22{,}703$ and each cell is denoted as PRS_b^m, where $m \in$ [1,22703] denotes the m individual and $b \in$ [1,100] denotes the b set of PRS.

Based on Ding et al.[37], the individual PRS uncertainty for individual m for empirical analyses is calculated as $var(PRS^m)$, that is the variance of 100 sets of PRS. The PRS accuracy for individual m is defined as $1 - \frac{var(PRS^m)}{h^2 var(y_{residue})}$, where h^2 denotes the estimated heritability from LDSC and $var(y_{residue})$ denotes the variance of residue phenotype in training data after regressing out age, sex, age^2, age*sex, age^2 * sex, and the first 10 PCs. The PRS accuracies for four blood panel traits (neutrophil count, white blood cell count, mean corpuscular volume, and mean corpuscular hemoglobin), for which $PRS_{AoU\text{-}Multi}$ performed best, and two additional polygenic traits (height and BMI) for comparison were scaled using min-max normalization ranging from 0 to 1, where the minimum and maximum values correspond to the smallest and largest PRS accuracy observed among individuals across all six traits, respectively. The correlation coefficient *R* was measured by Pearson correlation.

Genetic distance: We used the same strategy as described in Ding et al.[37] to calculate genetic distance between each individual and the discovery population. Briefly, we calculated the Euclidean distance of the PCs of the individuals in the target groups from the center of the discovery data, i.e. either the EUR or all groups in AoU.

Acknowledgements

We acknowledge helpful comments from Mark Daly and Konrad Karczewski. A.R.M is funded by the K99/R00MH117229. K.T. is funded by F31HL167378 and supported by the ECOR Claflin Award to A.R.M. A.R.M. and Y.W. are funded by U01HG011719. Additional support for this work to A.R.M. and Y.W. also comes from the European Union's Horizon 2020 research and innovation program under grant agreement 101016775 (INTERVENEConsortium). B.P. is supported by U01HG011715.

Declaration of Interests

A.R.M. has received speaker fees from Novartis.

References

1. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Abul-Husn, N. S. & Kenny, E. E. Personalized Medicine and the Power of Electronic Health Records. *Cell* **177**, 58–69 (2019).
3. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *Nature* **591**, 211–219 (2021).
4. Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci* **5**, 293–320 (2022).
5. Vassy, J. L. *et al.* The GenoVA study: Equitable implementation of a pragmatic randomized trial of polygenic-risk scoring in primary care. *Am. J. Hum. Genet.* **110**, 1841–1852 (2023).
6. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk

equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
7. Lennon, N. J. *et al.* Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. *Nat. Med.* **30**, 480–487 (2024).
8. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
9. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
10. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
11. Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. & Tang, H. Leveraging Multi-ethnic Evidence for Risk Assessment of Quantitative Traits in Minority Populations. *Am. J. Hum. Genet.* **101**, 218–226 (2017).
12. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* **25**, 8–25 (2024).
13. Wang, Y. *et al.* Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genom* **3**, 100408 (2023).
14. Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* **17**, e1009021 (2021).
15. Thompson, D. J. *et al.* UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *bioRxiv* (2022) doi:10.1101/2022.06.16.22276246.
16. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
17. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3**, e3395 (2008).
18. de Vlaming, R. *et al.* Meta-GWAS Accuracy and Power (MetaGAP) Calculator Shows that Hiding Heritability Is Partially Due to Imperfect Genetic Correlations across Studies. *PLoS Genet.* **13**, e1006495 (2017).
19. Karczewski, K. J. *et al.* Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. *medRxiv* 2024.03.13.24303864 (2024) doi:10.1101/2024.03.13.24303864.
20. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genom* **2**, 100192 (2022).
21. Suzuki, K. *et al.* Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, 347–357 (2024).
22. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
23. Verma, A. *et al.* Diversity and Scale: Genetic Architecture of 2,068 Traits in the VA Million Veteran Program. *medRxiv* (2023) doi:10.1101/2023.06.28.23291975.
24. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
25. Wang, Y. *et al.* Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genom* **3**, 100241 (2023).
26. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* **29**, 1793–1803 (2023).
27. Ge, T. *et al.* Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Med.* **14**, 70 (2022).
28. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves performance across

598 diverse populations. *Nat. Genet.* **55**, 1757–1768 (2023).
599 29. Chen, R., Petrazzini, B. O., Malick, W. A., Rosenson, R. S. & Do, R. Prediction of Venous
600 Thromboembolism in Diverse Populations Using Machine Learning and Structured Electronic Health
601 Records. *Arterioscler. Thromb. Vasc. Biol.* **44**, 491–504 (2024).
602 30. Zhang, J. *et al.* An ensemble penalized regression method for multi-ancestry polygenic risk prediction.
603 *Nat. Commun.* **15**, 3238 (2024).
604 31. Lehmann, B., Mackintosh, M., McVean, G. & Holmes, C. Optimal strategies for learning multi-ancestry
605 polygenic scores vary across traits. *Nat. Commun.* **14**, 4023 (2023).
606 32. He, Y. & Martin, A. R. We need more-diverse biobanks to improve behavioural genetics. *Nat Hum Behav*
607 (2023) doi:10.1038/s41562-023-01795-3.
608 33. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9**,
609 (2020).
610 34. Hou, K., Xu, Z., Ding, Y., Harpak, A. & Pasaniuc, B. Calibrated prediction intervals for polygenic scores
611 across diverse contexts. *medRxiv* (2023) doi:10.1101/2023.07.24.23293056.
612 35. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank
613 Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
614 36. Monti, R. *et al.* Evaluation of polygenic scoring methods in five biobanks shows larger variation between
615 biobanks than methods and finds benefits of ensemble learning. *Am. J. Hum. Genet.* **111**, 1431–1447
616 (2024).
617 37. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**,
618 774–781 (2023).
619 38. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions.
620 *Nat. Rev. Genet.* **9**, 255–266 (2008).
621 39. Ding, Y. *et al.* Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk
622 stratification. *Nat. Genet.* **54**, 30–39 (2022).
623 40. Majara, L. *et al.* Low and differential polygenic score generalizability among African populations due
624 largely to genetic diversity. *HGG Adv* **4**, 100184 (2023).
625 41. Schultz, L. M. *et al.* Stability of polygenic scores across discovery genome-wide association studies. *HGG*
626 *Adv* **3**, 100091 (2022).
627 42. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–
628 209 (2018).
629 43. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**,
630 668–676 (2019).
631 44. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science*
632 **319**, 1100–1104 (2008).
633 45. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–
634 74 (2015).
635 46. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets.
636 *Gigascience* **4**, 7 (2015).
637 47. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits.
638 *Nat. Genet.* **53**, 1097–1103 (2021).
639 48. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association
640 scans. *Bioinformatics* **26**, 2190–2191 (2010).
641 49. Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional
642 genomic categories. *Nat. Commun.* **12**, 1164 (2021).
643 50. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-
644 wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

51. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
52. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).