nature communications

Article

https://doi.org/10.1038/s41467-024-54727-8

Mapping the relative accuracy of cross-ancestry prediction

Received: 5 December 2023

Accepted: 20 November 2024

Published online: 02 December 2024

Check for updates

Alexa S. Lupi [1,2], Ana I. Vazquez [1,2] & Gustavo de los Campos [1,2,3]

The overwhelming majority of participants in genome-wide association studies (GWAS) have European (EUR) ancestry, and polygenic scores (PGS) derived from EURs often perform poorly in non-EURs. Previous studies suggest that between-ancestry differences in allele frequencies and linkage disequilibrium are significant contributors to the poor portability of PGS in cross-ancestry prediction. We hypothesize that the portability of (local) PGS varies significantly over the genome. Therefore, we develop a method, MC-ANOVA, to estimate the loss of accuracy in cross-ancestry prediction attributable to allele frequency and linkage disequilibrium differences between ancestries. Using data from the UK Biobank we develop PGS relative accuracy (RA) maps quantifying the local portability of EUR-derived PGS in non-EUR ancestries. We report substantial variability in RA along the genome, suggesting that even in ancestries with low overall RA of EUR-derived effects (e.g., African), there are regions with high RA. We substantiate our findings using six complex traits, which show that EUR-derived effects from regions where MC-ANOVA predicts high RA also have high empirical RA in real PGS. We provide software implementing MC-ANOVA and RA maps for several non-EUR ancestries. These maps can be used to interpret similarities and differences in GWAS results between groups and to improve cross-ancestry prediction.

In the last fifteen years, thousands of genome-wide association studies (GWAS) have been published[1]. Increasingly, single nucleotide polymorphisms (SNPs) that these studies reported to be associated with specific phenotypes or disease outcomes are used to build polygenic scores (PGS). The availability of biobank-sized data has led to unprecedented improvements in PGS prediction accuracy[2,3]. However, the overwhelming majority of participants in GWAS (approximately 80%) are of European (EUR) descent[4], leading to issues with generalizability and exacerbating existing health disparities. Consistently, studies across various traits/diseases and target ancestry groups have shown that PGS derived with data from EURs have poor predictive performance when used to predict among individuals of non-EUR ancestry (African [AF] in particular)[4–15].

Several factors can contribute to the poor portability of PGS across ancestries. At causal loci, unaccounted gene-by-gene (G×G) and genetic-by-environment (G×E) interactions can lead to ancestry differences in the additive effects of causal alleles. Furthermore, differences across ancestry groups in allele frequencies and linkage disequilibrium (LD) patterns can lead to heterogeneity in marker effects even for loci without such heterogeneity at causal loci[16]. The relative contribution of G×G, G×E, allele frequency differences, and LD differences to the poor portability of PGS remains largely unknown and can be expected to vary across traits and ancestries. However, several studies suggest that allele frequency and LD differences between ancestries are significant factors contributing to the poor portability of PGS, possibly explaining up to 75% of the empirical loss of accuracy (LOA) in cross-ancestry prediction[7,8,17,18].

[1]Department of Epidemiology and Biostatistics, Michigan State University (MSU), East Lansing, MI 48824, USA. [2]Institute for Quantitative Health Science and Engineering, Systems Biology, MSU, East Lansing, MI 48824, USA. [3]Department of Statistics and Probability, MSU, East Lansing, MI 48824, USA.
e-mail: lupialex@msu.edu; gustavoc@msu.edu

Many studies have investigated the portability of PGS across ancestries from a whole-genome perspective[7,8]. However, no previous study has quantified how the portability of local PGS varies over the genome and how this information can be used to identify genomic regions of low and high relative accuracy (RA, the ratio of cross-ancestry to within-ancestry variance explained and functions thereof) between ancestral groups. We hypothesize that the degree of allele frequency and LD differences between ancestries (and therefore the local portability and RA of PGS) varies along the genome. Therefore, we developed an algorithm, Monte Carlo ANOVA (MC-ANOVA), to map the RA of local linear functions of SNP genotypes.

In this work, we apply the MC-ANOVA method to data from the UK Biobank[19] and the ARIC (Arteriosclerosis in Risk Communities) study[20] to generate portability maps of the local RA of PGS between EUR and non-EUR ancestry groups. Using PGS for six quantitative traits (height, high-density lipoprotein [HDL], low-density lipoprotein [LDL], serum urate, body mass index [BMI], and serum glucose), we show that the portability maps we develop are predictive of the empirical local RA of EUR-derived PGS for the prediction of the same traits in African (AF), Caribbean (CR), East Asian (EA), and South Asian (SA) ancestry groups. We illustrate how the RA maps we develop can be used, together with GWAS results, to improve prediction accuracy in underrepresented ancestry groups. Our study is accompanied by the software needed to develop RA maps for other ancestries or data sets.

Results

The MC-ANOVA method estimates the impact of differences in allele frequencies and LD patterns between ancestries on the local relative accuracy (RA, and functions thereof) of PGS. To define RA, let us consider a scenario where the same causal additive model holds in two ancestry groups:

$$y_i = \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_i \tag{1}$$

where y_i ($i = 1,\ldots,n$ is an index for subjects) is a phenotype, $\mathbf{z}_i$ is the (centered) vector of SNP genotypes at causal loci (QTL), and $\boldsymbol{\alpha}$ is the vector of effects. Now, let us consider an instrumental model where phenotypes are regressed on SNPs that may not necessarily have a causal effect (markers):

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i \tag{2}$$

where $\mathbf{x}_i$ is a vector of (centered) SNP genotypes at markers.

For a single marker-QTL pair j, the (population) marker effect is defined as:

$$\beta_j = \frac{\mathrm{Cov}\left(x_{ij}, z_{ij}\right)}{\mathrm{Var}\left(x_{ij}\right)}\alpha_j \tag{3}$$

where $\mathrm{Var}\left(x_{ij}\right)$ is the marker variance and $\mathrm{Cov}(x_{ij}, z_{ij})$ is the marker-QTL covariance (both scalars). Extending this to a multilocus model[21], we have that the vector of population marker effects is defined as:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XZ}\boldsymbol{\alpha} \tag{4}$$

where $\boldsymbol{\Sigma}_X$ is the covariance matrix of marker genotypes and $\boldsymbol{\Sigma}_{XZ}$ is the covariance matrix between marker and QTL genotypes.

Within-ancestry R-squared

Within an ancestry group, the maximum proportion of variance of the genetic values that can be explained by a regression on SNPs (assuming SNP effects are known with certainty) depends on the extent of LD between the SNPs used in [1] and those in [2], specifically (see the Supplementary Methods for a step-by-step derivation of [5]):

$$R^2 = \mathrm{Corr}\left(\mathbf{x}_i'\boldsymbol{\beta}, \mathbf{z}_i'\boldsymbol{\alpha}\right)^2 = (\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{ZX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XZ}\boldsymbol{\alpha})/(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_z\boldsymbol{\alpha}) \tag{5}$$

Under perfect LD between markers and QTL (something that will occur if the causal loci are genotyped or are perfectly predicted by markers), [5] would be equal to one. However, if there is imperfect LD between markers and QTL, R^2 would be less than one. Thus, the R-squared in [5] captures the impact of imperfect LD between markers and QTL on the proportion of variance at causal loci that can be explained by a regression on SNPs[22] within a population.

Cross-ancestry R-squared

An R-squared similar to [5] can be derived for cross-ancestry prediction by using marker effects from an ancestry (ancestry 1, $\boldsymbol{\beta}_1$ [4]) to predict genetic scores in a different ancestry group (ancestry 2). Thus, introducing ancestry group notation, we can define cross-ancestry R-squared as:

$$R^2_{1\to 2} = \mathrm{Corr}\left(\mathbf{x}_{i_2}'\boldsymbol{\beta}_1, \mathbf{z}_{i_2}'\boldsymbol{\alpha}\right)^2 \tag{6}$$

where $\mathbf{x}_{i_2}$ and $\mathbf{z}_{i_2}$ are the marker and QTL genotype vectors of an individual from the target ancestry (ancestry 2), $\boldsymbol{\beta}_1$ is the vector of marker effects from ancestry 1, and $\boldsymbol{\alpha}$ is the vector of QTL effects in the target ancestry. The Supplementary Method present a step-by-step derivation of [5] and [6], expressing the within- and cross-ancestry R-squared parameters as a function of (co)variance matrices of alleles at markers and QTL loci and the QTL effects.

It is important to highlight that the R-squared defined above (expressions [5] and [6], as well as the expressions presented in the Supplementary Methods) are not directly comparable to empirical PGS R-squared values commonly reported in the literature because empirical PGS R-squared values quantify the proportion of variance of a phenotype that can be explained by a PGS (and such, its upper limit is the genomic heritability). The R-squared defined above capture the proportion of genetic (not phenotypic) variance at causal loci that can be explained by regression on SNPs (as such, the upper limit for [5] and [6] is one; this will happen under perfect LD between markers and causal variants).

Relative accuracy

Following Wang et al.[7], we define the RA of a PGS as:

$$\mathrm{RA} = \frac{R^2_{1\to 2}}{R^2_{1\to 1}} \tag{7}$$

where $R^2_{1\to 1}$ is a within-ancestry R-squared (i.e., the proportion of variance at causal loci that can be explained by regression on markers within-ancestry group [5]), and $R^2_{1\to 2}$ is a cross-ancestry R-squared [6]. Under the assumption that the effects of the causal loci are the same in both ancestries and in the absence of allele frequency or LD differences between ancestries, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. In this case, the RA will equal one. However, if there are allele frequency or LD differences between ancestries and imperfect LD between markers and causal variants (QTL), $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ and the RA will be less than one. Thus, the RA captures the proportion of the reduction in PGS prediction R-squared attributable to allele frequency and LD differences between ancestries.

Monte Carlo Analysis of Variance (MC-ANOVA)

Estimating the R-squared parameters ([5] and [6]) and the RA ([7]) requires knowledge of the QTL positions and effects ($\boldsymbol{\alpha}$), which are both unknown. Therefore, we propose a Monte Carlo (MC) algorithm (Fig. 1) that, for a given chromosome segment, estimates the distribution of these R-squared values by computing R-squared values

https://doi.org/10.1038/s41467-024-54727-8

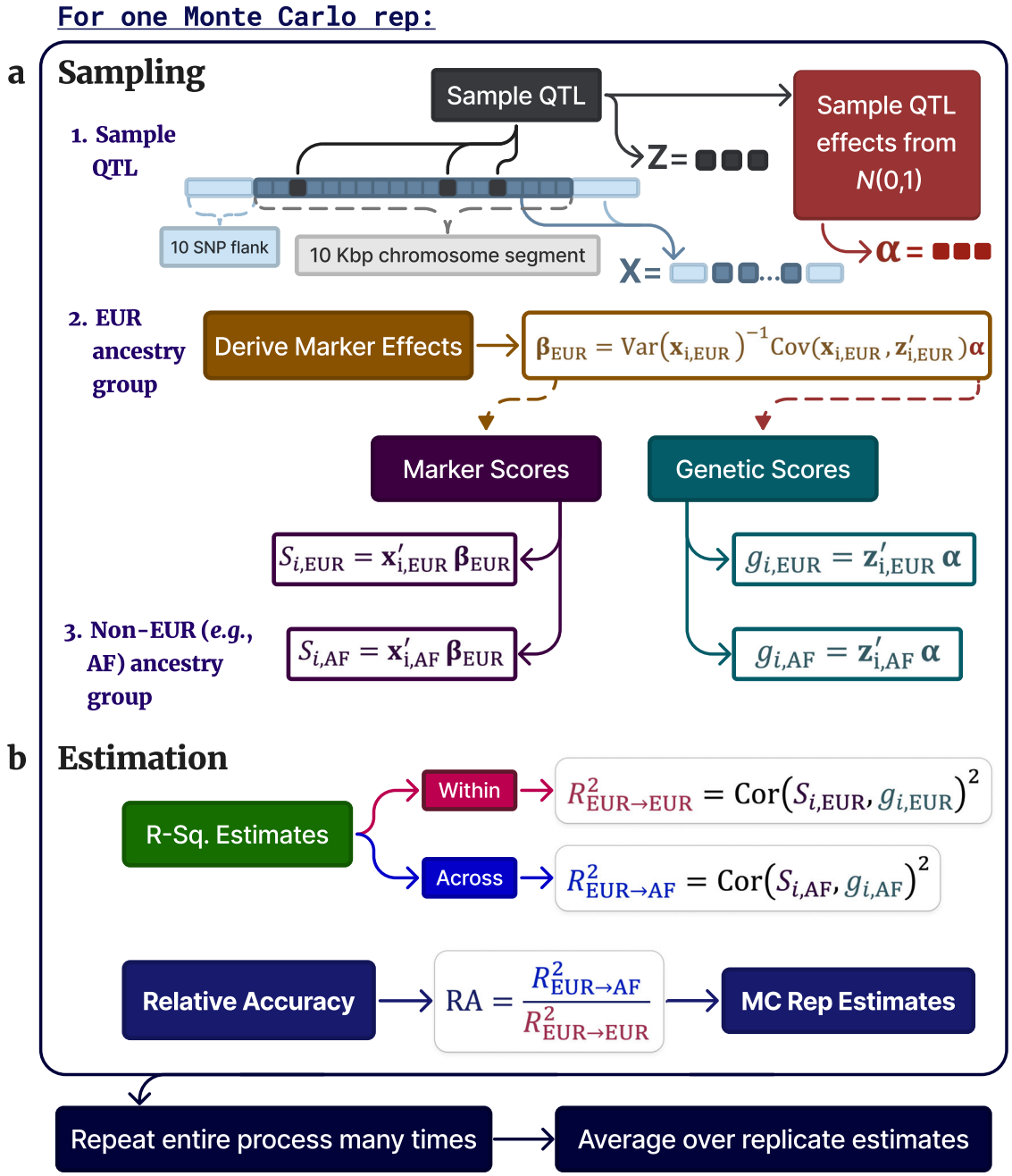

Fig. 1 | A representation of the MC-ANOVA algorithm. MC-ANOVA uses genetic data from two or more ancestry groups (here, to illustrate we consider European [EUR] and African [AF] ancestry) to estimate the proportion of variance at causal loci explained by EUR-derived marker effects in testing data from EUR and non-EUR (e.g., AF) ancestry groups. To estimate the relative accuracy (RA) for a given chromosome segment (e.g., all loci in a ten Kbp segment), MC-ANOVA assumes that the same additive genetic model ($g_{i.} = \mathbf{z}'_{i.}\boldsymbol{\alpha}$, * = EUR or AF) holds in both ancestry groups. Within a short chromosome segment, for one Monte Carlo replicate (MC rep), we sample quantitative trait loci (QTL) (e.g., three) positions ($\mathbf{z}_{i.}$ for $i = 1,\ldots,n$) at random (**a**). The remaining SNPs in the segment plus those in short flanking regions form a marker genotype vector $\mathbf{x}_i$. After sampling QTL effects ($\boldsymbol{\alpha}$) from a standard normal distribution, $N(0,1)$, genetic scores are computed as $g_{i.} = \mathbf{z}'_{i.}\boldsymbol{\alpha}$. Marker effects are derived from the EUR ancestry group using $\boldsymbol{\beta}_{EUR} = \mathrm{Var}(\mathbf{x}_{i_{EUR}})^{-1}\mathrm{Cov}(\mathbf{x}_{i_{EUR}}, \mathbf{z}'_{i_{EUR}})\boldsymbol{\alpha}$, where Var is the variance and Cov is the covariance, and these effects are used to obtain local marker scores for both ancestry groups ($S_{i.} = \mathbf{x}'_{i.}\boldsymbol{\beta}_{EUR}$). The squared correlations (Cor) between genetic ($g_{i.}$) and marker ($S_{i.}$) scores are used to derive cross-ancestry and within-ancestry R-squared (R-sq.) values, and the RA is computed as the ratio between the two (**b**). This procedure is repeated a large number of times for each segment, resampling QTL positions and their effects every time. For each segment, the R-squared and RA values are averaged across MC replicates. The procedure is applied to each chromosome segment.

over possible configurations of marker and causal loci and their effects. The algorithm is an extension of a method proposed by us previously[23] to estimate the proportion of variance of a high-dimensional set by a regression on another high-dimensional set (in our case, the QTL by the SNPs). Additional details of the MC-ANOVA algorithm can be found in the Methods.

Maps of the relative accuracy of European-derived PGS in non-Europeans

We used the MC-ANOVA method to develop maps of the RA of EUR-derived marker effects in non-EUR ancestry groups from the UK Biobank. We developed RA maps using SNPs from the UK Biobank arrays (~610,000 SNPs with minor-allele frequency ≥1%) as well as using

Table 1 | Average R-squared and relative accuracy (RA) by testing set (based on SNPs from the UK Biobank arrays)

Ancestry Group	Sample Size	Fst with EUR	R-squared ($R^2_{1\to2}$)[a]	Relative Accuracy ($R^2_{1\to2}/R^2_{1\to1}$)	Standard Error of RA[c]	Variance in RA Across Segments[c]
European (EUR)	230,000	---	0.648[b]	1.000	---	---
African (AF)	3083	0.120	0.182	0.268	0.016	0.033
Caribbean (CR)	3343	0.102	0.228	0.340	0.017	0.030
East Asian (EA)	1329	0.095	0.379	0.564	0.030	0.043
South Asian (SA)	7919	0.022	0.506	0.771	0.017	0.016

[a]Subscript 1 always indicates an EUR training or testing set; 2 indicates non-EUR testing.
[b]$R^2_{1\to1}$.
[c]Median.

~ 1.3 million HapMap SNPs (with minor-allele frequency ≥0.1%) that were present in the imputed UK genotypes (see Methods for further details on the QC and filtering steps).

To develop each of these portability maps, we partitioned the genome into short nonoverlapping segments that were at least ten Kbp long and had at least ten SNPs. We chose to use short chromosome segments to capture the proportion of variance at causal loci that can be explained (in both within- and cross-ancestry prediction) by SNPs that are physically close to causal variants. The average segment was 45 Kbp long (containing 12 core SNPs) in the case of the map derived using SNPs from the UK Biobank arrays and 22 Kbp long (containing 13 core SNPs) in the case of the map using HapMap variants. Though the base-pair length differed, the average and median number of SNPs per segment in each map were very similar. The results from the map developed using SNPs from the UK Biobank arrays are presented in the main body of the article, and those based on HapMap variants are provided as supplementary data. Whenever pertinent, we discuss the differences between the two maps.

The derivation of marker effects (Fig. 1) and within-ancestry R-squared [5] used the genotypes of 230,000 distantly related EUR ancestry individuals from the UK Biobank. To estimate the cross-ancestry R-squared [6] and the RA [7], we used data from the UK Biobank of individuals of African (AF), Caribbean (CR), East Asian (EA), and South Asian (SA) ancestry (Table 1 and Supplementary Fig. 1). Further details about sample selection and SNP QC are offered in the Methods section. An interactive R Shiny app that displays RA estimates (from the UK Biobank arrays or HapMap variants) for user-specified genome positions (or SNP IDs) was created and is available via an R package and also on a website (see Supplementary Notes for more information). In addition, the portability map based on the UK Biobank arrays is provided in Supplementary Data 1 and the portability map based on the HapMap variants is provided in Supplementary Data 2.

MC-ANOVA predicts low relative accuracy of PGS between ancestry groups

Averaged over the genome, the within-EUR R-squared, $R^2_{1\to1}$ [5], was 0.65. This suggests that within-EUR ancestry SNPs from the UK Biobank arrays could explain roughly two-thirds of the genetic variance at ungenotyped causal loci that have a similar allele frequency distribution to the SNPs in the UK Biobank arrays (Table 1). The cross-ancestry R-squared [6] estimates were much lower, ranging from 0.182 (AF) to 0.506 (SA), which resulted in RA estimates ranging from 0.268 (AF) to 0.771 (SA). As expected, the RA was inversely related to the genetic distance between the testing ancestry and the EUR group (Table 1 and Supplementary Fig. 1). For example, the AF ancestry group had the highest Fst[24] with the EUR group (0.120) and the lowest whole-genome RA (0.268), while the SA group had the lowest Fst (0.022) and the highest RA (0.771). The estimated R-squared values were significantly higher when the map was produced using HapMap variants (Supplementary Fig. 2 and Supplementary Table 1). The variance of RA between segments was slightly smaller when the map was produced with the HapMap variants (Table 1 and Supplementary Table 1). The increase in RA with the HapMap variant-based map was expected, given that this map had twice as many SNPs as the one using array SNPs.

Predicted versus empirical RA

We used data from distantly related EURs from the UK Biobank and a Bayesian shrinkage variable-selection prediction method (BayesC[25]) to develop PGS for six complex traits: height, HDL, LDL, serum urate, BMI, and serum glucose (see Methods for details about the phenotypes and methods used to derive the PGS). Using testing data from the EUR and non-EUR ancestry groups (Supplementary Table 2), we estimated the empirical prediction R-squared for each trait, the corresponding empirical RA (i.e., the ratio of the PGS R-squared in non-EURs relative to the PGS R-squared in EURs), and the loss of accuracy (LOA) attributable to allele frequency and LD differences between ancestries[7] ($\text{LOA}\,\% = \frac{1-\text{predicted RA}}{1-\text{empirical RA}} \times 100$, where the predicted RA is defined in [7]).

For most traits, the empirical RA estimates were smaller than the predicted RA (Fig. 2 for UK Biobank arrays and Supplementary Fig. 3 for HapMap variants). This is expected because the MC-ANOVA-predicted RA captures the LOA attributable to allele frequency and LD differences, which together are only one source of LOA. In general, for any given trait, ancestries with higher predicted RA also had higher empirical RA (Fig. 2). This suggests that, as noted earlier by Wang et al.[7] allele frequency and LD differences between ancestries are a substantial factor affecting the portability of PGS and that the MC-ANOVA estimates capture that. For most traits and ancestry groups, allele frequency and LD differences alone explained more than 50% of the empirical LOA. However, for glucose, the proportion of reduction in accuracy explained by allele frequency and LD differences was smaller. This could suggest that differences in the genetic architecture (including both heritability and polygenicity) of traits between ancestries and G×E interactions may play a more important role in glucose than the other traits evaluated. For example, height is highly heritable and highly polygenic, and BMI is also highly polygenic and moderately heritable. On the other end, glucose has a moderately low heritability and is less polygenic than height or BMI[26–28].

We compared our PGS-predicted RA and LOA estimates with those reported by Wang et al[7]., who developed a method to predict RA and LOA for specific PGS. Overall, except for LDL, our results were similar to those published by Wang et al. in terms of both RA and LOA (Supplementary Fig. 4), although, unlike Wang et al.'s method, MC-ANOVA does not use trait-specific SNP effect estimates.

The (local) relative accuracy of PGS varies along the genome

The results presented above were based on the estimated R-squared and RA averaged across the genome or the segments of the genome represented in a PGS. However, in line with our main hypothesis, we found sizable variability in cross-ancestry R-squared [6] and RA between chromosome segments (Fig. 3 and Supplementary Fig. 5 [UK Biobank arrays]; Supplementary Fig. 6 and Supplementary Fig. 7 [HapMap variants]), suggesting that even for ancestries with a low overall RA (e.g., AF), there are still chromosome segments with high RA and portability of EUR-derived PGS. The distribution of the within-EUR R-squared [5] values was symmetric; however, for ancestries with a

https://doi.org/10.1038/s41467-024-54727-8

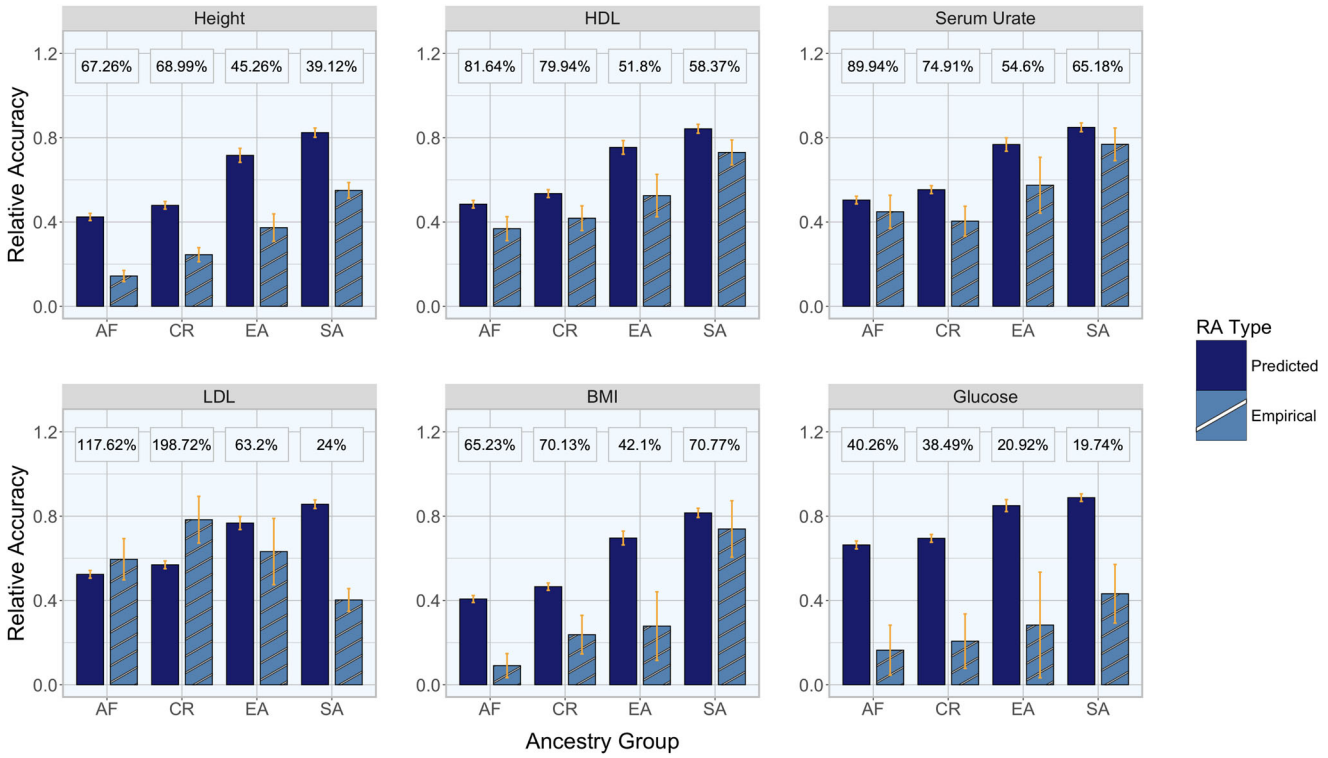

Fig. 2 | Predicted relative accuracy versus empirical relative accuracy with UK Biobank array SNPs. MC-ANOVA predicted relative accuracy (RA) versus empirical RA of European (EUR)-derived polygenic scores when used to predict phenotypes of individuals of non-EUR ancestry (AF, CR, EA, and SA denote African, Caribbean, East Asian, and South Asian ancestry, respectively). Each panel displays a different phenotype (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose). The loss of accuracy (LOA, %) attributable to allele frequency and LD differences between ancestries is shown on top of each bar set. A standard error bar of each mean RA estimate is shown and derivation details are in the Supplementary Methods and details for the empirical RA are in the Methods. The sample sizes used to derive the standard errors are in Supplementary Table 2. These results are based on SNPs from the UK Biobank arrays; see Supplementary Fig. 3 for results obtained using HapMap SNPs.

strong African ancestry influence, the distribution of the cross-ancestry R-squared [6] was heavily right-skewed, with most of the chromosome segments having a low cross-ancestry R-squared.

The estimates presented in Fig. 3 correspond to average results across MC runs of the MC-ANOVA algorithm. In our maps, we also provide the standard deviation (SD) of the distribution of the R-squared and RA parameters across MC replicates, along with the standard error of the means (Supplementary Methods). The median cross-ancestry R-squared [6] standard error was 8.0% (median) of the point estimates and the within-ancestry R-squared [5] variance was 1.7%. To illustrate the uncertainty associated with the reported R-squared estimates, we sampled 100 segments for each ancestry group and displayed the within- and cross-ancestry R-squared point estimates with their corresponding standard error bars in Supplementary Fig. 8.

MC-ANOVA estimates are predictive of the local RA of empirical PGS

The results shown in Fig. 3 suggest that in any ancestry group, but particularly for those that are more genetically distant from the EUR ancestry, the predicted cross-ancestry R-squared and RA vary substantially over the genome. To evaluate whether MC-ANOVA estimates are predictive of the local RA of real PGS, we first grouped SNPs into sets according to their MC-ANOVA predicted cross-ancestry R-squared [6] and used this to define four portability groups: Very Low, Low, Medium, and High (Table 2 for AF; Supplementary Table 3 for CR, EA, and SA). Then, we decomposed the trait-specific PGS into subscores, each using the SNPs in a predicted portability group. Finally, we computed the correlation between each subscore and their corresponding adjusted phenotype in testing sets for EUR and non-EUR, as well as the difference in the correlations of within- and cross-ancestry PGS prediction.

For most traits, we observed that the difference in empirical PGS correlation (non-EUR PGS correlation subtracted from EUR PGS correlation) decreased as the predicted portability of the SNP set increased (Fig. 4 for AF; Supplementary Fig. 9 for CR, EA, and SA). For instance, for individuals of AF ancestry, the difference in the within- and cross-ancestry PGS and phenotype correlations for height ranged from 0.30 for the Very Low portability group of SNP segments to just 0.06 for the High portability group of SNP segments (top-left panel in Fig. 4). Similar patterns were observed for the other traits (and ancestry groups; Supplementary Fig. 9). For serum urate and HDL cholesterol, there was near-perfect portability of PGS between EUR and AF for SNPs in the High portability group. Furthermore, the LOA attributable to allele frequency and LD differences estimated within each SNP portability group was lowest in the High portability group for most traits and ancestry groups (Supplementary Fig. 10). For example, in the AF group, we achieve a LOA for height of just 9.2% for the High portability group, but in the Very Low portability group, the LOA is 88.3%. This indicates that MC-ANOVA-predicted portability is predictive of the empirical RA and LOA of chromosome segments.

Using HapMap SNPs did not notably improve PGS local portability over using the called genotypes set (Supplementary Fig. 11). Overall, the validation results obtained with the HapMap-based map were similar to the ones reported for the map based on SNPs of the UK Biobank arrays; however, the grouping of SNPs based on the HapMap-based map was not as effective at reducing the empirical difference in prediction correlation between EUR and non-EUR ancestry groups as

https://doi.org/10.1038/s41467-024-54727-8

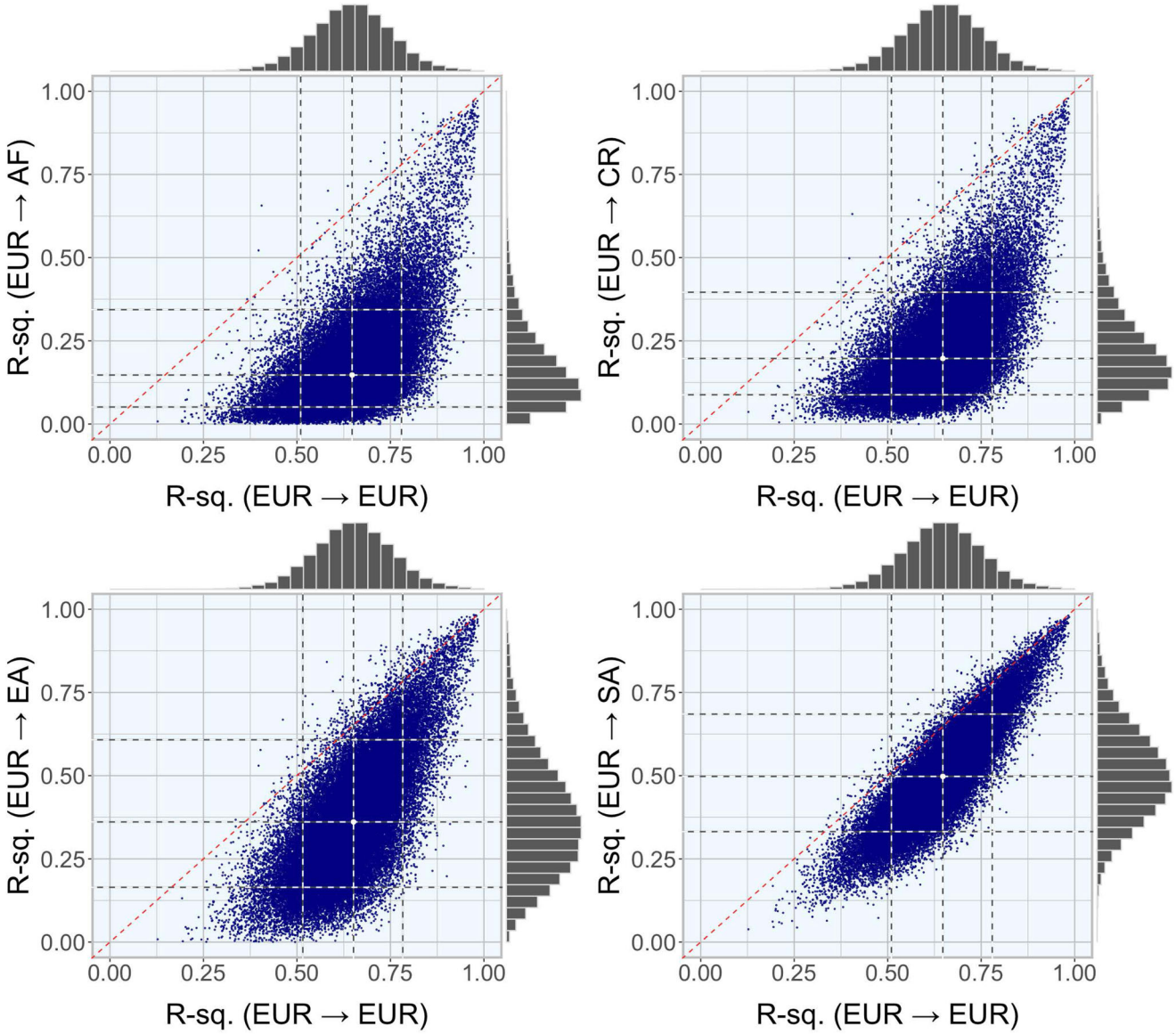

Fig. 3 | Within- and cross-ancestry R-squared distributions based on UK Biobank array SNPs. Distribution of the cross-ancestry R-squared (R-sq.) versus the within-European (EUR) R-squared for the African (AF), Caribbean (CR), East Asian (EA), and South Asian (SA) ancestry groups obtained when using SNPs from the UK Biobank arrays (see Supplementary Fig. 6 for results based on HapMap SNPs). Each panel displays a different non-EUR ancestry group. Each point represents a small chromosome segment (45 Kbp) and a histogram of the distribution of the points is also shown along each axis. Each subplot has dashed gray lines at the 10th, 50th, and 90th percentiles of the distribution and a red dashed 45-degree reference line (slope of one and intercept at zero). There is a white point at the intersection of the within-ancestry R-squared median and the cross-ancestry R-squared median. See Supplementary Fig. 6 for results based on HapMap SNPs.

Table 2 | Estimated relative accuracy (RA) of the SNP segments across the genome grouped by their estimated portability in terms of cross-ancestry R-squared ($R^2_{1\to 2}$ for 1 = EUR and 2 = AF testing set). Results were obtained using SNPs from the UK Biobank arrays

Testing Group	Portability Group	Quantile Group Cutoff	$R^2_{1\to 2}$ Range	Number of SNPs	Average $R^2_{1\to 1}$	Average $R^2_{1\to 2}$	Average RA ($R^2_{1\to 2}/R^2_{1\to 1}$)
African (AF)	High	(0.8,1]	(0.26,0.97]	122,135	0.751	0.40	0.529
	Medium	(0.6,0.8]	(0.18,0.26]	122,131	0.674	0.215	0.323
	Low	(0.5,0.6]	(0.15,0.18]	61,065	0.646	0.162	0.255
	Very Low	[0,0.5]	[0,0.15]	305,352	0.597	0.086	0.144

(See Supplementary Table 3 for other ancestry groups).

https://doi.org/10.1038/s41467-024-54727-8

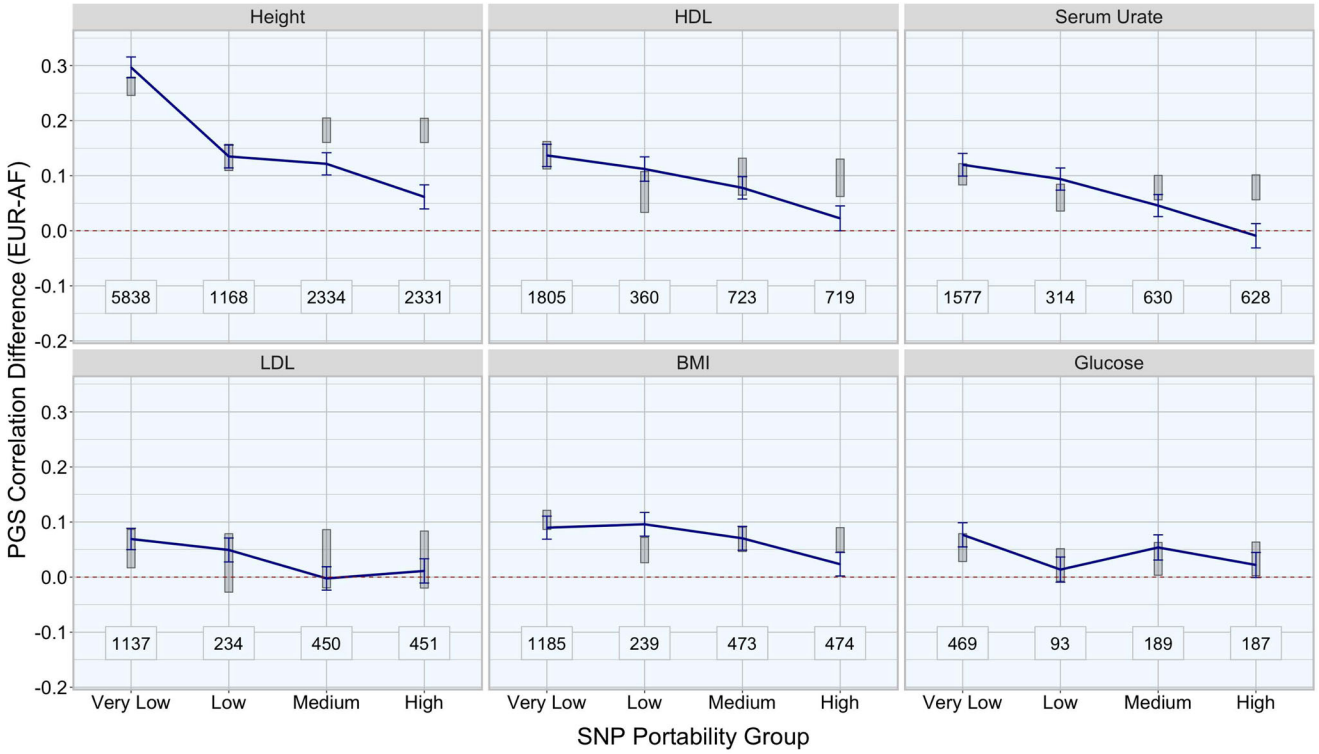

Fig. 4 | The difference between polygenic score prediction correlation by SNP portability group based on UK Biobank array SNPs. The vertical axis represents the difference between the within- and cross-ancestry polygenic score prediction correlations of European (EUR) derived polygenic scores (PGS) for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability ($R^2_{1\to 2}$ groupings, Table 2) by trait. Each panel displays a different phenotype (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose). A positive difference in PGS prediction correlation indicates that the PGS of the SNP set had a higher prediction correlation in EUR (within-ancestry prediction) than in individuals of African (AF, cross-ancestry prediction) ancestry. The number of SNPs entering each PGS is annotated toward the bottom of each subplot. A standard error bar for each prediction correlation difference is shown and details for the calculation can be found in the Methods. The gray vertical bars are the simulated null distribution (mean +/- standard error of 2,000 iterations) for the correlation difference, where SNPs were assigned to portability groups completely at random, maintaining the number of SNPs in each subgroup. The sample sizes for the simulated null distribution are in Supplementary Table 2. See Supplementary Fig. 9 for results for other ancestry groups (Caribbean, East Asian, and South Asian) and Supplementary Fig. 11 for results based on HapMap SNPs.

with the map based on SNPs from the UK Biobank arrays (Fig. 4, and Supplementary Figs. 9, 11). We believe this may partially reflect possible artifacts induced by the use of imputed SNPs which may lead to upwardly biased estimates of RA.

To benchmark the results of Fig. 4, we performed a similar analysis to that presented in Fig. 4, Supplementary Fig. 9, and Supplementary Fig. 11 classifying SNPs into portability groups using Fst[24] and Wang et al.'s RA method[7] (Supplementary Fig. 12). Overall, MC-ANOVA was considerably more effective at identifying SNP sets with varying levels of portability than Fst or Wang et al.'s RA. Fst was very poor at predicting the RA of trait-specific local PGS, and Wang et al.'s RA was only effective at detecting SNP sets with different RAs for height (Supplementary Fig. 12). Conversely, both the High and Medium portability groups based on MC-ANOVA were different from the simulated null for height, and the High portability group based on MC-ANOVA was different from the simulated null for HDL, serum urate, and BMI (Fig. 4).

Genomic regions with high RA are enriched for GWAS hits and high SNP density

We investigated whether the MC-ANOVA estimates of R-squared and RA were associated with the presence of GWAS hits (p-value < 5e-8; Supplementary Table 4) in the EUR ancestry. We found that genomic regions with higher MC-ANOVA R-squared values were highly enriched for GWAS hits for all the traits investigated (Fig. 5) and tended to have higher marker density (which, in turn, leads to higher LD between markers and causal variants). However, for segments with similar marker density to each other, the R-squared estimates were relatively uniformly distributed across the entire range (Supplementary Fig. 13), especially for the EA and SA ancestry groups. This suggests that high marker density is a necessary but not sufficient condition to achieve high MC-ANOVA R-squared values.

Using RA to improve cross-ancestry prediction of transfer learning algorithms

To demonstrate how RA maps can be used to improve cross-ancestry PGS prediction accuracy, we evaluated PGS informed by the RA maps in the context of transfer learning. Gradient Descent with Early Stopping (GD-ES) is a widely employed technique for transfer learning (TL) in various machine learning algorithms. Recently, Zhao et al[29]. introduced the application of GD-ES in constructing PGS for cross-ancestry prediction. This approach uses EUR-derived SNP effect estimates as initial values for a GD-ES algorithm that updates these estimates iterating on data from the non-EUR target population. In GD-ES, a learning rate parameter is used to control the strength of the updates. In Zhao et al.[29], the learning rate was the same for all SNPs in the PGS. We took this concept one step further by using the cross-ancestry RA maps to inform the learning rate of the gradient descent algorithm, making it SNP-specific (see Methods). Specifically, we allowed for stronger learning rates for SNPs in regions with low predicted portability and weaker learning rates for SNPs with high cross-ancestry portability. We applied this approach to develop PGS for non-EUR ancestry groups from the UK

https://doi.org/10.1038/s41467-024-54727-8

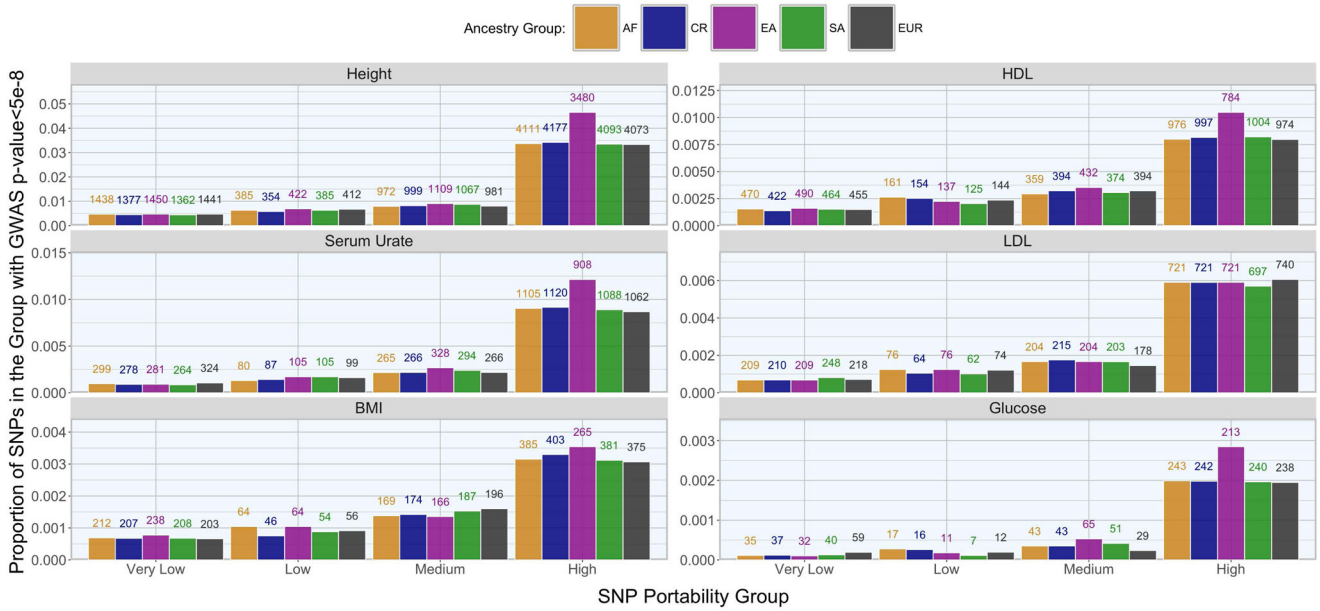

Fig. 5 | The proportion of UK Biobank array SNPs that were significantly associated with a trait for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability. The y-axes give the proportion of SNPs for which a European (EUR)-based genome-wide association study (GWAS) *p*-value (based on a two-sided test of a t-statistic, with the null hypothesis that the SNP effect is zero) was less than 5e-8 within each portability group (x-axes). Each panel displays a different phenotype (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose). For the EUR testing set, the grouping was based on the within-ancestry R-squared [5]. The grouping was based on the cross-ancestry R-squared [6] for the African [AF], Caribbean [CR], East Asian [EA], and South Asian [SA] testing sets. The number of SNPs is noted above each bar and is based on SNPs from the UK Biobank arrays.

Biobank, using EUR-derived effects as initial values. Our preliminary results (Supplementary Table 8) suggest that using RA-informed learning rates can improve cross-ancestry prediction accuracy over using a fixed learning rate in most traits evaluated for prediction in an external testing set (see Methods). The improvement is particularly clear in the CR and AF ancestry groups (Supplementary Table 8).

External validation

The results presented thus far were entirely based on UK Biobank data. Prediction across cohorts poses additional challenges (e.g., the use of different SNP arrays and G×E factors). Therefore, to assess the performance of MC-ANOVA in an external validation, we conducted an evaluation using data from the Atherosclerosis Risk in Communities (ARIC) study[20]. The validation involved 9628 European American (AEA) and 3130 African American (AAA) participants from the ARIC study. For these analyses, we utilized a set of 795,613 SNPs that were common between the genotypes of the ARIC study and the imputed genotypes from the UK Biobank[19]. The AEA group from the ARIC study served as a within-ancestry (cross-data set) testing set, while the AAA group from the ARIC study served as a cross-ancestry (and cross-data set) testing set. We evaluated global RA and LOA, as well as local PGS, based on the predicted portability groups based on the MC-ANOVA R-squared estimates [6] for height, serum urate, and BMI. The whole PGS empirical RA estimates were higher than those of the within data set (UK Biobank only) analysis for height (approximately 0.35) and BMI (approximately 0.25), and the predicted RA estimates were correspondingly higher as well. The whole PGS LOA attributable to allele frequency and LD differences across height, serum urate, and BMI was approximately 60% (Supplementary Fig. 14a), which is similar to what we estimated using the UK Biobank data. The assessment of empirical correlation difference (UK Biobank EUR → ARIC AEA minus UK Biobank EUR → ARIC AAA) within SNP sets grouped by MC-ANOVA portability estimates validated the results for height, as the empirical correlation difference deviated from the simulated null distribution in the High portability group (Supplementary Fig. 14b).

Discussion

Previous studies suggest that between-ancestry differences in allele frequencies and LD patterns are a major factor contributing to the loss of accuracy (LOA) in cross-ancestry PGS prediction[6–8]. For instance, Privé et al.[8] showed that the portability of PGS between ancestry groups worsens with the genetic distance between the groups, and Wang et al.[7] reported that much of the LOA in prediction from European (EUR) to African (AF) ancestry could be attributed to allele frequency and LD differences. However, no previous study has investigated whether the relative accuracy (RA) of cross-ancestry PGS varies along the genome. To address this knowledge gap, we developed a novel approach (MC-ANOVA) to estimate the RA of short chromosome segments. MC-ANOVA estimates the RA of randomly generated linear functions of genotypes within each chromosome segment, making MC-ANOVA a trait-agnostic method that is solely based on genome information. The methodology can be used to map regions of high and low (local) PGS portability between two or more ancestry groups. We applied MC-ANOVA to UK Biobank data to generate maps (with a mapping resolution of ~ 45 Kbp) of the maximum expected RA when EUR-derived SNP effects are used to predict phenotypes or disease risk of non-EURs, including individuals of AF, Caribbean (CR), East Asian (EA), and South Asian (SA) descent. Finally, we validated these RA maps by quantifying the empirical RA of real PGS for SNP sets with High, Medium, Low, and Very Low MC-ANOVA predicted portability for prediction within and across data sets.

Genome differentiation between populations has been a focus of population genetics for more than seven decades. The Fst[24] metric quantifies differentiation in allele frequencies. MC-ANOVA and Wang et al.'s RA method[7] capture both differences in allele frequencies and LD patterns, with the key difference being that Wang et al.'s RA method accounts for pairwise LD and MC-ANOVA uses a multilocus regression approach that accounts for the full patterns of conditional linear dependence/independence of loci within a segment and does not require assuming that causal variants are independent. Additionally, unlike Wang et al. method, MC-ANOVA is trait-agnostic in that it does

not use SNP effect estimates. This makes MC-ANOVA suitable to develop RA maps that can be used with any trait. We benchmarked MC-ANOVA against Fst and Wang et al.'s RA metric in terms of the ability of the methods to identify SNPs with Very Low, Low, Medium, and High portability. In the benchmark analysis, MC-ANOVA convincingly outperformed both Fst and Wang et al.'s RA method across traits and ancestry groups (Supplementary Fig. 12).

Consistent with previously reported LOA estimates[7], we found that, on average, allele frequency and LD differences between ancestries explained approximately half of the LOA genome-wide in the EA and SA ancestry groups and approximately two-thirds in the AF and CR groups. As expected, for the average chromosome segment, MC-ANOVA predicts lower RA for groups more genetically distant (e.g., EUR→AF or EUR→CR) relative to genetically closer groups (e.g., EUR→EA or EUR→SA). These results support the literature that allele frequency and LD differences between ancestries significantly affect the RA of PGS across ancestries. However, we also found significant variability in RA across chromosome segments. Indeed, even for the more genetically distant groups (e.g., EUR→AF), we found many segments with high predicted RA. This is important because it suggests that there are many genomic regions of the genome for which results from large EUR GWAS may be portable to non-EUR ancestries, which has the potential for improving cross-ancestry prediction.

MC-ANOVA estimates capture the components of LOA attributable to differences in allele frequencies and LD between ancestry groups, which together are only one of the factors affecting the RA of PGS in cross-ancestry prediction. Therefore, MC-ANOVA-predicted RA should be considered the maximum RA that one could achieve in cross-ancestry prediction, under the implicit assumption that causal variants are being tagged by SNPs within ~ 45 Kbp. The gap between the predicted empirical RA varied between traits. For example, among the traits we considered, the gap between the MC-ANOVA predicted RA and the empirical RA appeared to be largest for glucose (Fig. 2), a trait that is likely to be more affected by G×E exposures (e.g., diet, lifestyle, and exercise) that can be correlated with ancestry. Likewise, the ability of MC-ANOVA RA maps to identify regions of high and low RA varied between traits (Fig. 4). For traits with an extremely polygenic genetic architecture (e.g., height and BMI[26,27]), MC-ANOVA appeared to be more predictive of the empirical difference in the PGS prediction correlation between the EUR and non-EUR groups than for traits such as glucose. This is expected because MC-ANOVA estimates the RA of linear functions averaging over many possible randomly drawn linear combinations of SNP and QTL genotypes.

The MC-ANOVA algorithm is controlled by a few parameters, including the segment size, the number of causal variants within the segment, the number of SNPs in the flanking regions, and the distribution causal variant effects are drawn from. The RA maps that we present in this study are based on small (~ 45 Kbp) segments, each containing three causal variants (which are randomly chosen in each MC replicate) and ten SNPs in each of the flanking regions of the segment. We chose these parameters to achieve a relatively fine mapping resolution for segments that may hold more than one causal variant. To assess the robustness of our results with respect to the parameter values chosen, we performed sensitivity analyses first varying the number of causal variants in the segment, then varying the flank size for a given QTL and segment size, and finally changing the distribution used to sample effects from Gaussian to Gamma (Supplementary Figs. 15a, b, 16, respectively). Overall, in all sensitivity analyses, we found that the distribution of the RA measures, as well as the genomic regions where RA peaks, were reasonably robust to the parameters of the MC-ANOVA algorithm, except in cases involving just one causal variant or no flanking SNPs. In these two cases, we observed a systematic reduction in R-squared parameters and RA (Supplementary Fig. 15c, d).

The RAs of the map developed with UK Biobank array SNPs (~ 610,000 SNPs) were smaller (and the variance in RA was higher) than those estimated using twice as many HapMap variants (~ 1.3 million SNPs). This can be attributed to the higher marker density of the HapMap variant set and the stronger LD among those variants compared to those of the UK Biobank array. The higher LD among variants in the HapMap variant set was both a consequence of the higher marker density and of a distribution of the minor allele frequency (MAF) that was symmetric and with a mode near 0.24. On the other hand, the distribution of the MAF in the array set had an enrichment in the lower MAF which would impose limits on the maximum LD[30]. Furthermore, correlated imputation errors (which may result from a tendency to impute genotypes from certain haplotypes) may lead to a spurious increase in LD among imputed variants. Overall, the global MC-ANOVA predicted relative accuracy was more similar to the empirical relative accuracy with the UK Biobank array-based map (Fig. 2) than the HapMap-based map (Supplementary Fig. 3). Furthermore, the UK Biobank array-based RA map was slightly better than the HapMap-based map at predicting the empirical differences between the within- and cross-ancestry PGS prediction correlation (compare Fig. 4 and Supplementary Fig. 9 with Supplementary Fig. 11). Therefore, for PGS with SNPs within the allele frequency spectrum represented in the UK Biobank arrays, we recommend using the map based on UK Biobank array variants. Nevertheless, both maps are made available with this article.

When comparing RA estimates with GWAS results, we found that regions with high predicted portability are highly enriched for GWAS hits. This is expected because RA is expected to be high in regions with strong and long-spanning LD and, at the same time, high LD among variants also increases the power to detect associations when causal variants are not genotyped. Furthermore, selection can lead to higher LD for loci with large effects on fitness traits[31,32]. A good example of the overlap of high RA in regions that have been detected to be associated with many traits, including many fitness traits, appears on chromosome six between 25.84 and 33.29 Mbp (Supplementary Fig. 5), which had the largest cross-ancestry R-squared [6] values in all four non-EUR ancestry groups. This peak closely overlaps with the major histocompatibility complex (MHC) region[33]. An abundance of literature has established that the MHC region includes numerous loci (e.g., human leukocyte antigen [HLA] genes) associated with many traits and diseases, particularly autoimmune diseases (e.g., nephropathy), infections, cancers, and psychiatric conditions (e.g., autism and schizophrenia)[1,33–37]. The MHC region is also known to be highly polymorphic, has high gene density, and has very strong LD[33,34,38]. Interestingly, for all four ancestry groups, the majority of the genes with the highest predicted portability were within chromosome six and the MHC region (Supplementary Tables 5–7).

An important question is whether the RA maps that we developed can be used to improve PGS prediction accuracy for groups that are underrepresented in GWA studies. For example, in the construction of PGS for cross-ancestry prediction, one could filter out SNPs that are in regions with very low predicted RA. However, in our maps, there were almost no segments with negative cross-ancestry correlation estimates. Therefore, we don't expect that removing SNPs based on their low RA would result in improved cross-ancestry PGS prediction. Another possibility is to use cross-ancestry predicted R-squared [6] estimates to inform transfer learning algorithms used to develop PGS for non-EUR ancestry groups. We found that using cross-ancestry predicted R-squared [6] to inform learning rates in a GD-ES[29] algorithm resulted in improvements in PGS prediction accuracy compared to an algorithm that used a fixed learning rate; thus, demonstrating an important practical application of the RA maps developed in this study.

In conclusion, we developed and validated a method to map the RA of short chromosome segments and used data from the UK

Biobank and the ARIC study cohorts to develop RA maps for several ancestry groups. These maps can provide valuable information for explaining GWAS replication (or lack thereof) across ancestry groups and can help in prioritizing variants for the development of PGS for cross-ancestry prediction. Together with the methods and results presented in this study, we provide software that can be used to generate RA maps for other data sets and ancestry groups and share the maps of RA through an R-package and a web interface.

Methods

Data

In this study, we used data from the UK Biobank and the ARIC study cohorts. For model training, we leveraged the large sample size of Europeans (EUR) from the UK Biobank. We conducted an internal validation using testing data from EUR and non-EUR from the UK Biobank and an external validation using data from European Americans and African Americans from the ARIC study.

UK Biobank cohort. We used distantly related individuals (defined as individuals with a within-ancestry genomic relationship < 0.05) from the UK Biobank. We randomly split the 236,698 distantly related EUR ancestry individuals into a training set of size 230,000 and a testing set of 6698. Additionally, UK Biobank testing sets included individuals of African ([AF], $n = 3083$), Caribbean ([CR], $n = 3343$), East Asian ([EA], $n = 1329$), and South Asian ([SA], $n = 7919$) ancestry (Table 1). Ancestral groups were defined by the UK Biobank self-reported Ethnic background (Data-Field 21,000[39]), but individuals were only included in each ancestry group if they passed the UK Biobank's Sample QC (Resource 531[39]), not excluded from kinship inference, included in phasing, and not identified as an outlier in heterozygosity and missing rates. Samples were also excluded if they withdrew from the study, if they had a mismatch of reported and genetic sex, if they were missing all six phenotypes of interest (described below), or if they were related to other samples with relatedness ≥ 0.05. Relatedness was determined using genomic relationship matrices ($\mathbf{G} = \frac{\mathbf{ZZ}'}{\mathrm{tr}(\mathbf{ZZ}')/n}$, where $\mathbf{Z}$ is the centered genotype matrix) computed within an ancestry group.

The ARIC study cohort. An external validation utilized the ARIC study, consisting of a European American (AEA) testing set of 9628 and an African American (AAA) testing set of 3130 based on self-reported race, which is highly concordant with the ancestry group defined based on SNP-derived principal component analysis[16]. The previously described EUR training set from the UK Biobank was used as the training set again for this external validation.

UK Biobank genotypes. For analysis involving the SNPs from the UK Biobank arrays, we used 610,791 genotyped SNPs from the UK Biobank Affymetrix array[19] in autosomal chromosomes. SNPs with a minor allele frequency of < 1% or a missing call rate > 5% overall (all ancestry groups combined) were excluded, and monomorphic SNPs in a particular ancestry group were excluded from analyses involving that group (108 for AF and 47,390 for EA). The base pair positions provided are based on GRCh37[19]. The HapMap SNP set used was based on the intersection of the Northern and Western European ancestry HapMap 3[40,41] SNPs and the UK Biobank imputed SNP genotypes[19]. 1,297,917 SNPs with a quality score > 0.7, a minor allele frequency in the full dataset cohort ≥ 0.1%, and not monomorphic in either the EUR or non-EUR cohorts were retained for analysis.

The ARIC study genotypes. For analysis involving model training in the UK Biobank and model validation in the ARIC study, we identified a common set of 795,613 autosomal chromosome SNPs between the ARIC study genotyped SNPs and the UK Biobank imputed set (excluding multiallelic variants)[19]. SNPs were excluded if they were monomorphic in either the UK Biobank EUR training set or one of the testing sets from the ARIC study. We checked for consistency of the genotyped strand and the reference alleles. SNP effects for SNPs with different reference alleles in the UK Biobank and the ARIC study (estimated in the UK Biobank) were multiplied by −1 before PGS were computed in the ARIC study cohorts.

Mapping the relative accuracy (RA) of cross-ancestry PGS prediction

MC-ANOVA method. MC-ANOVA uses genomic data from two or more ancestry groups (here, we use 1 = EUR and 2 = AF groups to illustrate). The goal is to estimate the proportion of variance (R-squared) at causal loci that can be explained by EUR-derived marker effects in testing data from EUR ($R^2_{1\to1} = \mathrm{Corr}(\mathbf{x}'_{i_1}\boldsymbol{\beta}_1, \mathbf{z}'_{i_1}\boldsymbol{\alpha})^2$ [5]) and AF ($R^2_{1\to2} = \mathrm{Corr}(\mathbf{x}'_{i_2}\boldsymbol{\beta}_1, \mathbf{z}'_{i_2}\boldsymbol{\alpha})^2$ [6]) ancestries. Here, $\mathbf{z}_{i*}$ and $\mathbf{x}_{i*}$ are genotypes at causal variants and markers (including markers in the core and flanking regions, Fig. 1) of group * (* = 1 or 2), respectively, $\boldsymbol{\alpha}$ is the vector of QTL effects (which are assumed to be the same in both groups), and $\boldsymbol{\beta}_1$ is the vector of marker effects in group 1. The relative accuracy (RA) ratio is then defined and computed as $\mathrm{RA} = R^2_{1\to2}/R^2_{1\to1}$. For a chromosome segment, MC-ANOVA estimates RA by quantifying the portability of randomly generated linear functions of SNP genotypes within short chromosome segments. We have previously shown that for general settings, the MC-ANOVA algorithm provides unbiased estimates of [5][23].

RA maps. To develop our RA maps, we first grouped SNPs into disjoint segments. For each chromosome, we partitioned the SNPs into ten Kbp nonoverlapping segments with a minimum of ten core SNPs per segment, leading to 52,956 segments for the SNPs from the UK Biobank arrays and 100,311 segments for the SNPs from the HapMap variants. The average SNP segment was 45 (22) Kbp long and contained 12 (13) core SNPs for the UK Biobank array SNPs (HapMap variants). The code used to define the SNP segments for the RA maps can be found at https://github.com/lupiA/MCANOVA (Supplementary Notes).

For each segment and Monte Carlo (MC) replicate, we sampled three QTL positions at random ($\mathbf{z}_{i.}$). The remaining SNPs in the segment plus 20 flanking SNPs (ten for each flanking region) were used as markers ($\mathbf{x}_{i.}$). QTL effects were sampled from IID standard normal distributions. For the sensitivity analysis shown in Supplementary Fig. 16, QTL effects were sampled from IID Gamma distributions with a shape parameter equal to 1.5 and a rate parameter equal to one. We computed genetic scores for the causal model for individuals from ancestries 1 and 2 using $g_{i_1} = \mathbf{z}'_{i_1}\boldsymbol{\alpha}$ and $g_{i_2} = \mathbf{z}'_{i_2}\boldsymbol{\alpha}$. Marker effects in ancestry group 1 were computed as $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}'_1\mathbf{X}_1 + \mathbf{I}k)^{-1}\mathbf{X}'_1\mathbf{Z}_1\boldsymbol{\alpha}$, where $k = 1\mathrm{e}\text{-}8$ was a small constant added to the diagonal of $\mathbf{X}'_1\mathbf{X}_1$ to avoid numerical problems. For short chromosome segments, the resulting marker effect estimates ($\hat{\boldsymbol{\beta}}_1$) are almost identical to the true population effects ($\boldsymbol{\beta}_1$) because the response used to derive $\hat{\boldsymbol{\beta}}_1$ ($g_{i_1} = \mathbf{z}'_{i_1}\boldsymbol{\alpha}$) is not affected by errors and the sample size used vastly exceeded the number of markers. For each MC replicate, we estimated the within and across R-squared parameters ([5] and [6]) using data not used to derive marker effects by squaring the correlation of the marker and QTL predictions: $R^2_{1\to1} = \mathrm{Corr}(\mathbf{x}'_{i_1}\boldsymbol{\beta}_1, \mathbf{z}'_{i_1}\boldsymbol{\alpha})^2$ [5] and $R^2_{1\to2} = \mathrm{Corr}(\mathbf{x}'_{i_2}\boldsymbol{\beta}_1, \mathbf{z}'_{i_2}\boldsymbol{\alpha})^2$ [6]. For each segment, we conducted 300 MC replicates (each time resampling QTL positions and their effects) and reported the average (across MC replicates) R-squared and RA values in the RA maps. A visual representation of the MC-ANOVA estimation algorithm can be found in Fig. 1.

MC-ANOVA sensitivity analysis. To demonstrate the robustness of MC-ANOVA to its main parameters, we re-estimated the RA maps in the AF UK Biobank cohort, first varying the number of QTL sampled for a

given segment (one, two, three, four, five, and six QTL per segment). Second, we varied the number of flanking SNPs to each side of the segment to be included in the MC-ANOVA estimation (zero, five, ten, 15, 20, and 30 flanking SNPs to each side). These were both evaluated in the chromosome segments discussed above.

Phenotype preprocessing

UK Biobank phenotypes. We evaluated six phenotypes in the UK Biobank cohort (Supplementary Table 2): height, HDL, serum urate, LDL, BMI, and serum glucose. Each phenotype was preadjusted using an ordinary least squares (OLS) regression including sex, age, the first five genotyped principal components, center, and batch. We used records from the first visit or, when the first instance was missing, the second visit. Serum urate was log-transformed before preadjustment.

The ARIC study phenotypes. We evaluated three phenotypes that were common between the ARIC study and those evaluated in our main UK Biobank-based analyses: height, serum urate, and BMI. The ARIC study phenotypes were preadjusted within each ancestry group using OLS regressions including sex and age. Serum urate was log-transformed before preadjustment. The ARIC study subjects were removed from the PGS analyses if they were missing the phenotype of interest, sex, or age.

Relative accuracy map validation for real traits

GWAS. For each preadjusted phenotype (Supplementary Table 2), we conducted a GWAS in the training set described above – distantly related individuals of EUR ancestry (n = 230,000) from the UK Biobank (Supplementary Table 4). Each GWAS (a single marker regression) was carried out using the R package BGData[42] (the rayOLS option). This uses a t-statistic with the null hypothesis that the SNP effect is zero (a two-sided test). The GWAS p-values were used as a filtering step for the subsequent PGS, in that a SNP was included in the PGS if it had a p-value < 1e-5. Note that when referring to a GWAS hit, as in Fig. 5, we used the standard cutoff of p-value < 5e-8 for consistency with other literature.

SNP effects for polygenic scores (PGS) using real data. For each phenotype, effects ($\hat{\mathbf{b}}_1$) for the GWAS-filtered SNPs were estimated with a Bayesian shrinkage variable-selection method (BayesC[25], a mixture prior consisting of a point of mass at zero and a Gaussian slab). These models were fit using the BLRXy function from the R package BGLR[43], which generates posterior samples using a Gibbs sampler[44]. We estimated SNP effects using 50,000 posterior samples collected using five MCMC chains. SNP effects were averaged over the chains.

PGS prediction. For each phenotype, we computed PGS for each subject in each testing set (ancestry group 1 = EUR and 2 = AF, CR, EA, or SA) using $\hat{y}_{i_*} = \mathbf{x}'_{\mathbf{i}_*}\hat{\mathbf{b}}_1$, where * denotes group 1 or 2. The PGS prediction correlation was then defined as $\mathrm{Corr}(\hat{y}_{i_*}, y_{i_*})$, where y_{i_*} is the adjusted phenotype of the *ith* subject of the corresponding testing group. The empirical RA was then defined as $\mathrm{RA} = \frac{\mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2}{\mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2}$, where the numerator is the squared PGS correlation for a cross-ancestry PGS (e.g., 2 = AF, CR, EA, or SA), and the denominator is that for within-ancestry (1 = EUR). Comparing this empirical RA to the MC-ANOVA predicted RA, $\mathrm{RA} = \frac{R^2_{1\to 2}}{R^2_{1\to 1}}$ [7], we can also define the loss of accuracy[7] (LOA) percentage attributable to allele frequency and LD differences between ancestries: $\mathrm{LOA}\,\% = \frac{1-\text{predicted RA}}{1-\text{empirical RA}} \times 100$.

Standard error estimates. We obtained approximate standard error estimates for the PGS correlation coefficients, $\mathrm{Corr}(\hat{y}_{i_*}, y_{i_*})$, using $\sqrt{\frac{1-\mathrm{Corr}\left(\hat{y}_{i_*}, y_{i_*}\right)^2}{n_*-2}}$, where n_* is the sample size of the given testing set (* = 1 or 2). The standard error of the correlation difference between two ancestries (e.g., 1 = EUR and 2 = AF), $\mathrm{Corr}(\hat{y}_{i_1}, y_{i_1}) - \mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})$, was computed as $\sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$. Following Wang et al.[7], the standard error for the empirical RA was computed as $\mathrm{SE}(\text{empirical RA}) = \sqrt{(\text{empirical RA})^2 * \left(\frac{4(1-\mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2)}{n_1 * \mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2} + \frac{4(1-\mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2)}{n_2 * \mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2}\right)}$. A similar method was used to obtain standard errors for the predicted RA, with the addition of an MC error component. More details of this can be found in the Supplementary Methods.

PGS subscores. To validate the MC-ANOVA method, we computed four PGS subscores for each trait and ancestry group based on the MC-ANOVA cross-ancestry R-squared estimates [6] from the RA maps. For one ancestry group and trait, the High PGS subscore consisted of the SNPs in the PGS that were in the top 20th percentile of $R^2_{1\to 2}$ [6]. Similarly, the Medium subscore was the 60th-80th percentile SNPs, the Low the 50th−60th, and the Very Low the bottom 50th. The PGS correlations described above, $\mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})$, were then computed within each of those SNP sets. Note that in Supplementary Table 4, the trait-specific proportion of variance explained by the EUR ancestry-derived PGS was computed from the overall PGS R-squared (using all PGS SNPs). In the benchmark analysis described next, PGS subscores were computed in the same way as MC-ANOVA, with SNP sets for PGS subscores based on the quantiles of the respective method's RA map (Fst or Wang et al.'s RA). To obtain a simulated null distribution for the expected correlation difference based on the number of SNPs included in each PGS in Fig. 4 and Supplementary Figs. 9, 11, 12, 14b, we permuted the grouping labels over 2000 iterations for each trait and ancestry group and estimated the PGS correlation difference between EUR and non-EUR within each permuted grouping.

Benchmarks

We benchmarked MC-ANOVA against Fst[24] and the RA method described in Wang et al.[7]. Both of these benchmark RA methods were evaluated in the same SNP segments described above (which were defined based on a minimum length of ten Kbp and at least ten SNPs) to build cross-ancestry RA maps for MC-ANOVA, ultimately building RA maps for each benchmark method as well.

Fixation index (Fst). Derived from Wright's F-statistic, Fst[24] has been the traditional metric used in population genetics to quantify genome differentiation in terms of allele frequency differences between populations. For a given locus, Fst decomposes the genetic variance as the proportion of between-population variation out of the total population variation, such that a value of zero corresponds to no differentiation between the populations. We computed the Fst for the *qth* window as the average Fst of all core SNPs in that segment, where the Fst for a single SNP is:

$$\frac{\left(\left(p1\cdot\frac{n1}{n1+n2}+p2\cdot\frac{n2}{n1+n2}\right)\cdot\left(1-\left(p1\cdot\frac{n1}{n1+n2}+p2\cdot\frac{n2}{n1+n2}\right)\right)\right)-\left(\frac{n1}{n1+n2}\cdot p1\cdot(1-p1)+\frac{n2}{n1+n2}\cdot p2\cdot(1-p2)\right)}{\left(p1\cdot\frac{n1}{n1+n2}+p2\cdot\frac{n2}{n1+n2}\right)\cdot\left(1-\left(p1\cdot\frac{n1}{n1+n2}+p2\cdot\frac{n2}{n1+n2}\right)\right)} \quad (8)$$

where p_* is the minor allele frequency and n_* is the sample size for population*.

Wang et al. RA method. The second RA method was described by Wang et al.[7] to quantify the proportion of prediction accuracy loss across ancestries attributable to allele frequency and LD differences. We modified Wang et al.'s method to make it trait-invariant. For each core SNP j (i.e., those in a chromosome segment, excluding the SNPs in flanking regions) in a single segment (see the section 'Mapping the relative accuracy (RA) of cross-ancestry PGS prediction' for segment details), we computed the SNPs in pairwise LD ($R^2 \geq .45$) from SNPs in the core or buffer of that window. The local RA of Wang et al. for SNP j was then defined by:

$$\left(\frac{\overline{r_{1,j}r_{2,j}}\sqrt{\frac{p_{2,j}(1-p_{2,j})}{p_{1,j}(1-p_{1,j})}}}{r_{1,j}^2}\right)^2 \cdot \frac{p_{1,j}(1-p_{1,j})}{p_{2,j}(1-p_{2,j})} \quad (9)$$

Here, $p_{*,j}$ is the allele frequency for the *jth* SNP, and $r_{*,j}$ is the mean correlation between the *jth* SNP and the SNPs in pairwise LD with it, for ancestry group * = 1,2 (for this analysis 1 = EUR and 2 = AF). The overall RA estimated for a segment by Wang et al. is the average of [9] over each core SNP j in the segment.

Validation in the ARIC study

RA maps were developed using the UK Biobank EUR training set and the data from the AEA and AAA participants from the ARIC study for external validation. For this validation, the MC-ANOVA procedure was carried out as described above for the UK Biobank (a minimum segment length of ten Kbp and at least ten SNPs, and three QTL), and 65,525 nonoverlapping SNP segments (an average of 36 Kbp and containing 12 core SNPs) were defined for the RA maps. Global predicted RA, empirical RA, and LOA were estimated for height, serum urate, and BMI. First, portability measures (cross-ancestry R-squared [6] and predicted RA [7]) were estimated within each segment with MC-ANOVA. In this case, predicted RA is defined as $R^2_{\mathrm{EUR}\to\mathrm{AAA}}/R^2_{\mathrm{EUR}\to\mathrm{AEA}}$ [7], where EUR is the UK Biobank EUR training set, AAA is the ARIC study African American testing set, and AEA is the ARIC study European American testing set. Similarly, the global PGS (using all SNPs meeting the GWAS p-value threshold of 1e-5) was evaluated for each trait. The same procedure as above was used to estimate SNP effects (see 'SNP effects for polygenic scores (PGS) using real data'), which are derived from the UK Biobank EUR training set: $\hat{\mathbf{b}}_1$. Then, the PGS prediction is $\hat{y}_{i*} = \mathbf{x}'_{i*}\hat{\mathbf{b}}_1$, for * = 1 or 2 now denoting either AEA (within-ancestry) or AAA (cross-ancestry), respectively. The PGS correlation calculation, $\mathrm{Corr}(\hat{y}_{i*}, y_{i*})$, was also the same as above for the UK Biobank (* = 1 [AEA] or 2 [AAA]; see 'PGS prediction'). Thus, empirical RA in this case was computed as $\frac{\mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2}{\mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2}$. When evaluating the RA map validation estimating PGS subscores based on SNP groups defined by the RA maps, the correlation difference was computed as $\mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2 - \mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2$, and the portability groupings were based on the same $R^2_{1\to 2}$ [6] quantiles as for the UK Biobank (see 'PGS subscores').

Integrating RA maps into a gradient descent algorithm

Gradient descent with early stopping (GD-ES) is an approach commonly used for transfer learning in machine learning algorithms. Recently, Zhao et al.[29] proposed using GD-ES to build PGS for cross-ancestry prediction. In Zhao's GD-ES algorithm, effects are estimated by minimizing a residual sum of squares evaluated in a data set (D2) from a target population (e.g., African ancestry), using an iterative procedure that uses an external estimator ($\hat{\boldsymbol{\beta}}_1$ derived from D1 of, e.g., European ancestry) as the initial value. Thus, GD-ES produces a sequence of estimates, $\{\tilde{\boldsymbol{\beta}}_{2(0)}, \tilde{\boldsymbol{\beta}}_{2(1)}, \ldots, \tilde{\boldsymbol{\beta}}_{2(s)}\}$, starting with $\tilde{\boldsymbol{\beta}}_{2(0)} = \hat{\boldsymbol{\beta}}_1$ (pure cross-ancestry prediction) and moving toward the solution that one would obtain only using D2 ($\hat{\boldsymbol{\beta}}_2$) after s iterations. Early stopping of the GD algorithm renders estimates that are a compromise between $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ and have been shown to improve cross-ancestry PGS prediction compared to using either a purely external ($\hat{\boldsymbol{\beta}}_1$) or a purely internal ($\hat{\boldsymbol{\beta}}_2$) estimate[29]. We extended this approach by allowing for a SNP-specific learning rate (LR) that is based on MC-ANOVA relative accuracy estimates.

In a GD algorithm, coefficients are updated one at a time using $\beta_{2j}^{\mathrm{new}} = \beta_{2j}^{\mathrm{current}} - \mathrm{LR} \times \mathrm{d}L/\mathrm{d}\beta_{2j}$, where LR is a learning rate parameter (controlling how fast the algorithm moves in the direction that minimizes the loss function, in our case the residual sum of squares loss function) and $\mathrm{d}L/\mathrm{d}\beta_{2\mathrm{j}}$ is the gradient of the loss function with respect to the *jth* coefficient of $\boldsymbol{\beta}_2$. In Zhao et al.[29], the same LR was used for all SNPs. We modified the algorithm by introducing an adaptive (SNP-specific) LR: $\mathrm{LR_j} = 0.01 \times e^{-3R^2_{1\to 2,j}}$, where $R^2_{1\to 2,j}$ is the estimate presented in equation [6]. With this approach, a SNP with a high MC-ANOVA cross-ancestry R-squared estimate will have a low learning rate, staying closer to the initial external estimate ($\hat{\beta}_{1j}$) and a SNP with a low $R^2_{1\to 2,j}$ will have a higher learning rate, thus moving further away from the EUR-derived estimated effect.

For a EUR ancestry group effect, $\hat{\boldsymbol{\beta}}_1$, we used the same PGS effects ($\hat{\mathbf{b}}_1$) described above (see the Methods section 'SNP effects for polygenic scores (PGS) using real data'). This was then employed as an initial value in a gradient descent algorithm run on data from either AF, SA, or CA ancestry from the UK Biobank (the EA group was excluded due to the small sample size for this group). To obtain an unbiased estimate of the out-of-sample R-squared, we split the data into training and testing sets (n-testing=300). We then conducted a five-fold cross-validation within the training data to select the optimal number of iterations of the GD algorithm (which acts as the parameter controlling how much effects are shrunk towards the initial values). Then, we ran the GD algorithm with that number of iterations on the entire training data and used the resulting effects to predict in the excluded testing data. This was repeated 50 times, each time with a different random partition of training and testing. The average results for the 18 trait-ancestry group combinations are reported in Supplementary Table 8. The adaptive (SNP-specific) learning rate was compared to using a fixed learning rate, which was the mean of the adaptive learning rate for each trait-ancestry group pair. Additionally, for each trait-ancestry pair, we compute the percentage of times (across training-testing partitions) for which the prediction R-squared for the adaptive learning rate method compared to the fixed learning rate is higher (Supplementary Table 8), excluding partitions that had identical R-squared. The R-code implementing the GD algorithm is included in the GD.R function in the GitHub repository https://github.com/lupiA/MCANOVA.

Genetic distance

The genetic distance reported between the ancestry groups in Table 1 and Supplementary Fig. 1 was computed as the overall (genome-wide) Fst[24] between pairwise ancestries using PLINK (v1.90b6.24)[45]: --fst –within. We used a random sample of 20,000 individuals from the EUR ancestry group.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The relative accuracy maps generated in this study have been deposited in the Zenodo database at https://doi.org/10.5281/zenodo.

13769713 and are provided as Supplementary Data. The GWAS summary statistics are available through Zenodo at https://doi.org/10.5281/zenodo.13785877. The UK Biobank data is available under restricted access and access can be obtained by applying at https://www.ukbiobank.ac.uk/. The ARIC Study data is available from dbGaP (https://www.ncbi.nlm.nih.gov/gap/) under accession code phs000280.v3.p1. The raw UK Biobank and the ARIC study data are protected and are not available due to data privacy laws. The protocol and consent were approved by the UK Biobank's Research Ethics Committee and were conducted under the application number 15326. Data from the ARIC study usage was approved by Michigan State University's Institutional Review Board under Study ID LEGACY15-745. Source data for Figures are provided with this paper. Source data are provided with this paper.

Code availability

The software presented and described in this study (the MC-ANOVA algorithm, a function to obtain the chromosome segments, the portability maps, and an interactive Shiny App) along with examples of how to use the MC-ANOVA algorithm can be found in an R package described and installable from https://github.com/lupiA/MCANOVA (Zenodo: https://doi.org/10.5281/zenodo.13769713). An identical web-based Shiny app is also available at https://lupia.github.io/Cross-Ancestry-Portability/ (Zenodo: https://doi.org/10.5281/zenodo.13769723) which will run slower than the R package app but does not require R software or package installation.

References

1. Sollis, E. et al. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
2. Lello, L. et al. Accurate genomic prediction of human height. *Genetics* **210**, 477–497 (2018).
3. Kim, H., Grueneberg, A., Vazquez, A. I., Hsu, S. & de los Campos, G. Will big data close the missing heritability gap? *Genetics* **207**, 1135–1145 (2017).
4. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
5. Dikilitas, O. et al. Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *Am. J. Hum. Genet.* **106**, 707–716 (2020).
6. Scutari, M., Mackay, I. & Balding, D. Using genetic distance to infer the accuracy of genomic prediction. *PLOS Genet.* **12**, e1006288 (2016).
7. Wang, Y. et al. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
8. Privé, F. et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).
9. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
10. Belsky, D. W. et al. Development and evaluation of a genetic risk score for obesity. *Biodemogr. Soc. Biol.* **59**, 85–100 (2013).
11. Domingue, B. W., Belsky, D., Conley, D., Harris, K. M. & Boardman, J. D. Polygenic influence on educational attainment: new evidence from The National Longitudinal Study of Adolescent to Adult Health. *AERA Open* **1**, 1–13 (2015).
12. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
13. Vassos, E. et al. An examination of polygenic score risk prediction in individuals with first-episode psychosis. *Biol. Psychiatry* **81**, 470–477 (2017).
14. Li, Z. et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).
15. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
16. Veturi, Y. et al. Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. *Genetics* **211**, 1395–1407 (2019).
17. Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hum. Genet. Genomics Adv.* **2**, 100017 (2021).
18. Hou, K. et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).
19. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
20. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129**, 687–702 (1989).
21. de los Campos, G., Sorensen, D. & Gianola, D. Genomic heritability: what is it? *PLOS Genet.* **11**, e1005048 (2015).
22. de los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C. & Sorensen, D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* **9**, e1003608 (2013).
23. de los Campos, G. et al. ANOVA-HD: Analysis of variance when both input and output layers are high-dimensional. *PloS One* **15**, e0243251 (2020).
24. Wright, S. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**, 395–420 (1965).
25. Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinforma.* **12**, 186 (2011).
26. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
27. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).
28. Qiao, Z. et al. Estimation and implications of the genetic architecture of fasting and non-fasting blood glucose. *Nat. Commun.* **14**, 451 (2023).
29. Zhao, Z., Fritsche, L. G., Smith, J. A., Mukherjee, B. & Lee, S. The construction of cross-population polygenic risk scores using transfer learning. *Am. J. Hum. Genet.* **109**, 1998–2008 (2022).
30. VanLiere, J. M. & Rosenberg, N. A. Mathematical properties of the measure of linkage disequilibrium. *Theor. Popul. Biol.* **74**, 130–137 (2008).
31. Bulmer, M. G. The effect of selection on genetic variability. *Am. Nat.* **105**, 201–211 (1971).
32. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
33. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323 (2013).
34. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
35. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
36. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000

individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* **8**, 21 (2017).
37. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
38. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).
39. UK Biobank - UK Biobank. https://www.ukbiobank.ac.uk/.
40. International HapMap 3 Consortium *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
41. HapMap 3. *Broad Institute* https://www.broadinstitute.org/medical-and-population-genetics/hapmap-3 (2008).
42. Grueneberg, A. & de los Campos, G. BGData - A suite of R packages for genomic analysis with big data. *G3amp58. GenesGenomesGenetics* **9**, 1377–1383 (2019).
43. Pérez, P. & de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**, 483–495 (2014).
44. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).
45. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Acknowledgements

Data from the UK Biobank was acquired from application number 15326 and data from the ARIC study was acquired through dbGaP under accession code phs000280.v3.p1 and project number 9191. We would like to thank the participants and those who developed the UK Biobank and ARIC data sets, as well as Michigan State University and the Institute for Cyber-Enabled Research at Michigan State University for providing funding and computing resources, respectively. We also thank Wen Huang for the comments provided when A.L. presented the preliminary results of this study. The authors received funding from NIH grants R01DK119836 (A.L., G.D.L.C., and A.V.), R03HG011674 (A.L., G.D.L.C., and A.V.), and R01HG013794 (A.L., G.D.L.C., and A.V.).

Author contributions

Conceptualization: A.L., G.D.L.C., and A.V.; Methodology: A.L. and G.D.L.C.; Software Development: A.L. and G.D.L.C.; Investigation: A.L. and G.D.L.C.; Writing – Original Draft: A.L.; Writing – Review & Editing: A.L., G.D.L.C., and A.V.; Project Administration and Supervision: A.V. and G.D.L.C.; Funding Acquisition: A.L., G.D.L.C., and A.V.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54727-8.

Correspondence and requests for materials should be addressed to Alexa S. Lupi or Gustavo de los Campos.

Peer review information *Nature Communications* thanks Clive Hoggart, Doug Speed, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.