Leveraging genetic ancestry continuum information to interpolate PRS for admixed populations

Yunfeng Ruan[1], Rohan Bhukar[1,2], Aniruddh Patel[1,2,3], Satoshi Koyama[1,2,3,4], Leland Hull[5,6], Buu Truong[1,2,3,8], Whitney Hornsby[1,2], Haoyu Zhang[7], Nilanjan Chatterjee[9,10], Pradeep Natarajan[1,2,3,6]

Affiliations:

1. Program in Medical and Population, Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA, USA
2. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA
3. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, USA
4. Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
5. Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA
6. Department of Medicine, Harvard Medical School, Boston, MA, USA
7. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA
8. Department of Genetic Epidemiology and Statistical Genetics, Harvard T.H. School of Public Health, Cambridge, MA, US
9. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
10. Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Please address correspondence to:
Pradeep Natarajan, MD MMSc
185 Cambridge Street, CPZN 5.238
Boston, MA 02446
617-726-1843
pnatarajan@mgh.harvard.edu

Funding:
P.N. is supported by NHGRI U01HG011719 and NHLBI R01HL127564. N.C is supported by NHGRI U01HG011719 and R01HG010480. A.P.P. is supported by NHLBI K08HL168238.

Disclosures:
P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Genentech / Roche, and Novartis, personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Bristol Myers Squibb, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co, Esperion Therapeutics, Foresite Capital, Foresite Labs, Genentech / Roche, GV, HeartFlow, Magnet Biomedicine, Merck, Novartis, Novo Nordisk, TenSixteen Bio, and Tourmaline Bio, equity in Bolt, Candela, Mercury, MyOme, Parameter Health, Preciseli, and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. All other

authors report no conflicts.

Abstract

The relatively low representation of admixed populations in both discovery and fine-tuning individual-level datasets limits polygenic risk score (PRS) development and equitable clinical translation for admixed populations. Under the assumption that the most informative PRS weight for a homogeneous sample varies linearly in an ancestry continuum space, we introduce a Genetic **Dis**tance-assisted PRS **Co**mbination Pipeline for **Div**erse Genetic **A**ncestrie**s** (**DiscoDivas**) to interpolate a harmonized PRS for diverse, especially admixed, ancestries, leveraging multiple PRS weights fine-tuned within single-ancestry samples and genetic distance. DiscoDivas treats ancestry as a continuous variable and does not require shifting between different models when calculating PRS for different ancestries. We generated PRS with DiscoDivas and the current conventional method, i.e. fine-tuning multiple GWAS PRS using the matched or similar ancestry samples. DiscoDivas generated a harmonized PRS of the accuracy comparable to or higher than the conventional approach, with the greatest advantage exhibited in admixed individuals.

Introduction/Main

Individuals who are not of European ancestry remain underrepresented in genome-wide association studies (GWAS), which at least partly explains why polygenic risk score (PRS) performance is generally reduced in this population when compared with individuals of European ancestry[1]. Within the constraints of existing data, the current principal solution to increase the PRS accuracy among non-European individuals is to fine-tune a combination of PRS derived from multiple populations or multiple traits with the individual-level data of a training cohort[2–6]. However, PRS accuracy decays as the genetic distance between the fine-tuning and testing samples increases[7]. Relative to the vast diversity across the genetic ancestry continuum, the existing and near-term individual-level datasets that can be used for fine-tuning PRS combinations remains very sparse. Most existing individual-level genotype data are mainly collected from single-ancestry populations and therefore admixed populations are left underrepresented or are largely excluded from analysis [8–11]. Additionally, fine-tuning and testing samples that are labeled as “from the same superpopulation” are often truly genetically heterogeneous [10,12–15], leading to variable accuracy within such samples.

PRS analysis across diverse ancestries may also be limited by inconsistency. The raw PRS distributions of the same model varies by ancestry and therefore the raw PRS values for individuals of different genetic ancestries should not be directly compared without ancestry correction[16–18]. Although prior research[16,18,19] has shown that regressing out the top principal components of ancestry (PCA) from the PRS can unify the PRS distributions of different ancestries (i.e., the mean and standard deviation of corrected PRS sampled from different populations can become very close), the inconsistency is only partially solved. In the application of PRS across diverse ancestries, one would have to use one PRS model for all the individuals, causing inconsistent PRS accuracy, or use several discrete PRS models for different individuals approximating superpopulations also causing inconsistent PRS modelling and accuracy.

Given these issues and the increasing clinical use of PRS[20–22], PRS generation for diverse and admixed genetic ancestries with more consistent accuracy and more unified PRS distributions is critically needed. We devised a method, DiscoDivas, a Genetic ***Dis***tance-assisted PRS ***Co***mbination Pipeline for ***Div***erse Genetic ***A***ncestrie***s***, to generate PRS across the genetic ancestry continuum. This method is based on the recent observation[7] that the PRS accuracy in the testing data decays approximately linearly as the genetic distance between the fine-tuning and samples increases, and that the genetic distance can be approximated by Euclidian distance of PCA based on the global ancestries[7]. Based on this observation, we assumed that the most informative PRS weights for a sample can be linearly interpolated from the currently available PRS weights that are fine-tuned in the ancestries surrounding it in the global ancestry-based PCA space with the interpolation weights based on the Euclidian distance of the PCA. In summary, DiscoDivas calculates PRS for diverse and admixed genetic ancestries whose genetic data may not be sufficiently powerful alone to train a PRS model by linearly interpolating the multiple PRS fine-tuned in ancestries whose genetic data are more available. We evaluated its performance in simulated and empiric data.

Results

Overview of DiscoDivas

DiscoDivas combines PRS fine-tuned in different fine-tuning samples - generally from different single-ancestry populations - to linearly interpolate PRS for individuals of diverse genetic ancestries, treating ancestry as a continuous variable. The rationale for PRS combination is based on the observation that the correlation of the most informative PRS weight for two samples of different ancestry drops as the genetic distance, represented by Euclidean distance of global ancestry-based PCA, increases[7]. Therefore, the best PRS weight for an ancestry representation can be linearly interpolated from other PRS weights fine-tuned in other ancestries with the additional consideration of the genetic distance between the samples (Figure 1).

Under the same principle of interpolating the PRS weight, the best PRS can be interpolated from several PRS calculated using the weight fine-tuned in other ancestries. Since generating individual-specific PRS weights in a testing dataset causes redundant calculation and given the difficulty of normalizing information from different datasets, we combine the PRS instead of the SNP weights. The PRS of individuals in the testing sample is a linear combination of PRS based on the SNP weights fine-tuned in different fine-tuning samples:

$$PRS_i = \sum w_{i,k} PRS_{i,k}$$

where $PRS_{i,k}$ is the PRS of testing individual i calculated using the weight fine-tuned in the fine-tuning sample k; $w_{i,k}$ is the combination coefficient mainly based on the reciprocal of the PCA Euclidean distance between the testing individual and median point of the fine-tuning sample $D_{i,k}$. Note that the input PRS and PCA should be of the same scale: all the individuals are projected to the same PCA space based on a global ancestry reference panel and the PRS input $PRS_{i,k}$ is the raw PRS regressed out the top PCs and then standardized. Additionally, we recommend including all available discovery GWAS for PRS in each PRS model fine-tuned in the single-ancestry sample to maximize the PRS accuracy, as indicated in Figure 1. Nevertheless, DiscoDivas is a flexible framework that allows the different sets of discovery GWAS and fine-tuning method used in different fine-tuning samples.

In addition to the PCA distances, other factors are included in the model. First, since some fine-tuning samples are more correlated than others (e.g., EAS and SAS are more correlated than AFR and EUR), the combination coefficients should be further modified by these correlations, which can also be extracted from the PCA Euclidean distances. Second, since PRS fine-tuned in each of the fine-tuning samples may be of differing qualities (e.g., when the PRS model fine-tuned in different samples are based on GWAS of different sample sizes or populations), the quality of the PRS trained with each of the training data will vary and should be taken into account when combining the PRS. Thus, the combination coefficient $w_{i,\,k}$ in the previous formula is a function of multiple factors:

$$w_{i,\,k} = f\left(\frac{1}{D_{i,k}}, G, r_k\right)$$

where $\frac{1}{D_{i,k}}$ is the reciprocal of PCA Euclidean distance between the individual i and the fine-tuning sample k; G is the matrix of PCA Euclidean distance between fine-tuning samples; r_k is the parameter describing the quality of training fine-tuning samples. A more detailed description of defining $w_{i,\,k}$ is given in the supplementary method section entitled ‘Methodological Details of DiscoDivas’.

The PRS input for DiscoDivas in this study was the multi-GWAS PRS fine-tuned in AFR, EAS, EUR, and SAS fine-tuning samples with the conventional method pipeline as mentioned above (see the following section titled “Overview of multi-population GWAS PRS model” for more detailed information of the input PRS). The interpolation of these four PRS is based on the PCA calculated using the 1000 Genomes reference panel. For most of the PRS analysis conducted in in the present study, the input PRS of DiscoDivas are based on the same set of discovery GWAS and the fine-tuning datasets are sufficiently large to generate a stable result. Therefore, we assumed that all the input PRS can be viewed as of equal quality and their parameter for PRS quality r_k can be viewed as a constant value in the present study.

Overview of multi-population GWAS PRS model

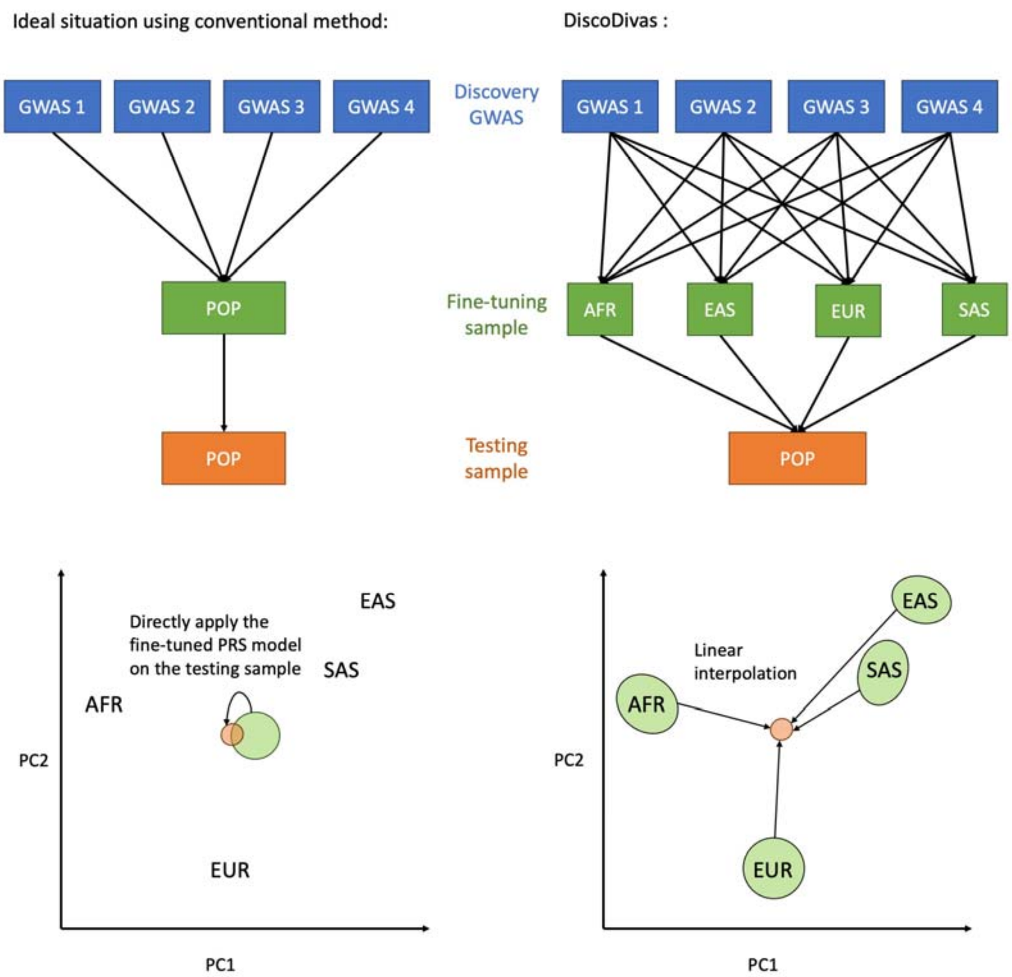


Figure 1: The workflow of comparing DiscoDivas with the existing method. Left: The ideal situation for the existing method is to fine-tune a PRS model that contains multiple GWAS with matched fine-tuning data, which is not currently available for many under-represented populations. Right: DiscoDivas first fine-tunes the PRS in the available ancestries, which are currently AFR, EAS, EUR, and SAS, and interpolates PRS for diverse ancestry groups based on these fine-tuned PRS. In this plot, POP refers to any ancestry for which the PRS is to be calculated.

A common approach for constructing PRS is to include as much genome-wide association study (GWAS) summary statistic data as possible in the discovery data[5,23,24]. The GWAS data is typically then processed by PRS methods that will adjust the SNP effect size using a set of hyper-parameters. Individual-level data of an independent fine-tuning sample is used to fine-tune the hyper-parameters across PRS methods and the combination of the fine-tuned PRS. The resulting PRS is expected to perform the best in samples of matched ancestry with the fine-tuning sample.

The current approach, as shown in the left panel of Figure 1, is to use the multi-GWAS PRS fine-tuned in the matched sample or the closest approximation when the matched sample is unavailable. The pipeline of adjusting SNP effect sizes and combining information from different GWAS varies widely. Without loss of generality, we built the following pipeline as a representation of the current conventional method: we first adjusted the SNP effect size of each of the summary statistical GWAS datasets by a Bayesian method and then chose the most predictive PRS from all the PRS generated under different hyper-parameters. For simulated GWAS data, we used PRS-CS[25] to adjust the SNP effect size and LDpred2[26] for real GWAS. Then we used the fine-tuning data to first select the most predictive PRS based on each GWAS and then to train the linear combination of the most predictive single-GWAS PRS with a linear regression model. The final PRS model generated from each of the fine-tuning datasets is a linear combination of PRS. For the empiric data set, the PRS were fine-tuned controlling for the following covariates: top 20 PCA, sex, and age. We used AFR, EAS, EUR, SAS, AMR, and admixed samples to fine-tune the PRS. A more detailed description of generating PRS weight from the one fine-tuning data was given in Supplementary method section entitled ‘Methodological Details of PRS Construction Using a Single Fine-tuning Dataset’.

Simulated data results

Summary-level GWAS used as discovery data were generated based on simulated genotype of AFR, EAS, EUR, and SAS population based on 1000 Genomes[27] reference as described in the previous publication provided by Zhang et al[6]. Fine-tuning and testing samples were simulated based on UKBB genotype data. From each ancestry group of AFR, EAS, EUR, SAS, and other (OTH) for admixed individuals whose PCA information was not matched with any of the five ancestries by 1000 Genome reference definition, 1.3k individuals were used as the training fine-tuning datasets (See supplementary method section entitled ‘Generating data for simulation analysis’). The phenotype of discovery, fine-tuning, and testing data were generated using the same pipeline and parameters: the phenotypes of 100, 300, 1,000, or 10,000 causal SNPs and heritability = 0.6 were simulated. Scenarios of shared causal SNP with effect size constant across different ancestries and shared causal SNP with effect size varying across population are both simulated. We used up to 100,000 simulated individuals from AFR, EAS, EUR, and SAS to generate the discovery summary statistic GWAS dataset with PLINK2[28] and left the remaining samples out for other downstream analyses.

We primarily focused on the PRS performance in the admixed testing cohort. DiscoDivas, which is based on PRS fine-tuned in AFR, EAS, EUR, and SAS, was compared with the conventional PRS fine-tuned in the matched admixed fine-tuning sample in scenarios of different causal SNP numbers, different discovery GWAS sample sizes, and different causal SNP distribution across ancestry (See Figure 2)

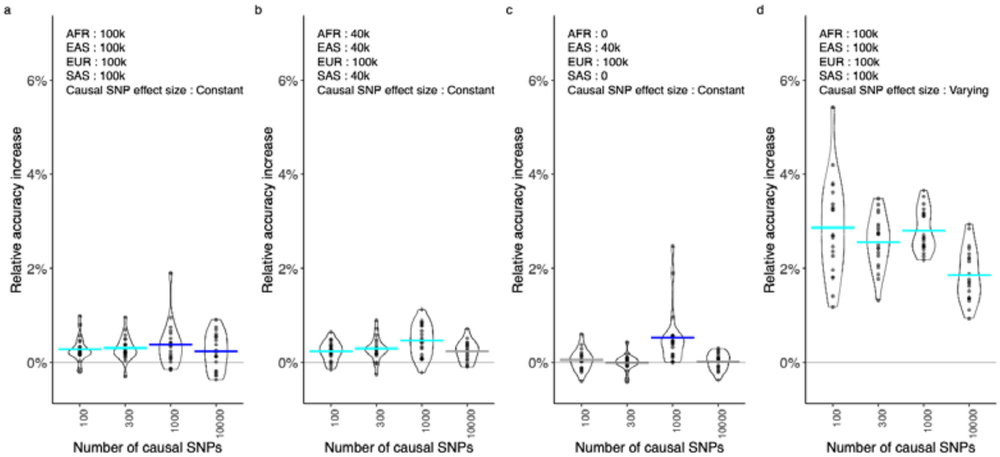


Figure 2 Relative R^2 increase of DiscoDivas over the conventional PRS fine-tuned in a matched sample when tested in admixed individuals. The x-axis shows the simulated number of causal SNPs. The horizontal bar shows the mean relative R^2 increase and the color of the horizontal bar indicates the *p*-value of the paired t-test of DiscoDivas PRS R^2 and conventional PRS R^2, with cyan being *p*-value<0.0005, dark blue being *p*-value<0.05 and grey being *p*-value>0.05. In panels a, b, and c, the causal SNP effect sizes are constant across different populations. The annotation texts on the top of each panel shows the sample size of discovery GWAS of different populations and the distribution of causal SNP effect sizes.

Although the comparison between DiscoDivas and the conventional method of fine-tuning PRS with matched ancestry sample in a single test iteration usually showed no statistical significance due to the small numeric differences, the paired t-test of DiscoDivas R^2 and the conventional PRS R^2 over the 20 iterations better clarified significant differences. When effect sizes of causal SNPs were held constant across different ancestries (Figure 2 panel a, b, and c), the PRS generated by DiscoDivas had comparable accuracy with the PRS fine-tuned using matched data. We noticed that when the sample size of non-European discovery GWAS dropped and the dataset was relatively more Eurocentric, the advantage of DiscoDivas became less statistically significant. In Figure 2 panel d, we compared DiscoDivas and the conventional PRS method of fine-tuning the PRS with matched ancestry in the scenario where causal SNPs were shared across all populations, but the effect sizes varied linearly in the PCA space. The advantage of DiscoDivas over conventional PRS method was more obvious in this scenario than when the effect sizes were constant across populations (Figure 2 panel a and d), presumably because personalized PRS combination with DiscoDivas better captured the changing effect sizes for the admixed testing sample. In all the scenarios tested, the advantage of DiscoDivas was least statistically significant when the number of causal SNPs was 10,000 but still significant when the number of causal SNPs was 1,000. Notably, the accuracy of both DiscoDivas and the conventional PRS method was the lowest when the number of causal SNPs was 10,000 (Supplementary Figure 1), indicating that the difference of the two PRS methods became less obvious when the input data became increasingly underpowered.

When predicting the individuals that are usually classified as single ancestries, i.e. AFR, EAS, EUR, and SAS, DiscoDivas showed no statistically significant difference or a slight advantage over the conventional PRS method (Supplementary Figure 2). When predicting AMR individuals, we used admixed fine-tuning data (OTH) to fine-tune the conventional PRS due to the small sample size of the AMR dataset. The PRS performance when testing in the AMR dataset was similar as in admixed data but the statistical significance was weaker, potentially due to the small sample size and the high heterogeneity of the AMR dataset. In general, DisocDivas showed its clearest advantage over the conventional method of fine-tuning PRS with matched PRS when the testing data and the fine-tuning data for the conventional method were of different ancestries.

Biobank data results

We downloaded publicly available summary statistical data of body-mass index (BMI), high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), total cholesterol (TC), triglycerides (TG), systolic blood pressure (SBP), diastolic blood pressure (DBP), coronary artery disease (CAD), and diabetes mellitus (DM2) and adjusted the SNP effect size using LDpred2 as described previously[5].

For the quantitative traits, we used the fine-tuning samples of AFR, EAS, EUR, SAS, and admixed (OTH) ancestry to fine-tune the model. The remaining UKBB samples were used as the testing data. The results for empiric quantitative trait data were highly aligned with the simulation results (Figure 4): DiscoDivas showed a robust advantage over the conventional PRS method of fine-tuning PRS with matched or similar ancestry samples when compared across the 7 traits in the admixed testing dataset. When predicting AFR, EAS, EUR, and SAS, DiscoDivas and the conventional PRS method had similar performance. The results of both methods in AMR testing dataset had large deviations due to the small sample size and greater genetic heterogeneity of the AMR data.

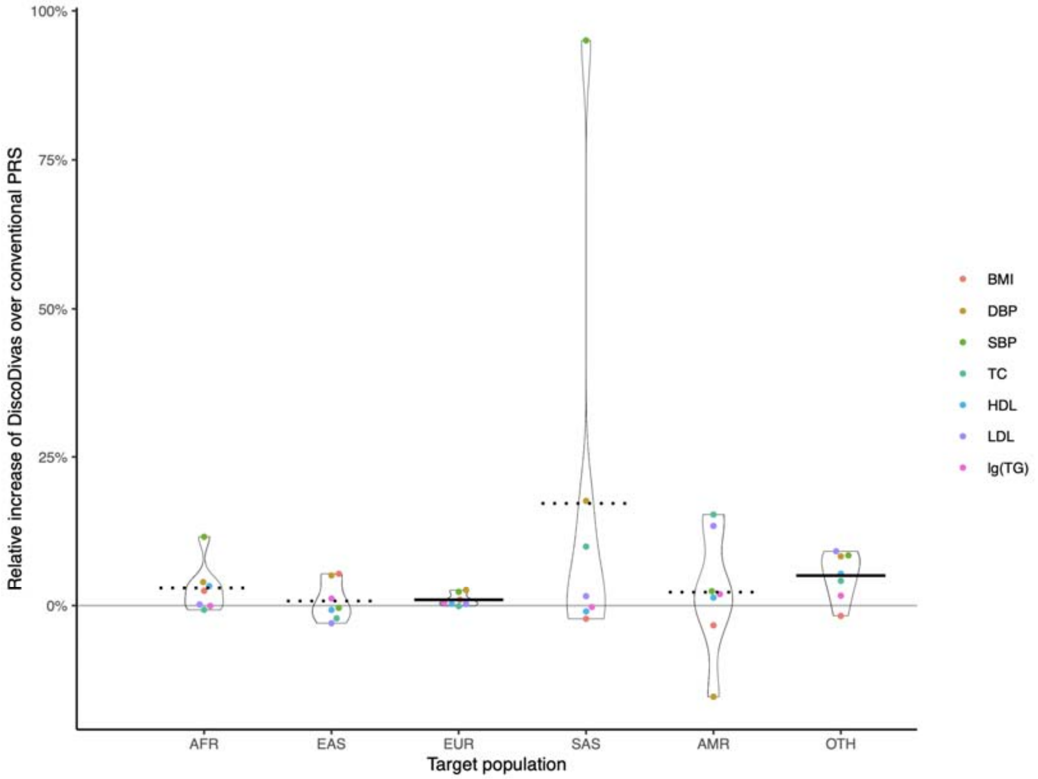
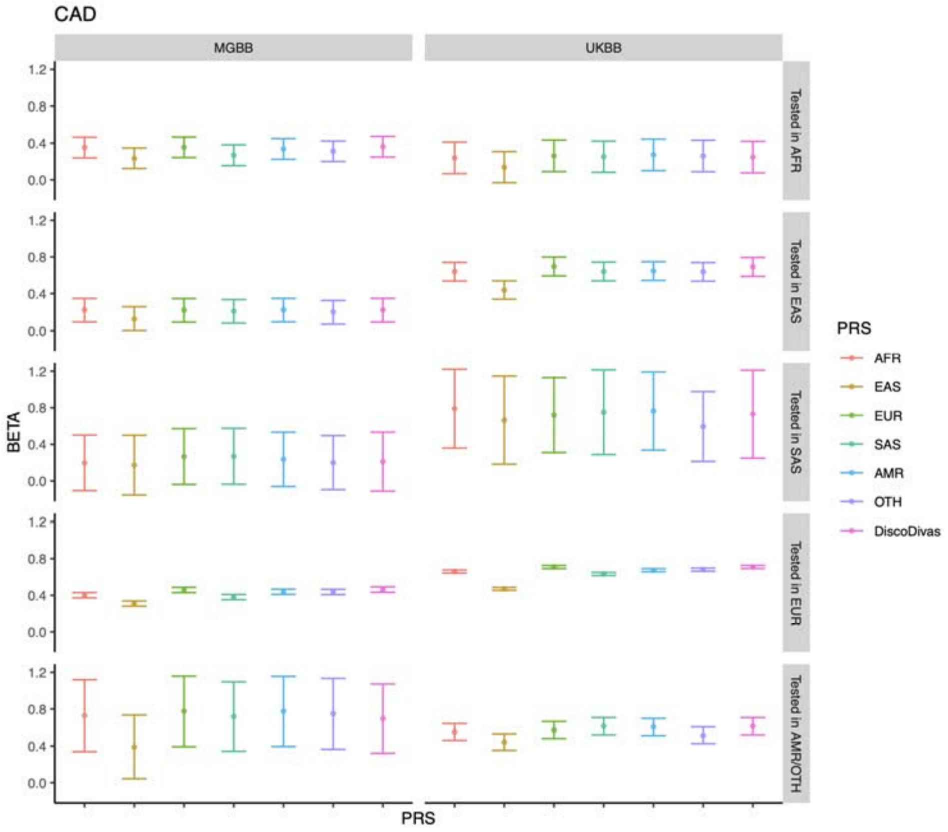


Figure 3 Relative R^2 increase of DiscoDivas over the conventional PRS fine-tuned in a matched sample. The x-axis shows the population in which the PRS was tested. We used OTH as the fine-tuning

dataset for the test in both OTH and AMR due to the absence of matched AMR training data. The horizontal bar shows the mean of relative increase, and the line-type of the bar indicates the *p*-value of paired t-test of DiscoDivas PRS R^2 and conventional PRS R^2, with the solid bar being *p*-value <0.05 and dotted bar being *p*-value>0.05.

For the binary traits coronary artery disease (CAD) and type 2 diabetes (DM2) (Figure 4), we used the AFR, EAS, EUR, SAS, AMR, and OTH (i.e., unclassified) samples from AoU as the fine-tuning data and tested in AFR, EAS, EUR, SAS, and OTH individuals in UKBB and AFR, EAS, EUR, SAS, and AMR individuals in MGBB. The DiscoDivas PRS were based on the PRS fine-tuned in AFR, EAS, EUR, and SAS and used the default assumption that the PRS fine-tuned from all the samples were of similar quality even though the sample sizes of both discovery GWAS and the fine-tuning samples were not balanced across different ancestries. AMR in UKBB was excluded because of the small sample size (N=669).

The PRS fine-tuned in different single samples and the DiscoDivas PRS had similar performances. It also appeared that some of the fine-tuning sample could be underpowered: generally, we expect the PRS fine-tuned in the matched sample to perform the best in the testing samples, but PRS fine-tuned in larger fine-tuning data performed better than PRS fine-tuned in smaller fine-tuning data in general. For example, the PRS fine-tuned in EAS AoU data performed worse than other PRS in both MGBB and UKBB EAS data and had low accuracy in other testing data as well; the CAD PRS fine-tuned in EUR performed better than all the other PRS in all the testing data and the effective sample size of EUR CAD fine-tuning data was much larger than all the other fine-tuning data.



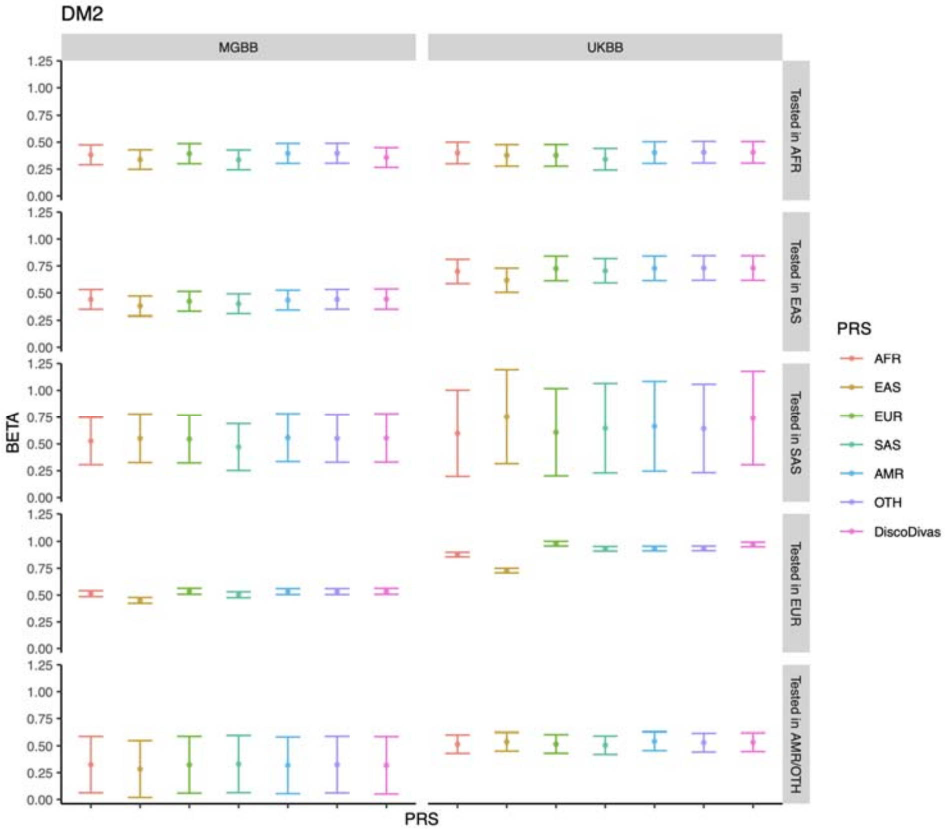


Figure 4 PRS performance for coronary artery disease (CAD) and type 2 diabetes (DM2) tested in UKBB and MGBB. The plot shows OR per SD with the error bar showing 95% CI. The sub-panels show that population of the testing sample and the different colors show the method for generating the PRS, either fine-tuning in a single sample or combining the PRS using DiscoDivas.

Discussion

We propose a new method, DiscoDivas, to interpolate the PRS for diverse, especially admixed, ancestries with a generalized framework that does not requiring binning into discrete ancestries. Our results shows that the accuracy of DiscoDivas was comparable to or greater than the conventional method, i.e. fine-tuning using the matched population sample when available. In addition, when generating PRS for a wide range of ancestries, DiscoDivas did not require shifting from several sets of PRS weights fine-tuned in discrete samples while remaining matched with the ancestry information. Our method provides a new solution to generate PRS for underrepresented, generally admixed, populations and as well as generate a harmonized PRS model across different ancestries.

The performance of our method depends on the quality of both the discovery GWAS data and the fine-tuning data. As shown in the simulation test, discovery GWAS datasets that represent diverse ancestries with sufficient sample size will increase the accuracy of interpolated PRS generated by DiscoDivas. On the contrary, Eurocentric and underpowered discovery GWAS datasets would limit the advantage of DiscoDivas over the conventional PRS method. This might partly explain the limited advantage of DiscoDivas when predicting binary traits: the discovery GWAS datasets were highly Eurocentric and the GWAS, especially the non-European cohorts, could be more underpowered than quantitative trait GWAS. Furthermore, the PRS fine-tuned in fine-tuning datasets of insufficient sample size will be overfitted and cannot be used to fairly evaluate the performance of either the conventional PRS method or DiscoDivas. We aimed to address this issue by only using traits that 1) had effective sample sizes larger than 200 in all the fine-tuning samples, and 2) had high-quality phenotyping data in both the fine-tuning datasets and the testing datasets, However, Asian populations were largely under-represented in the current public biobanks: the effective sample size of many binary traits in EAS or SAS can be as small as <200 even in AoU, the most diverse and large-scale largely publicly-available biobank we had access to. This limited our choice for binary traits to only CAD and DM2. One additional limitation of our method is that DisoDivas does not consider the local ancestry information, which improve PRS predictions in various research[24,29,30], especially PRS prediction of newly admixed populations[31].

Our research underscores the notion that non-European populations, both admixed and singe-ancestry populations, remain largely under-represented in the existing genetic data. Furthermore, some potential extensions of our method will not become possible until we collect more diverse and larger datasets. First, our method has not been designed nor tested for extrapolating data, e.g. generating PRS for continental African samples based on African American, European, and Asian samples. Even though it is mathematically plausible to alter our method to extrapolate the PRS, we lack data such as continental African samples to test the method. Secondly, we currently only consider the assumption that the most informative genome-wide PRS weight shifts linearly in the PCA space. Although more complicated PRS interpolation, e.g. interpolation guided by local ancestry information [24,29,30], pathway-specific[32,33] and annotation-guided[34] PRS weights and polynomial interpolation[35,36], can possibly further improve the PRS accuracy, training such complicated models would require collecting much larger and more diverse datasets than the existing data. Finally, additional biological insights could be revealed by interpolating PRS if genetic data of all the involved diverse ancestries are of sufficient power. In this case, the differences between interpolated PRS and the PRS trained using the matched ancestry would indicate the population- or sample- specific factors absent in the interpolation model, e.g. population-specific genetic variance[37], complicated population stratification involving cofounding factors[38,39], sample/ancestry-specific modifiers like local adaptation[38], gene x environment interactions[40] or other factors that contribute to the genetic variant frequency or effect size in these samples/ ancestries.

In conclusion, our method provides a new option to treat the ancestry information as a continuous variable and interpolate a harmonized PRS for diverse ancestries. Notably, although our method was developed primarily to calculate PRS when the matched fine-tuning datasets were unavailable, our research showed that successfully interpolating PRS required sufficient input data and highlighted the need to collect genetic data for underrepresented populations. We believe that more diverse and larger data collected in the coming future will enable the development of new methods of interpolating PRS and the elucidation of the genetic basis of complex traits.

Methods

Calculation of 1000 Genomes-based PCA and Euclidean distance

We use 1000 Genomes as the reference panel for PCA calculation. The PCA should be based on SNPs that are constantly included in as many samples as possible to enable the use of wide-ranging discovery GWAS and fine-tuning datasets. We started with the Hapmap3 SNPs for this set of SNPs, which has been widely used as a subset of SNPs that approximates the feature of genome-wide common SNPs in many recent studies that involve multi-ancestry prediction [6,25,26,41]. We further filtered for the SNPs likely to be frequently genotyped or imputed with relatively high quality by most samples based on the 1000 Genome data: Hapmap3 SNPs were first extracted from the five super-populations, Africans (AFR), Admixed Americans (AMR), East Asians (EAS), Europeans (EUR) and South Asians (SAS) of the 1000 Genomes. Secondly, SNPs described as the following were excluded: 1) of minor allele frequency lower than 1% in any of the super-population, 2) of minor allele frequency lower than 5% in the combined 1000 Genomes data, and 3) in the long-range LD region (25Mb – 35Mb by hg19 assembly on chromosome 6 and 7Mb – 13Mb on chromosome 8). To calculating the PCA loading, the QC'ed SNPs of the five super-populations were merged then pruned using the PLINK2 function "indep-pairwise" with the parameter "200 100 0.1" - namely the pruning was performed using window size = 200kb, step size = 100, and phased-hardcall-r^2= 0.1. The principal components and the SNP loadings are calculated using PLINK2 function "pca" with the parameter "allele-wts" based on the pruned SNPs.

Based on the protocol suggested on the PLINK2 website (https://www.cog-genomics.org/plink/2.0/score#pca_project), we projected samples for fine-tuning and PRS testing into the PCA space as describe above by calculation the linear score, i.e. the sum of alternative alleles weighted by the SNP effect size, using the PLINK2 function "score" with the SNP loadings as effect size. The original online protocol suggested linear score should be first scaled to standard variation and then rescaled by multiplying the square root of eigenvalue. However, the actual standard deviation of a sample in the same PCA space varies with the homogeneity and the ancestry of the sample. Forcing the PCA of all the samples to have the same standard deviation will cause inconsistent scaling when the samples can be of different ancestries. Therefore, we directly calculated the PCA from sum basic linear score based on the SNP loadings as generated above without any further scaling. The PCA in this study was the sum basic linear score calculated using the PLINK2 function "score" with the parameter "cols=+scoresums'". For large samples whose genotype data were divided into per-chromosome files, the same commands were used to calculate per-chromosome score and the genome-wide score was the sum of the score of all the autosomes.

In DiscoDivas' default setting, the genetic distance between two individuals is defined as the Euclidian distance between the PCA of the two individuals. When the genetic distance calculation involves a sample, we use the median point to present the whole sample.

We also explored the relationship between number of PCs included in the calculation and the Euclidean distance calculated (Supplementary Figure 9) and the distance calculated converged when the number of PCs was larger than 6 in our tests. In our analysis we use the top 10 PCs to calculate the PCs.

Genetic ancestry reference

We noticed that the protocol of generating top PCs for ancestry references varied in previous publications. In our pilot test (see supplementary resuts section entitled 'Pilot test of generating PCA based on less QC'ed SNPs'), we compared the ancestry reference based on Hapmap3 SNP without any QC and found the result to be highly correlated. We used the same set of PCs based on QC'ed SNP as described in section 'Calculation of 1000 Genomes-based PCA and Euclidean distance' for both genetic ancestry reference and Euclidean distance calculation for data consistency.

Random forest model of 100 trees was trained based on the 1000 Genome data. The out-of-bag estimate of error rate stabilize at the level of 0.28% after the number of PCs passed 5. We used the model using the top 6 PCs to infer the genetic ancestry of UK Biobank individuals and the Mass General Brigham Biobank individuals. The genetic ancestry of an individual was assigned to any of the five ancestries represented in the 1000 Genomes reference data, i.e. AFR, AMR, EAS, EUR and SAS, if the highest probability of an individual belonging to that ancestry passed a threshold. If none of the ancestries had a probability above the threshold, the individuals were assigned as other (OTH), which indicated that the individual was of admixed ancestries. With the consideration of the sample size and confirmed by visual inspection, the threshold of probability for UK Biobank and the Mass General Brigham Biobank was 0.9 and 0.8 respectively.

Data

UK Biobank

The UK Biobank (UKBB) is a volunteer sample of approximately 500,000 adults aged 40-69 upon enrollment living in the United Kingdom recruited since 2006[42]. UKBB data used in this research were first QC'ed with the following process: Remove the individuals meeting the criteria that indicate low genotype quality or contamination: 1) have missing genotype rate larger than 0.02; 2) have genotype-phenotype sex discordance; 3) are identified as having excess heterozygosity and missing rates; 4) are identified as putatively carrying sex chromosome configurations that are not either XX or XY; 5) appeared to have unreasonably large numbers of relatives. From the remaining samples, individuals from a group of multiple individuals that are closer than 3rd-degree relatives were retained. 415,402 individuals were left after the QC. 390,037 were self-identified as EUR, 7,039 AFR, 8,652 non-Chinese Asian (ASN), 1430 Chinese (CHN) and 6572 unknown or not answered, and 1672 as admixed (MIX). The genetic ancestry referred from PC was largely correlated with the self-reported race, with 385,038 EUR, 7,450 AFR, 8,298 SAS, 2,163 EAS, 669 AMR and 11,784 other (OTH) or admixed.

In the PRS test, UKBB samples were grouped by their genetic ancestry (see section 'Genetic ancestry reference'). The fine-tuning datasets for the single-ancestry populations (AFR, EAS, EUR and SAS) were based on 1.3k randomly selected samples whose self-report ancestry matched with their genetic ancestry and the probability of random forest = 1. The fine-tuning dataset for admixed ancestry (OTH) is 1.3k randomly selected samples of individuals of OTH genetic ancestry (see Supplementary Figure 3). AMR didn't have its corresponding fine-tuning dataset due to its small sample size and we used OTH fine-tuning datasets as a proxy since the two genetic ancestries had similar PCA. The remaining individuals of UKBB were used as testing data.

The quantitative trait of the UKBB samples was the measurement collected after the participants enrolled. The lipid trait measurement was adjusted for cholesterol-lowering medication by dividing TC by 0.8 and LDL by 0.7 as before[43]. Cases of coronary artery diseases (CAD) are defined using the definition described previously[24]; Cases of diabetes are defined as ever report the following code: E10X, E11X, E12X, E13X, and E14X where X can be any integer between 0 to 9 in the ICD10 diagnosis code.

UKBB participants provided consent in accordance with the primary IRB protocol, and the Massachusetts General Hospital IRB approved the present secondary data analysis of the UKBB data under UKBB application 7089.

Mass General Brigham Biobank

The Mass General Brigham Biobank (MGBB) is a volunteer sample of approximately 142,000 participants receiving medical care in the Mass General Brigham health care system recruited starting 2010. 53,306 MGBB participants underwent genotyping via Illumina Global Screening Array (Illumina, CA). MGBB genotype data was quality controlled, imputed and assigned one of the populations AFR, AMR, EAS, EUR, SAS using K-nearest neighbor model as described previously[44]. The phenotype data of CAD and diabetes are drawn from PheCodes based on International Classification of Diseases codes, Nineth (ICD9)110 and Tenth (ICD10) revisions, from the EHR as described previously[32]. MGBB participants provided consent in accordance with the primary IRB protocol, and the Massachusetts General Hospital IRB approved the present secondary data analysis.

All of Us Research Program

The *All of Us* (AoU) Research Program is a volunteer sample of more than one million United States residents recruited starting 2016. AoU aims to engage communities previously underrepresented in biomedical research in the United States and beyond[45]. In the present analysis, genetic data from the v7 245,394 participants who were genotyped using short read whole genome sequencing (srWGS) data. Hapmap3 SNPs were extracted for the callset with the threshold of (AF) > 1% or population-specific allele count (AC) > 100. Related individuals were pruned according to the information provided by AoU. Due to the inclusive data collection, we didn't excluded individuals whose self-report gender were different with their assigned sex at birth and used the combination of self-report gender and assigned sex as one of the covariates. The predicted ancestry information was provided by AoU[46]. The phenotypes were defined as described in previous research by Buu *et al*[47].

Simulated data

The simulated GWAS summary statistics were based on simulated genotype data based on 1000 Genomes reference[6]. Only Hapmap3 SNPs were included in the simulation. Causal SNPs were randomly selected from the Hapmap3 SNPs and simulated per allele effect size following normal distribution. The ladder of causal

SNP number was 100, 1000, 3000, 10000 and the heritability in each of the population was 0.6. The causal SNP effect size was simulated as either constant across populations or varying linearly in the PCA space.

The phenotype is the sum of genetic burden and non-genetic factor:

$$phenotype_i = \sum \beta_j x_{j,i} + E_i$$

where the $phenotype_i$ and E_i were the phenotype and non-genetic factor of individual i; β_j was the effect size of causal SNP j, and $x_{j,i}$ was the number of risk alleles of individual i in SNP j.

We used the PLINK2[28] to calculate the genetic burden based on the simulated causal SNPs and effect size and used R to simulate the non-genetic factors, scale the genetic burden and non-genetic factor, and generate a phenotype of heritability set to be 0.6. We used up to 100k individuals per population to generate the summary statistical GWAS as the discovery data for the PRS test. The rest simulated data were left out for the fine-tuning and testing datasets. The summary statistics GWAS were generated based on the simulated genotype data and phenotype data using the '--glm' function of PLINK2.

In addition to the completely simulated data, we generated more realistic fine-tuning and testing datasets of a wider ancestry range by using the QC'ed genotype data from UKBB described in the section 'Biobank data.' We simulated the genetic burden, non-genetic factor, and phenotype based on the real-life UKBB genotype data with the same pipeline and parameters. A more detailed description of simulating the data were given in section 'Generating data for simulation analysis' in the supplementary text.

Acknowledgement

We thank the participants and staff of the UK Biobank (application 7089), Mass General Brigham Biobank, and *All of Us* Research Program. We thank Ying Wang, Paul O'Reilly, Raymond Walters for helpful discussion and advice.

Contributions

Y.R. and P.N. designed the project; Y.R. developed the statistical methods and programmed the code for DiscoDivas. A.P curated the summary statistical GWAS. R.B, S.K, L.H, B.T, and W.H participated in application for the access to the Biobank data and the data curation. Y.R and R.B performed the data analysis. R.B, B.T, H.Z, and N.C contributed to the method development. Y.R. and P.N wrote the manuscript. A.P. and N.C. provided critical revision for the manuscript. All the authors reviewed and approved the final version of the manuscript.

Data Availability

The access to biobank data (UK Biobank, Mass General Brigham Biobank, and All of

US Research Program) were gained upon application. The simulated genotype data based on 1000 Genomes were downloaded from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP ;The resource of summary statistics GWAS data used to generate PRS were given in the supplementary file. The 1000 Genome raw reference genotype data were downloaded from https://cncr.nl/research/magma/.

Code Availability

The scripts of running DiscoDivas and other supporting files can be found at https://github.com/YunfengRuan/DiscoDivas; PRS-CS: https://github.com/getian107/PRScs; LDpred2: https://privefl.github.io/bigsnpr/articles/LDpred2; PLINK2: https://www.cog-genomics.org/plink/2.0/.

Reference

1. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019).
2. Miao, J. *et al.* Quantifying portable genetic effects and improving cross-ancestry genetic prediction with GWAS summary statistics. *Nat Commun* **14**, 832 (2023).
3. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nat Genet* **54**, 573–580 (2022).
4. Jin, J. *et al.* MUSSEL: Enhanced Bayesian Polygenic Risk Prediction Leveraging Information across Multiple Ancestry Groups. *bioRxiv* 2023.04.12.536510 (2023) doi:10.1101/2023.04.12.536510.
5. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat Med* **29**, 1793–1803 (2023).
6. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nat Genet* **55**, 1757–1768 (2023).
7. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
8. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L. & Bonham, V. L. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Aff* **37**, 780–785 (2018).
9. The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
10. Fatumo, S. *et al.* Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. *Genome Medicine* vol. 15 Preprint at https://doi.org/10.1186/s13073-023-01245-9 (2023).
11. Dokuru, D. R., Horwitz, T. B., Freis, S. M., Stallings, M. C. & Ehringer, M. A. South Asia: The Missing Diverse in Diversity. *Behav Genet* **54**, 51–62 (2024).
12. Stefflova, K. *et al.* Dissecting the Within-Africa ancestry of populations of African descent in the Americas. *PLoS One* **6**, (2011).
13. Harlemon, M. *et al.* A custom genotyping array reveals population-level heterogeneity for the genetic risks of prostate cancer and other cancers in Africa. *Cancer Res* **80**, 2956–2966 (2020).

14. Anagnostou, P. *et al.* Inter-individual genomic heterogeneity within European population isolates. *PLoS One* **14**, (2019).
15. Pan, Z. & Xu, S. Population genomics of East Asian ethnic groups. *Hereditas* vol. 157 Preprint at https://doi.org/10.1186/s41065-020-00162-w (2020).
16. Khera, A. V *et al.* Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation* **139**, 1593–1602 (2019).
17. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun* **10**, 3328 (2019).
18. Wang, M. *et al.* Validation of a Genome-Wide Polygenic Score for Coronary Artery Disease in South Asians. *J Am Coll Cardiol* **76**, 703–714 (2020).
19. 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).
20. O'Sullivan, J. W. *et al.* Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* vol. 146 E93–E118 Preprint at https://doi.org/10.1161/CIR.0000000000001077 (2022).
21. Lewis, C. M. & Vassos, E. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine* vol. 12 Preprint at https://doi.org/10.1186/s13073-020-00742-5 (2020).
22. Xiang, R. *et al.* Recent advances in polygenic scores: translation, equitability, methods and FAIR tools. *Genome Medicine* vol. 16 Preprint at https://doi.org/10.1186/s13073-024-01304-9 (2024).
23. Truong, B. *et al.* Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genomics* **4**, (2024).
24. Wang, Y. *et al.* Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genomics* **3**, (2023).
25. 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).
26. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
27. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
28. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, (2015).
29. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet* **53**, 195–204 (2021).
30. Sun, Q. *et al.* Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nat Commun* **15**, (2024).
31. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat Commun* **11**, (2020).
32. Xu, Y. *et al.* Effect of Pathway-Specific Polygenic Risk Scores for Alzheimer's Disease (AD) on Rate of Change in Cognitive Function and AD-Related Biomarkers Among Asymptomatic Individuals. *Journal of Alzheimer's Disease* **94**, 1587–1605 (2023).
33. Tubbs, J. D. *et al.* Pathway-Specific Polygenic Scores Improve Cross-Ancestry Prediction of Psychosis and Clinical Outcomes. Preprint at https://doi.org/10.1101/2023.09.01.23294957 (2023).
34. Miao, J. *et al.* Quantifying portable genetic effects and improving cross-ancestry genetic prediction with GWAS summary statistics. *Nat Commun* **14**,

(2023).

35. Kumar, R., Bhattacharya, S. & Murmu, G. Exploring Optimality of Piecewise Polynomial Interpolation Functions for Lung Field Modeling in 2D Chest X-Ray Images. *Front Phys* **9**, (2021).

36. Womersley, R. S. & Sloan, I. H. *How Good Can Polynomial Interpolation on the Sphere Be? Advances in Computational Mathematics* vol. 14 (2001).

37. Choudhury, A. *et al.* Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC Genomics* **15**, (2014).

38. Rees, J. S., Castellano, S. & Andrés, A. M. The Genomics of Human Local Adaptation. *Trends in Genetics* vol. 36 415–428 Preprint at https://doi.org/10.1016/j.tig.2020.03.006 (2020).

39. Mas-Sandoval, A., Mathieson, S. & Fumagalli, M. The genomic footprint of social stratification in admixing American populations. **12**, 84429 (2023).

40. Patel, R. A. *et al.* Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. *Am J Hum Genet* **109**, 1286–1297 (2022).

41. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

42. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

43. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

44. Koyama, S. *et al.* Decoding Genetics, Ancestry, and Geospatial Context for Precision Health. *medRxiv* (2023) doi:10.1101/2023.10.24.23297096.

45. Kathiresan, N. *et al.* Representation of Race and Ethnicity in the Contemporary US Health Cohort All of Us Research Program. *JAMA Cardiol* **8**, 859–864 (2023).

46. Bick, A. G. *et al.* Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).

47. Truong, B. *et al.* Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *medRxiv* (2023) doi:10.1101/2023.02.21.23286110.