nature communications

Article

https://doi.org/10.1038/s41467-023-38930-7

Optimal strategies for learning multi-ancestry polygenic scores vary across traits

Received: 7 April 2022

Accepted: 22 May 2023

Published online: 07 July 2023

Check for updates

Brieuc Lehmann[1], Maxine Mackintosh[2,3], Gil McVean[4,6] & Chris Holmes[3,4,5,6]

Polygenic scores (PGSs) are individual-level measures that aggregate the genome-wide genetic predisposition to a given trait. As PGS have predominantly been developed using European-ancestry samples, trait prediction using such European ancestry-derived PGS is less accurate in non-European ancestry individuals. Although there has been recent progress in combining multiple PGS trained on distinct populations, the problem of how to maximize performance given a multiple-ancestry cohort is largely unexplored. Here, we investigate the effect of sample size and ancestry composition on PGS performance for fifteen traits in UK Biobank. For some traits, PGS estimated using a relatively small African-ancestry training set outperformed, on an African-ancestry test set, PGS estimated using a much larger European-ancestry only training set. We observe similar, but not identical, results when considering other minority-ancestry groups within UK Biobank. Our results emphasise the importance of targeted data collection from underrepresented groups in order to address existing disparities in PGS performance.

Polygenic scores (PGS) are composite, quantitative measures that aim to predict complex traits from genetic data. As well as providing insights into the genetic architecture of complex traits, PGS have considerable clinical potential for screening and prevention strategies[1,2]. Largely driven by significant increases in sample sizes, the predictive utility of PGS has improved substantially in recent years for a variety of traits[3], including cardiovascular disease[4], breast cancer[5], and type I diabetes[6].

These improvements, however, have largely been limited to populations of European ancestry[7–9], reflecting the lack of diversity in genomic samples collected to date[9]. Moreover, predictive performance decreases with genetic distance from the training population[10–12]. The lack of transferability of PGS across ancestries may be due to a number of factors, including population differences in allele frequencies and linkage disequilibrium (LD)[9,11]. While there is some evidence that common causal variants have similar effects across ancestries[9,13], other studies have suggested that the underlying variant effects may in fact differ across ancestries[14–16], which may be due to gene-by-gene or gene-by-environment (GxE) interactions[17,18]. In addition, GxE interactions for people of African ancestry may be different between those living in the UK and those living in South Africa, say. This lack of transferability raises one of the most important technical and ethical challenges in the clinical utility and applications of PGS due to their potential to exacerbate health inequalities[9].

There are major, ongoing initiatives to collect genomic data from traditionally under-represented groups, such H3Africa[19], that aim to address the lack of global genetic diversity in research data. However, it may take many years to collect sufficient data to reduce the disparities in PGS performance. Statistical methods may provide an alternative, short-term, cost-effective and complementary potential solution to mitigate against the negative effects of the lack of diversity in genomic datasets, by using modelling techniques to make use of all the existing data available, while allowing for some differences between groups.

There has been a growing interest in statistical methods to improve the transferability of PGS, which have thus far focused on

[1]Department of Statistical Science, University College London, London, UK. [2]Genomics England, London, UK. [3]The Alan Turing Institute, London, UK. [4]Big Data Institute, University of Oxford, Oxford, UK. [5]Department of Statistics, University of Oxford, Oxford, UK. [6]These authors contributed equally: Gil McVean, Chris Holmes. e-mail: b.lehmann@ucl.ac.uk

GWAS-derived PGS, i.e. PGS based on summary statistics from a genome-wide association study (GWAS). These integrate results from GWAS trained on distinct populations at different stages of the PGS estimation. For example, Grinde et al. use European GWAS results to select variants, and then estimate the variant weights using the non-European GWAS results[20]. Márquez-Luna et al. propose a 'multi-ethnic' PGS by combining scores trained separately on different populations[21]. Weissbrod et al. expand on this approach by further leveraging functionally informed fine-mapping to estimate the PGS weights[22]. In a related approach, Ruan et al. incorporate LD information to directly estimate effect sizes using GWAS results from two or more populations[23]. To account for admixture, Cavazos & Witte propose local ancestry weighting to construct individualised PGS[24]. These efforts have yielded promising improvements in PGS performance, though there remains a significant gap in predictive accuracy between European and non European target populations.

Here, we investigate the use of multiple-ancestry datasets, such as UK Biobank[25], to estimate PGS for a range of anthropometric, blood-sample, and disease traits, with the explicit objective of improving predictive accuracy for under-represented ancestries. Specifically, we ask whether there are consistent optimal strategies for borrowing information across ancestry groups to maximize prediction accuracy in groups that have small sample sizes in available resources. For each trait, we construct training sets with varying numbers of individuals from each ancestry to assess the effect of sample size and composition on PGS accuracy, using both simulated data and data from UK Biobank[25]. Moreover, and to counteract the imbalance of ancestries in a multiple-ancestry training set, we investigate the use of an importance reweighting approach that places more weight on under-represented ancestries during training. Importance reweighting is a standard statistical technique used in survey sampling that aims to account for differences between the sampling population and the population of interest[26]. Given the availability of individual-level information in biobanks, we estimate PGS using regularised regression applied to full genotype and covariate data (as opposed to genome-wide association summary statistics) to avoid introducing additional artefacts into our analysis via the reliance on assumptions about genotype and covariate correlation structure (including LD) and GWAS methodology.

Our results show that the impact of sample size and composition on predictive performance is highly variable across traits. For some traits, polygenic scores estimated using a relatively small number of minority-ancestry individuals outperformed on a minority-ancestry test set scores estimated using a much larger number of European-ancestry individuals. Moreover, adding European-ancestry individuals to the training set did not always improve performance and in some cases even led to poorer performance. Although importance weighting yields moderate improvement in performance for some traits, we find that sample size is a much more prominent factor, highlighting the limitations of statistical corrections and the importance of collecting more data from a more diverse range of participants.

Results

Overview of methods

To investigate the effect of sample size and ancestry composition on polygenic score performance, we used both simulated data and real data from UK Biobank[25]. We initially focus on PGS strategy for African-ancestry groups, given that predictive accuracy using European-ancestry PGS is consistently worst out of all the major ancestry groups[9,12]. Simulated data was generated using the simulation engine `msprime`[27] in the standard library of population genetic simulation models `stdpopsim` v0.1.2[28] to generate African-ancestry and European-ancestry genotypes. We also considered a range of quantitative and binary traits from UK Biobank, using imputed genotypes along with inferred genetic ancestry labels made available by the Pan-UKBB initiative[29].

For both the simulated and real data, we constructed a range of training sets controlling the number of European-ancestry and minority-ancestry individuals. We considered three types of training sets: a single-ancestry set consisting only of European-ancestry individuals, a single-ancestry set consisting of minority-ancestry individuals, and a dual-ancestry set consisting of both European-ancestry and minority-ancestry individuals. We denote a PGS trained on a XYZ-ancestry training set as PGS_{XYZ}—for example PGS_{EUR}. Similarly, we refer to a PGS trained on a dual-ancestry (respectively minority-ancestry) training set as PGS_{dual} (respectively PGS_{min}).

For each training set, we estimated PGS using L1-regularised regression, also known as the LASSO[30], which has previously been used in the context of genetic risk prediction (see for example[31–34]). To account for the imbalance in sample size numbers in the dual-ancestry training sets, we also estimated PGS using an importance reweighted LASSO, upweighting the non-European-ancestry individuals. Following Martin et al.[9], we assessed the predictive performance of a PGS using partial r^2 relative to a covariate-only model. See Fig. 1 for a schematic diagram of the methods used and "Methods" for full details.

Single-ancestry PGS can outperform dual-ancestry PGS despite being trained on fewer individuals

We first set out to evaluate the relative performance of single-ancestry PGS versus dual-ancestry PGS through simulation. An important factor in the lack of transferability of PGS across ancestries is the difference in causal effect sizes[11], with the correlation of causal effect sizes between ancestry groups, ρ, has been estimated to be significantly less than one across a range of common traits[15,16]. We simulated traits by randomly selecting $p_0 = 100$ SNPs to be causal and then generating effect sizes for these SNPs such that that overall heritability across the entire population was $h_2 = 0.3$. To assess the impact of differences in causal effect sizes on the relative predictive performance of PGS strategies, we varied the trans-ancestry causal effect correlation, $\rho = 0.5, 0.6, \ldots, 1.0$, generating 10 quantitative traits for each value. We assume that all causal variants are genotyped, thus avoiding any differences in PGS accuracy that may arise from imperfect tagging[11]. We note that in practice imputation does not fully address the issue of imperfect tagging, due to differences in imputation quality across ancestry groups[7,9].

For each trait we created five African-ancestry training sets with $n_{AFR} = 2000, 4500, 7714, 12{,}000, 18{,}000$ African-ancestry individuals. We also created five dual-ancestry training sets made up of the corresponding African-ancestry training set supplemented with $n_{EUR} = 18{,}000$ European-ancestry individuals. To obtain PGS from the African-ancestry training sets, we used unweighted LASSO regression, while for the dual-ancestry training sets, we used the importance reweighted LASSO with $\gamma = 0, 0.1, \ldots, 1.5$. The case $\gamma = 0$ corresponds to no reweighting, that is, the standard LASSO. The case $\gamma = 1$ corresponds to inverse proportion reweighting, so that in total, African-ancestry individuals and European-ancestry individuals have equal weight. Note that the African-ancestry training set is equivalent to the limiting case $\gamma \to \infty$ whereby zero weight is placed on European-ancestry individuals. We quantified predictive performance of a PGS for a given ancestry group in terms of the predictive gap: the difference between variance explained by the PGS, r^2, and SNP heritability h^2 (the maximal variance explained by any PGS). See "Simulation study" for full details.

For both PGS_{AFR} and PGS_{dual}, the predictive gap for African-ancestry individuals decreased substantially as the number of African-ancestry individuals in the training sets increased (Fig. 2). For example, when $n_{AFR} = 2000$ and the correlation between genetic effects ρ was 0.7, the mean predictive gap of the unweighted PGS_{dual} ($\gamma = 0$) was just over 0.21 for the African-ancestry test sets, compared to 0.03 for the European-ancestry test sets. While the performance on Europeans did

https://doi.org/10.1038/s41467-023-38930-7

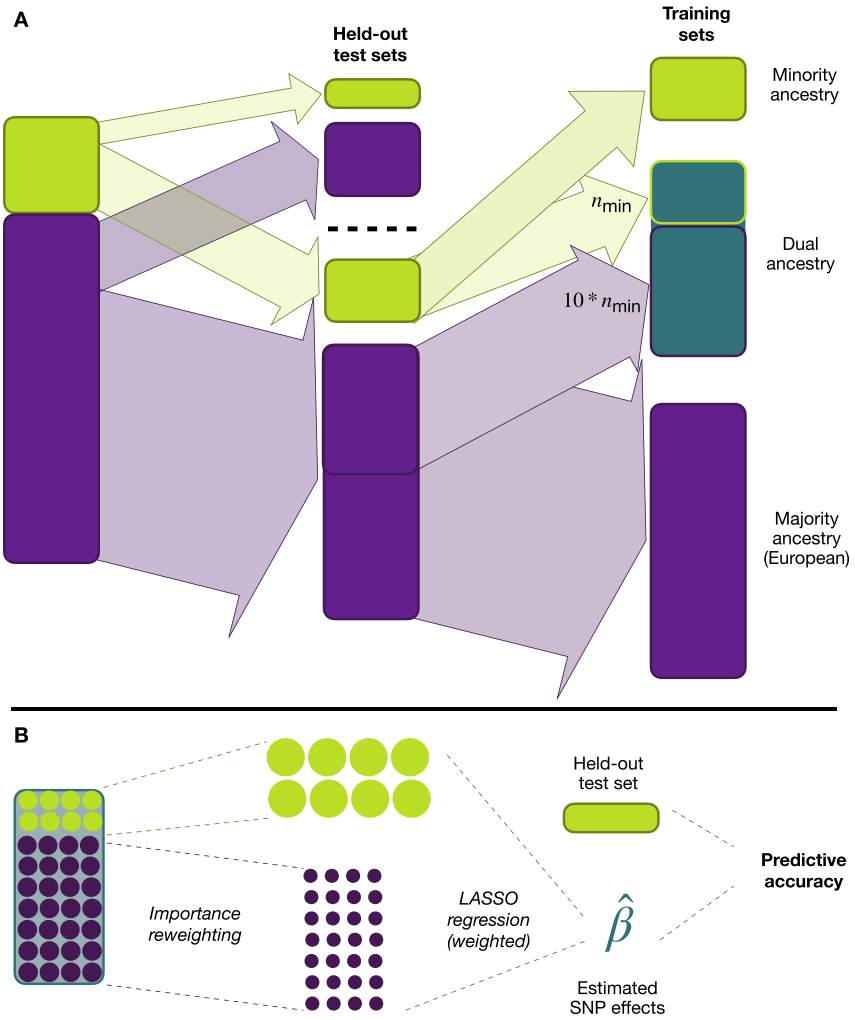

Fig. 1 | Overview of methods. A To evaluate the different PGSs, we performed various splits of the available data. Firstly, we held out test sets of 20% of individuals in each ancestry group. From the remaining 80%, we constructed three types of training sets: a single-ancestry set consisting only of European-ancestry individuals (purple block), a single-ancestry set consisting of non-European-ancestry individuals (yellow block), and a dual-ancestry set consisting of both European-ancestry and non-European-ancestry individuals (blue block). For each training set, we used another 20% of the data to select the regularisation parameter in the LASSO. **B** For the dual-ancestry training set, we used an importance weighted LASSO, assigning higher weights to individuals in the minority-ancestry group. See Methods for full details.

not change markedly as the number of African-ancestry training set individuals increased, the African-ancestry predictive gap fell to ~0.11 and 0.06 when $n_{AFR} = 7\,714$ and $n_{AFR} = 18{,}000$ respectively. As expected, the discrepancy between the two ancestry groups decreased as the correlation ρ between the ancestry-specific genetic effects increased.

In the absence of reweighting ($\gamma = 0$), PGS_{AFR} generally outperformed PGS_{dual}, despite the latter being trained on more individuals—the dual-ancestry training sets consist of the African-ancestry training sets along with 18,000 European-ancestry individuals. This was particularly evident when the correlation between ancestry-specific genetic effects was lower ($\rho = 0.5$).

Importance reweighting generally had a positive impact on predictive performance, reducing the difference between PGS_{AFR} and PGS_{dual}. When the number of African-ancestry training individuals was low ($n_{AFR} = 2000$), the importance-reweighted PGS_{dual} outperformed PGS_{AFR} when the correlation between ancestry-specific genetic effects was relatively high ($\rho \geq 0.7$). When $\rho = 0.7$, the relative amount of improvement with reweighting was stable as n_{AFR} increased: for $n_{AFR} = 2000$, the predictive gap was reduced by 27% (from 0.21 at $\gamma = 0$, to 0.15 at $\gamma = 1.4$), while for $n_{AFR} = 18{,}000$, the predictive gap was reduced by 29% (from 0.060 at $\gamma = 0$, to 0.042 at $\gamma = 1.4$). In contrast, when $\rho = 1$, the relative amount of improvement with reweighting was decreased as n_{AFR} increased: for $n_{AFR} = 2000$, the predictive gap was reduced by 16% (from 0.14 at $\gamma = 0$, to 0.11 at $\gamma = 0.5$), while for $n_{AFR} = 18{,}000$, the predictive gap was only reduced by 7% (from 0.040 at $\gamma = 0$, to 0.038 at $\gamma = 0.4$). We highlight that importance reweighting reduced the predictive gap even when the effect sizes were the same across ancestries ($\rho = 1$), indicating that the procedure can partially correct for ancestry-specific differences in allele frequencies and LD structure.

The effect on predictive performance depended on the degree of reweighting, quantified by γ. As γ increased from 0 to 1.5, the predictive gap typically decreased. When the correlation between ancestry-specific genetic effects was high ($\rho \geq 0.9$) and the number of African-ancestry samples low ($n_{AFR} = 2000$), the predictive gap increased again for larger values of γ. This reflects a crucial trade-off of importance reweighting in this context: while the bias of African-ancestry genetic effect estimates may be lower with more reweighting, the increased variance of the weights results in a lower effective sample size.

https://doi.org/10.1038/s41467-023-38930-7

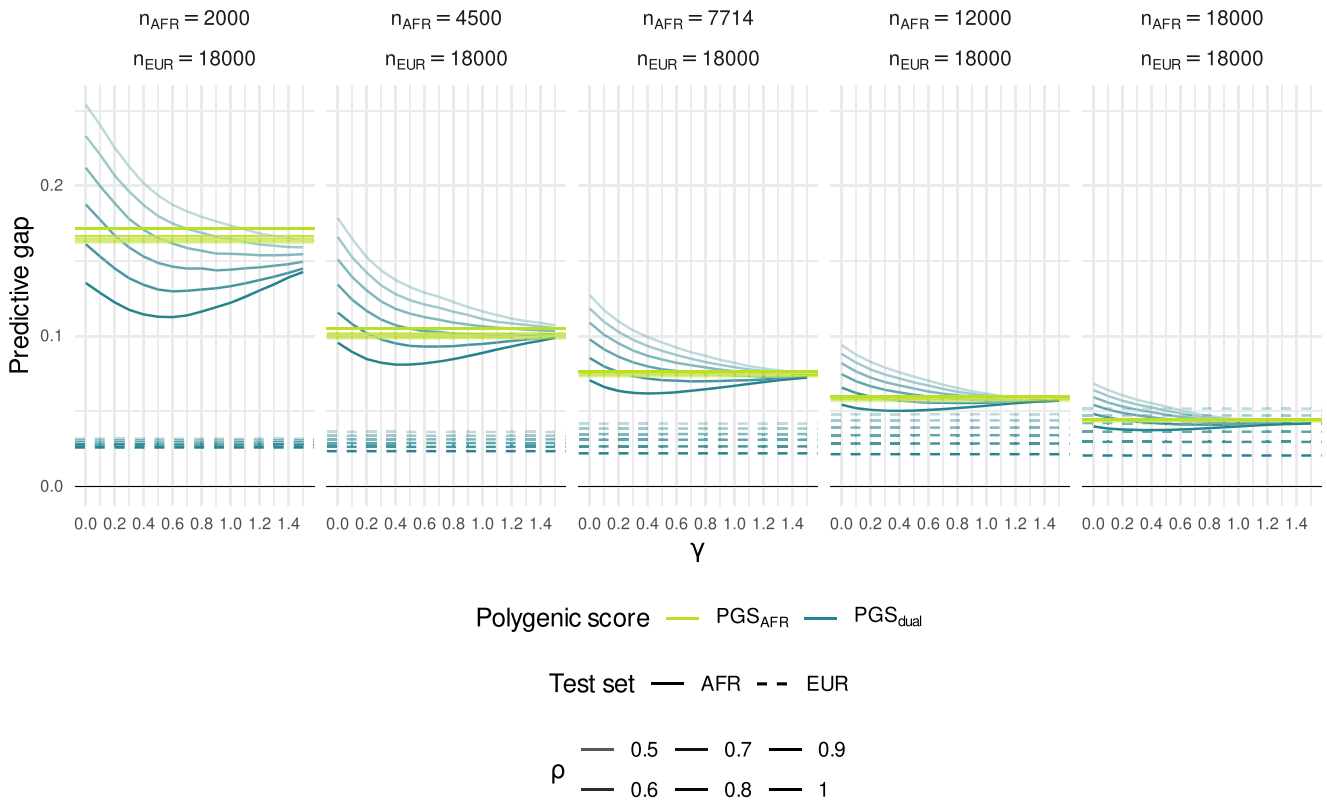

Fig. 2 | Simulation study: predictive gap against number of African-ancestry individuals in training set. Each panel corresponds to a different number of African-ancestry training set individuals from $n_{AFR} = 2000$ to $n_{AFR} = 18{,}000$. The training sets for PGS_{dual} (blue lines) consisted of the corresponding African-ancestry training set for PGS_{AFR} (yellow lines), along with $n_{EUR} = 18{,}000$ European-ancestry individuals. Each line represents the mean predictive gap across 50 repetitions. The horizontal dashed lines correspond to the predictive gap for European-ancestry (EUR) test sets based on an unweighted LASSO, while the solid lines correspond to the predictive gap for African-ancestry (AFR) test sets. The parameter γ corresponds to the degree of reweighting used in the reweighted LASSO for PGS_{dual}. The correlation of genetic effects between ancestries ρ was varied from 0.5 (lighter lines) to 1 (darker lines).

The effect of reweighting was relatively small compared to the effect of increased sample size. For $n_{AFR} = 2000$ and $\rho = 0.7$, reweighting reduced the predictive gap from 0.21 to 0.15. The same improvement for the unweighted PGS_{dual} was seen when increasing the number of African-ancestry individuals to $n_{AFR} = 4500$, though we note that the PGS_{AFR} with $n_{AFR} = 4500$ had an even lower predictive gap of 0.10.

To investigate the impact of different genetic architectures on these results, we performed supplementary studies varying the heritability $h_2 = 0.1, 0.3, 0.5$, and the polygenicity (i.e. number of causal SNPs) $p_0 = 10, 100, 1000$ (Supplementary Figs. 1 and 2 respectively). We observed a similar pattern of results, with importance reweighting having a positive impact on predictive performance when the number of African-ancestry training individuals was low. For a fixed genetic correlation $\rho = 0.8$, the impact of reweighting was generally larger when heritability was high, and when the number of causal SNPs was lower. In particular, we note that the impact of reweighting was minimal when $h_2 = 0.1$. This is likely a reflection of the LASSO's reduced ability to estimate small effect sizes as opposed to large effect sizes.

This simulation study illustrates the relative performance of PGS_{AFR} and PGS_{dual} as a function of (i) the number of minority-ancestry individuals in the respective training sets and (ii) the correlation of genetic effects between ancestries. For both PGS_{AFR} and PGS_{dual}, the between-ancestry disparity in predictive performance decreased as the number of minority-ancestry individuals in the training set increased. Despite being trained on fewer individuals, the PGS_{AFR} tended to outperform PGS_{dual}, particularly when the correlation between ancestry-specific genetic effects was low. Importance reweighting generally reduced the predictive gap of PGS_{dual}, especially when the sample size imbalance between ancestries was high. As this imbalance decreased, however, the reweighted PGS_{dual} did not outperform PGS_{AFR}.

Adding individuals from one ancestry does not always improve PGS performance for a different ancestry

Next, we sought to examine whether, even without reweighting, prediction performance in an underrepresented ancestry group can be boosted by including individuals from a different ancestry among the training samples. To do so, we used data from UK Biobank, estimating a range of PGS based on varying numbers of training individuals for a variety of traits (Table 1; see Methods for a description of how the traits were selected). In this analysis we focused on African-ancestry and European-ancestry individuals, where genetic ancestry labels were taken from Pan-UKBB. Note that the number of individuals from each ancestry with non-missing trait values varied across traits (see Table 1). For example, there are 6178 African-ancestry individuals and 361,699 European-ancestry individuals with non-missing data for height, compared to 5748 African-ancestry individuals and 345,862 European-ancestry individuals with non-missing data for high light scatter reticulocyte count.

First, we held the number of European-ancestry individuals fixed, and constructed six training sets by varying the number of African-ancestry individuals from 0% to 80% of the total number of African-ancestry individuals with non-missing data for the trait in question. The number of European-ancestry individuals was chosen such that the largest training set had a 90-10 split between European- and African-ancestry individuals. Considering height again as an example, this corresponded to keeping the number of European-ancestry individuals fixed at 44,487 and varying the number of African-ancestry

https://doi.org/10.1038/s41467-023-38930-7

Table 1 | Number of individuals with non-missing trait value in UK Biobank by genetic ancestry group

Ancestry	Trait	Total	Cases	h^2
AFR	Height	6178		0.4603
AMR	Height	981		
CSA	Height	8082		0.5709
EAS	Height	2689		0.4415
EUR	Height	361699		
MID	Height	1543		
AFR	Mean corpuscular volume (MCV)	5912		0.4465
AMR	Mean corpuscular volume (MCV)	958		0.8142
CSA	Mean corpuscular volume (MCV)	7956		0.3716
EAS	Mean corpuscular volume (MCV)	2622		0.2414
EUR	Mean corpuscular volume (MCV)	351672		0.1215
MID	Mean corpuscular volume (MCV)	1513		0.7393
AFR	Asthma	6257	741	0.0255
AMR	Asthma	989	106	0.0610
CSA	Asthma	8285	1017	0.0340
EAS	Asthma	2700	236	0.0170
EUR	Asthma	362511	42248	0.0748
MID	Asthma	1567	174	0.0301
AFR	Female genital prolapse (FGP)	3665	93	0.1222
AMR	Female genital prolapse (FGP)	641	27	
CSA	Female genital prolapse (FGP)	3793	170	
EAS	Female genital prolapse (FGP)	1786	31	
EUR	Female genital prolapse (FGP)	194734	11043	0.0675
MID	Female genital prolapse (FGP)	662	33	
AFR	Body mass index (BMI)	6168		0.3838
EUR	Body mass index (BMI)	361306		0.1090
AFR	Eosinophill percentage	5893		0.0739
EUR	Eosinophill percentage	351038		0.0958
AFR	Erythrocyte distribution width	5912		0.1331
EUR	Erythrocyte distribution width	351672		0.0936
AFR	Lymphocyte count	5893		0.1896
EUR	Lymphocyte count	351033		0.0962
AFR	Monocyte count	5893		0.2451
EUR	Monocyte count	351033		0.1105
AFR	Mean platelet volume (MPV)	5912		0.3315
EUR	Mean platelet volume (MPV)	351669		0.2062
AFR	Platelet crit	5912		0.2090
EUR	Platelet crit	351670		0.1298
AFR	High light scatter reticulocyte count	5748		0.3758
EUR	High light scatter reticulocyte count	345862		0.0967
AFR	Atrial fibrillation (AFib)	6257	125	0.0797
EUR	Atrial fibrillation (AFib)	362511	16476	0.0880
AFR	Diverticular disease of the intestine (DDI)	6257	285	0.0247
EUR	Diverticular disease of the intestine (DDI)	362511	30413	0.0615
AFR	Hypothyroidism	6257	124	0.0308
EUR	Hypothyroidism	362511	17496	0.1243

SAIGE ancestry-specific heritability estimates[72] were obtained from the Pan-UKBB initiative website[29].

individuals from 0 to 4 942. For each training set, we used the unweighted LASSO to generate PGSs, and evaluated predictive performance in terms of variance explained on held-out test sets for each ancestry.

The predictive performance on African-ancestry individuals increased for five traits: height, mean corpuscular volume (MCV), female genital prolapse (FGP), high light scatter reticulocyte count (reticulocyte), and platelet crit (platelet) (Fig. 3a). For the remainder of the traits, predictive performance stayed largely stable.

Next, we performed complementary analyses in which we held the number of African-ancestry individuals fixed at 80% of the total number of African-ancestry individuals with non-missing data for the trait in question. We again constructed six training sets, now varying the number of European-ancestry individuals so that the proportion of European-ancestry individuals in the training set ranged from 0% to 90%.

Predictive performance on African-ancestry individuals only increased markedly for height and mean platelet volume (MPV), with the improvement for MPV appearing to tail off between the two largest training sets. For three traits (mean corpuscular volume, reticulocyte count, erythrocyte distribution width), predictive performance worsened as European-ancestry individuals were added to the training set. For the remainder of the traits, predictive performance stayed mostly constant.

These results demonstrate that, for some traits, the potential for increasing prediction performance by including samples from a different population can be limited. They also highlight substantial between-trait heterogeneity in response, suggestive of major differences in genetic architecture.

Optimal ancestry composition of training sets varies among traits in UK Biobank

Next, we set out to assess the relative performance of single- versus dual-ancestry PGS in empirical data by considering the UK Biobank traits analysed in the previous section. We first considered three training sets: (i) a European-ancestry training set of ~300,000 European-ancestry individuals, (ii) a African-ancestry training set of ~5000 African-ancestry individuals, and (iii) a dual-ancestry training set made up of the African-ancestry training set combined with European-ancestry individuals so that the proportion of African-ancestry individuals was 10%. For the latter, we opted for this 90-10 split—thus not using all the available European-ancestry individuals—following the previous analysis that indicated the limited benefit to African-ancestry predictive performance of adding European-ancestry individuals to the training set. A secondary motivation was to limit the imbalance between the two ancestries in order to evaluate the effect of importance weighting. We used the importance weighted LASSO with $\gamma = 0, 0.2, \ldots, 1$ to construct PGS_{dual}, while for the single-ancestry training sets we just considered the standard, unweighted LASSO. In a supplementary analysis on a subset of the traits, we also estimated dual-ancestry PGS trained on the combination of the European-ancestry training set and the African-ancestry training set, using both the unweighted and reweighted LASSO (Supplementary Fig. 3). We found that for this subset there was limited benefit over either the European-ancestry only training set, or the smaller dual-ancestry training set with only 50k European-ancestry individuals. Moreover, although the LASSO is a relatively efficient algorithm for constructing PGS using individual-level data, it nevertheless took over 210 CPU hours to fit the PGS for height on the full dual-ancestry training set, compared to 40 CPU hours on the smaller dual-ancestry training set.

Figure 4 illustrates the predictive performance of the three above PGS for the 15 UK Biobank traits. In terms of African-ancestry predictive utility, PGS_{AFR} outperformed PGS_{EUR} on two traits—MCV, and erythrocyte distribution width (erythrocyte)—despite the former sample size being orders of magnitude smaller ($n \approx 5000$ versus $n \approx 300{,}000$). For both of these traits, the unweighted PGS_{dual} performed the same as or slightly worse than PGS_{AFR}. While importance reweighting yielded a moderate improvement, the reweighted PGS_{dual} did not outperform PGS_{AFR}.

https://doi.org/10.1038/s41467-023-38930-7

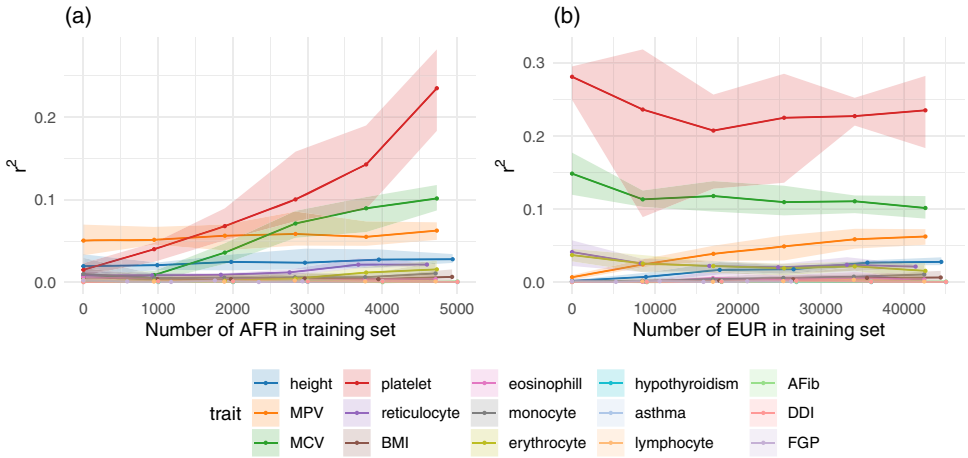

Fig. 3 | Predictive performance for African-ancestry individuals against sample size for 15 traits in UK Biobank. a We fixed the number of European-ancestry (EUR) individuals in the training set at ~50,000 (26,388 for female genital prolapse (FGP)) and varied the number of African-ancestry (AFR) individuals from 0 to ~4700 (2900). The predictive performance, evaluated in terms of partial r^2, on African-ancestry individuals increased markedly for mean corpuscular volume (MCV) and platelet crit; and stayed largely stable (or increased slightly) for the remainder. **b** Here, we instead fixed the number of African-ancestry individuals in the training set at ~4700 (2900 for FGP) for each trait and varied the number of European-ancestry individuals so that the proportion of European-ancestry individuals in the training set ranged from 0% to 90%. The effect on performance on African-ancestry individuals again varied by trait, showing a clear improvement for MPV and height, and a moderate decrease for MCV. Error bars correspond to the range across five cross-validation rounds of training set construction and PGS estimation. Phenotype acronyms: mean platelet volume (MPV), mean corpuscular volume (MCV), body mass index (BMI), atrial fibrillation (AFib), diverticular disease of the intestine (DDI), female genital prolapse (FGP).

For four other traits—high light scatter reticulocyte count (reticulocyte), hypothyroidism, atrial fibrillation (AFib), and and diverticular disease of intestine (DDI)—the predictive performance of all three PGS was largely similar. For the remaining traits—height, MPV, body mass index (BMI), eosinophill percentage, monocyte count, asthma and lymphocyte count—PGS_{EUR} performed best. We note that predictive accuracy of PGS_{EUR} for European-ancestry individuals was the same or higher than any of the PGS for African-ancestry individuals, consistent with previous investigations into the between-ancestry disparities of polygenic scores[7,9].

To explore whether the variability in optimal PGS strategy extended to other subpopulations, we focused in on four of the above traits (height, MPV, asthma, and erythrocyte distribution width) and ran the above analyses using different minority-ancestry groups: Admixed American ancestry (AMR), Central/South Asian ancestry (CSA), East Asian ancestry (EAS), and Middle Eastern ancestry (MID).

Figure 5 shows the predictive performance on the four traits across the five minority ancestry groups. For both height and asthma, PGS_{EUR} had the best predictive performance for each of ancestry groups, with partial r^2 for height varying from $r^2 = 0.08$ for the African-ancestry group to $r^2 = 0.25$ for the Admixed American ancestry group. For both erythrocyte distribution width and mean corpuscular volume in African-ancestry individuals, the PGS_{min} and the reweighted PGS_{dual} outperformed PGS_{EUR}, which offered almost no predictive utility. With the exception of mean corpuscular volume in East Asian ancestry individuals, where performance was similar between the different PGS, PGS_{EUR} was consistently the best strategy for the remaining minority-ancestry groups. Results were qualitatively identical when using only genotyped SNPs instead of the full imputed sequence (Supplementary Fig. 4).

Differences in trait architecture explain variable performance by ancestry

Finally, to investigate the reasons why optimal training approaches vary across traits, we investigated the contribution of variants at different allele frequencies to variance explained. Specifically, we measured the partial r^2 for different subsets of variants grouped by minor allele frequency (MAF) in a given ancestry group. For each ancestry group, we grouped variants into four sets: 'rare' (MAF ≤ 1%), 'uncommon' (1% < MAF ≤ 5%), 'intermediate' (5% < MAF ≤ 20%), and 'common' (MAF > 20%). Given a set of variants $\mathcal{K}$ and the genotype matrix of the test set X, let $X_{\mathcal{K}}$ denote the submatrix given by the columns of X that are in $\mathcal{K}$. For a set of variants $\mathcal{K}$, we define $\hat{\mathbf{g}}_{\mathcal{K}} = X_{\mathcal{K}}\hat{\boldsymbol{\beta}}_{\mathcal{K}}$ to be the score associated with the variant set $\mathcal{K}$, where $\hat{\boldsymbol{\beta}}$ is the vector of variant effects obtained from the LASSO. We then defined the partial r^2 attributable to variant set $\mathcal{K}$ to be the difference in r^2 between models

$$\mathbf{y} = M\boldsymbol{\theta}_1 + \hat{\mathbf{g}}_{\mathcal{K}}\eta + \boldsymbol{\epsilon}_1 \quad (1)$$

$$\mathbf{y} = M\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}_2, \quad (2)$$

where $\mathbf{y}$ is the vector of trait values, and M is the matrix of covariates (age, sex, the first ten genetic principal components (PCs), and interactions between sex and the ten genetic PCs) for the test set. We note that because we are considering the same set of variants for each trait, observed differences among traits in composition likely reflect differences in trait architecture, rather than differences in how the studies were performed.

Figure 6 illustrates the contribution to variance explained by different segments of the ancestry-specific allele frequency spectra for the PGS for mean corpuscular volume and height. We observe striking differences between these traits in the contribution of different allele frequency segments. For MCV in African-ancestry individuals, the majority of the variance explained in each of the polygenic scores was due to variants that were intermediate or common in African-ancestry individuals (see top row, left). Notably, over half the contribution to variance explained of the African-ancestry and PGS_{dual} could be attributed to variants that have MAF < 5% in the European-ancestry subgroup (second row, left). Conversely, variance explained in the European-ancestry subgroup was driven by variants that are intermediate or common in European-ancestry individuals (fourth row, left). And approximately one quarter of the variance explained could be attributed to variants with MAF < 5% in African-ancestry individuals.

https://doi.org/10.1038/s41467-023-38930-7

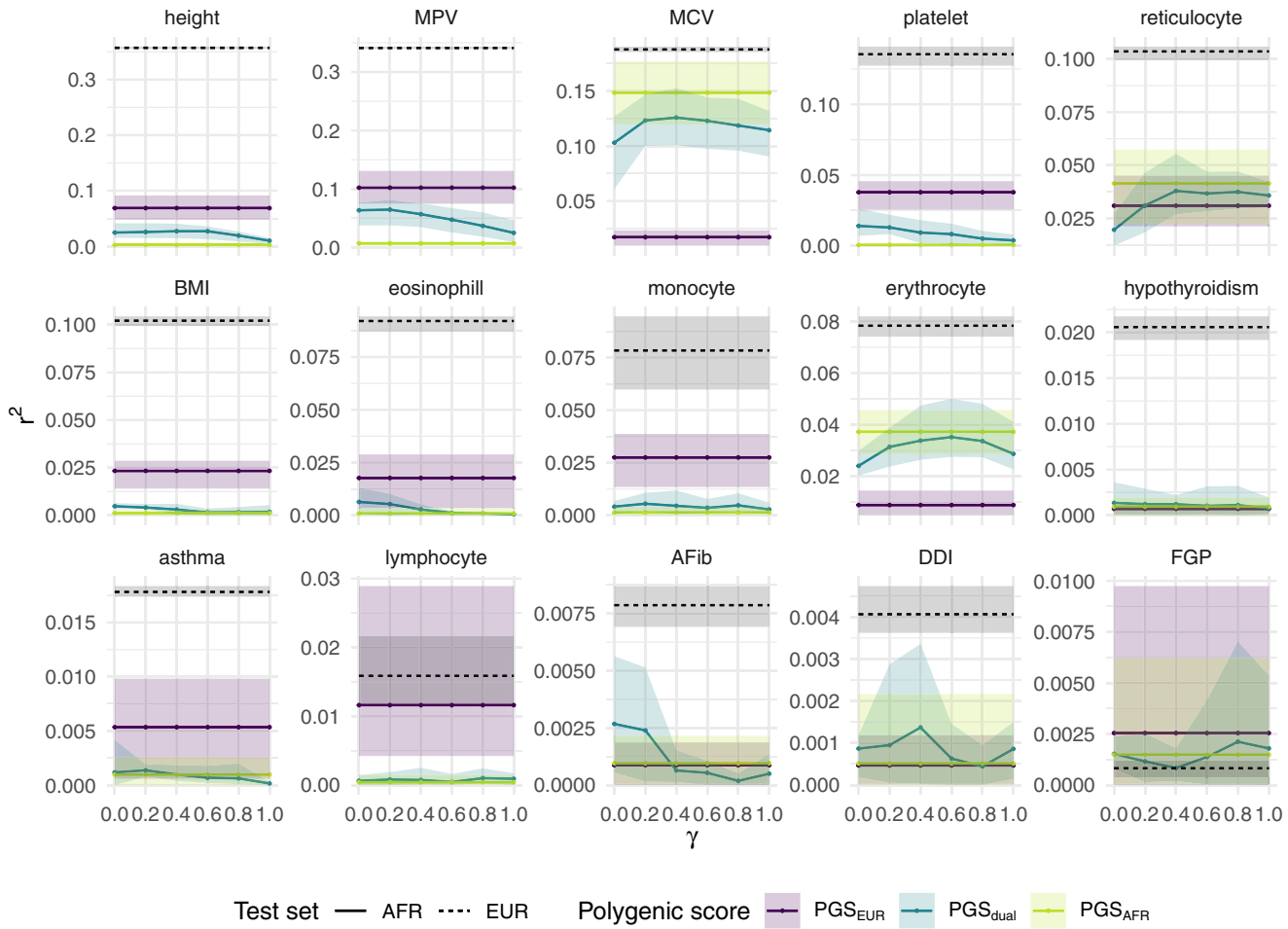

Fig. 4 | Partial r^2 for PGS_{EUR}, PGS_{dual}, and PGS_{AFR} on 15 traits in UK Biobank. Predictive performance on an African-ancestry (AFR) test set is shown by the solid lines. The dashed lines correspond to predictive performance on a European-ancestry (EUR) test set using PGS_{EUR}. The single-ancestry scores were estimated using a standard, unweighted LASSO. The dual-ancestry scores were constructed using an importance weighted LASSO with various degrees of reweighting γ. Traits are ordered according to partial r^2 of PGS_{EUR} on the European-ancestry test set (note the varying y-axes). Error bars correspond to the range across five cross-validation rounds of training set construction and PGS estimation. Phenotype acronyms: mean platelet volume (MPV), mean corpuscular volume (MCV), body mass index (BMI), atrial fibrillation (AFib), diverticular disease of the intestine (DDI), female genital prolapse (FGP).

In contrast, for height we find that variance explained in European-ancestry individuals has an allele frequency decomposition with a slightly greater contribution of variants that are common in European-ancestry individuals (bottom row, right). But, variance explained in African-ancestry individuals is dominated by variants that are common (MAF > 5%) in both populations (top two rows, right). The effect of reweighting follows a similar pattern; for MCV shifting variance explained towards variants that are higher-frequency in the African-ancestry individuals (and considerably rarer in European-ancestry individuals) and, for height, tending to favour variants that are common in both groups.

These results suggest differences in genetic architecture between the two traits. For mean corpuscular volume, the variance explained by PGS_{dual} and PGS_{AFR} could largely be attributed to variants that were relatively common in African-ancestry individuals but rare in European-ancestry individuals. Since the number of African-ancestry individuals in these training sets was relatively small, this suggests that the corresponding effect sizes are comparatively large. On the other hand, the variance explained by each of the height PGS could be attributed to variants that were common in both European-ancestry and African-ancestry individuals. We found similar patterns for the other minority ancestries (Supplementary Figs. 5–8). Specifically, for traits and ancestries where PGS_{min} outperformed the PGS_{EUR}, more variance could be attributed to variants that are more common in the minority ancestry group. We also found substantial differences across traits in the implied liability variance between PGS_{EUR} and PGS_{min} (Supplementary Table 1). Such differences in trait architecture are indicative of variation in the evolutionary and selective forces that have shaped trait variation[35–37].

Discussion

The lack of diversity in human genetic studies has been brought into focus by a number of articles (e.g.[9,38–40]), revealing that around 86% of all GWAS participants are of European descent. In the context of polygenic scores, this bias has been reflected in the lack of transferability of scores across ancestries: PGS developed using European-ancestry samples tend to perform poorly in non-European ancestry test sets[7–9]. Correspondingly, our results found that, for a range of traits, PGS estimated using European-ancestry individuals performed relatively poorly on other non-European ancestry groups. Although recent improvements in predictive performance across a number of traits indicate the potential clinical utility of PGS[2,3,5,6], the disparity in PGS accuracy across ancestry groups, raises serious technical, clinical and ethical issues, with likely substantial impacts on health inequalities if left unaddressed[9].

https://doi.org/10.1038/s41467-023-38930-7

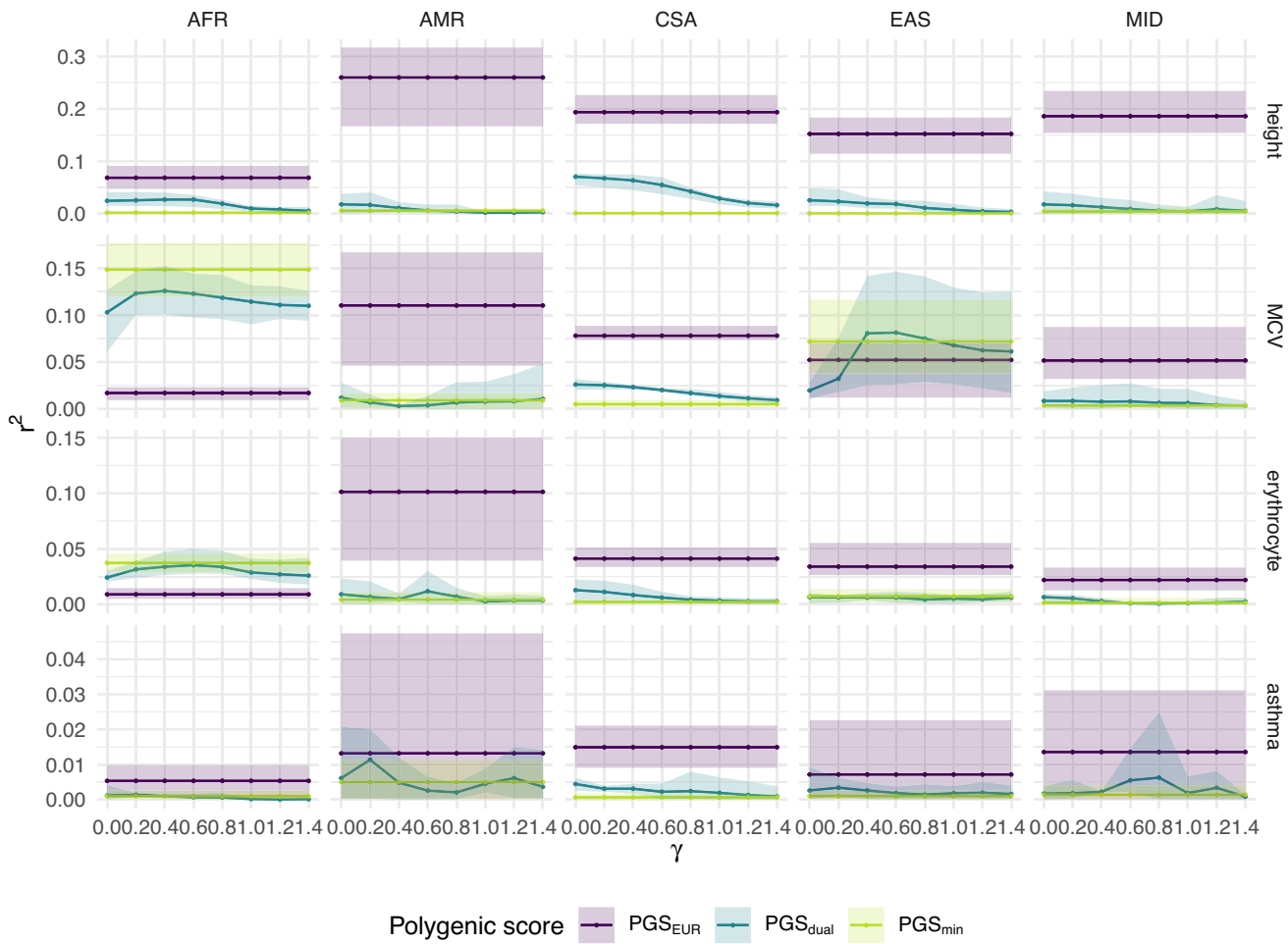

Fig. 5 | Partial r^2 for PGS_{EUR}, PGS_{dual}, and PGS_{min} on four traits in UK Biobank for five minority-ancestry groups. The single-ancestry scores were estimated using a standard, unweighted LASSO. The dual-ancestry scores were constructed using an importance weighted LASSO with various degrees of reweighting γ. Error bars correspond to the range across five cross-validation rounds of training set construction and PGS estimation. The four traits considered are height, MCV, asthma, and erythrocyte distribution width. We used inferred genetic ancestry labels from Pan-UKBB, with participants divided into six groups: European ancestry (EUR), African ancestry (AFR), Admixed American ancestry (AMR), Central/South Asian ancestry (CSA), East Asian ancestry (EAS), and Middle Eastern ancestry (MID).

Recent investigations into statistical methods have sought to construct improved PGS using the available data with the aim of reducing this disparity in predictive accuracy. Previous studies on the lack of transferability of PGS have generally estimated scores using summary statistics from genome-wide association studies of single-ancestry populations[7–9]. These summary statistic approaches are often highly efficient computationally and typically achieve highly competitive predictive performance relative to full genotype approaches[34,41–43]. A number of methods to construct PGS by combining GWAS results from distinct ancestry groups have recently been developed, demonstrating improved prediction accuracy on the minority-ancestry group[22–24].

However, the question of how to optimally construct PGS directly from multiple-ancestry training data has largely been unexplored. That is, to maximise predictive accuracy for a minority-ancestry group, is it better to employ a PGS trained on just the minority-ancestry individuals or a combination of the minority-ancestry individuals and the majority-ancestry individuals? To investigate this, we compared the prediction accuracy of PGS trained on different subsets of the available data. This was possible due to the availability of individual-level genetic data in UK Biobank, accompanied by a wide range of phenotypic measurements. This in turn allowed us to vary the number of individuals of each ancestry in each training set and thus to assess the effect of sample size and composition on PGS predictive performance for a range of quantitative and binary traits. Moreover, the use of individual-level data to estimate PGS, as opposed to the more standard approach of using external GWAS summary statistics, avoids making any assumptions such as common LD structure or phenotype definition from the GWAS population and target population (though see ref. 44 for a method that combines both approaches to improve polygenic prediction). A potential drawback of using an individual-level approach was the considerable computational cost associated with constructing the PGS, which may be impractical in the context of large-scale biobanks of increasing diversity.

It is worth noting that UK Biobank consists of an older subset of the UK population recruited on a voluntary basis in response to postal invitations. As a result, UK Biobank is not a representative sample of the wider UK population, displaying a bias towards individuals who are healthier, wealthier, and who self-report their ethnicity as white[45]. We further note that the absolute number of non-European-ancestry individuals in UK Biobank is considerably smaller than those of European-ancestry. Our findings are thus limited by the ancestral diversity of UK Biobank and more work is required to investigate optimal PGS strategies in more balanced multi-ancestry genomic datasets.

Our primary finding is that, in terms of minority-ancestry prediction performance, traits vary substantially in their optimal strategy for combining data across ancestries. Through simulation, we have

https://doi.org/10.1038/s41467-023-38930-7

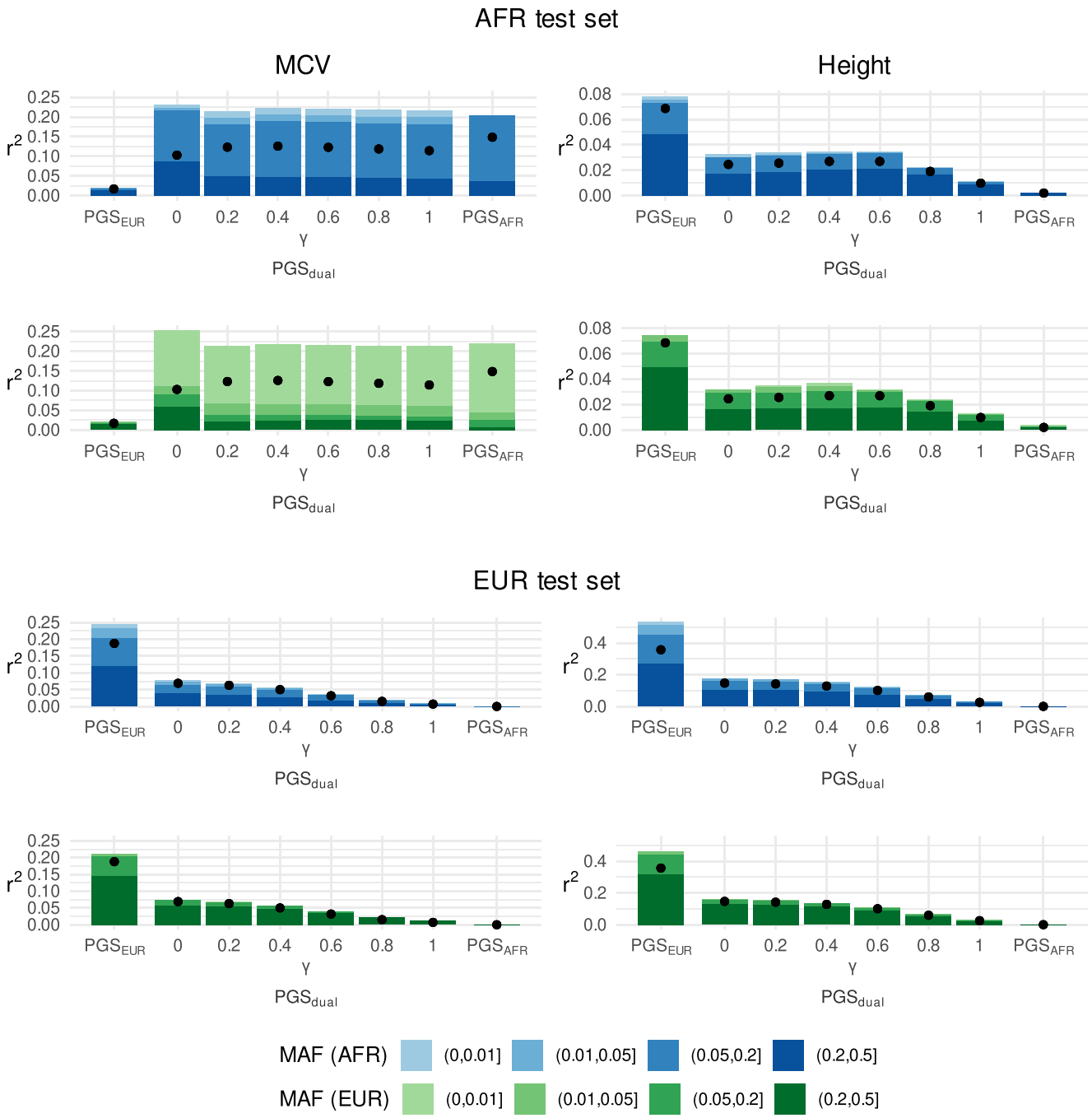

Fig. 6 | Allele frequency composition of variance explained by single- and dual-ancestry PGS. Results shown for mean corpuscular volume (left) and height (right) in a African-ancestry test set (AFR; top) and a European-ancestry test set (EUR; bottom). The black dots represent partial r^2 for all the variants, i.e. the entire polygenic score. Variants were grouped according to their minor allele frequency (MAF) in African-ancestry individuals (blue palette) or in European-ancestry individuals (green palette). Each bar represents the sum of the partial r^2 values for each subset of variants in a given polygenic score. Note that the bars are stacked, and the height of the bar is generally higher than corresponding dot due to LD between variants. The parameter γ corresponds to the degree of reweighting used in the reweighted LASSO for PGS$_{dual}$.

shown that a PGS trained on a small minority-ancestry training set may outperform a much larger European-ancestry training set. Moreover, there are plausible regions of parameter space, notably where effect sizes are correlated across ancestry groups but not identical, where reweighting strategies can boost prediction performance using a multiple-ancestry training set. However, when applied to empirical data from the UK Biobank, relatively weak benefit from reweighting was observed. Rather, traits fell into two broad categories for which the minority-ancestry PGS outperformed the majority-ancestry PGS, or vice versa. For those in the first category, the most valuable training data was the small dataset from the minority group alone (and where performance decreases by the inclusion of any individuals of the majority—here European-ancestry). For those in the second category (such as height), the best performance in the underrepresented group was achieved by training in the much larger majority group, though note this still represented a substantial decrease in absolute performance relative to the majority European-ancestry group. When the minority group was African-ancestry individuals, the first category, though smaller, included important biomarkers such as MCV. Moreover, for a given trait, the optimal strategy varied by ancestry. For MCV

for example, PGS_{EUR} outperformed the PGS_{min} for Admixed American-, Central/South Asian- and Middle Eastern-ancestry groups, while the reverse was true for the African-ancestry group. For the East Asian-ancestry group, the PGS displayed similar performance.

These findings raise the question of why the optimal strategy varies across traits and ancestry groups. Recent studies have ascribed the overall lack of PGS transferability to a number of factors, including population-level differences in allele frequencies, LD patterns, location of causal variants, and effect sizes[9–12,14,15]. We found that the variability across traits could be at least partially explained by differences in the allele frequency pattern of causal variants or, more precisely, variants in LD with a causal variant. Specifically, we found that for some traits, such as height, most predictive utility could be attributed to variants that were relatively common (MAF > 5%) in both European- and African-ancestry individuals. For other traits, such as MCV, the predictive utility could be attributed to variants that were common in African-ancestry individuals but rare in European-ancestry individuals. This points to why PGS_{EUR} for MCV performed much worse than the corresponding PGS_{AFR}, despite being trained on a much larger number of individuals. We also note that the minority-ancestry PGS and reweighted dual-ancestry PGS performed particularly well for blood cell traits. This may be due to the influence of relatively few loci that have large phenotypic effects and strong differences in allele frequency between populations indicative of strong evolutionary selective pressure (such as has been shown for the DARC (Duffy) locus[46]). For such traits, a strategy that emphasises the minority-ancestry data may be favoured. Further work is needed, however, to determine the precise nature of differences in optimal strategy and the relation with trans-ancestry differences in genetic architecture.

Our findings highlight the limitations of statistical corrections alone in reducing the disparities of PGS performance between ancestries. There are numerous sources of potential bias that adversely affect the application of an analytical pipeline to individuals of minority and/or underrepresented groups, including problem selection, data collection, outcome definition, model development and lack of real-world impact assessments and considerations[47]. Therefore, statistical solutions must be considered in combination with efforts to address the global health research funding gap[48], diversity of the bioinformatics workforce[49] and assessing the impact and translation of analyses to real-world data[50] in addition to efforts to increase the number of non-European-ancestry participants in genetic research. Each of these partial solutions, in combination, provide essential contributions to reduce and remove opportunities for negative effects on health inequalities, particular amongst those from different ancestry groups[51].

It is becoming increasingly clear that the vast disparities in PGS performance can only be bridged by improving the diversity of human genetic datasets[9]. An important consideration surrounds future sampling strategy: whose genomes should we aim to sequence to reduce existing disparities in polygenic prediction accuracy? Our findings indicate that increasing the number of samples from minority-ancestry groups can lead to significant improvements in prediction performance. Moreover, even within the same ancestry group, PGSs do not always generalise across other characteristics such as age, sex and socio-economic status[52], pointing to the need of more diversity across multiple axes both alongside and within genetic ancestry.

Perhaps counter-intuitively, with more diverse data, statistical tools such as importance reweighting may eventually play a more important role as we seek to boost predictive utility by using all the available data. Reweighting strategies have the benefit of overcoming the lack of universal definitions of race, ethnicity and ancestry, which causes considerable confusion and imprecision[53,54]. The categorisation of individuals into discrete ancestry groups to explain differences between behaviours and exposures may be unhelpful[55], whereas continuous representations of genetic ancestry[56,57] enable identification of areas of genetic ancestry where performance is stronger or weaker as well as how the trait itself varies across ancestry. Ultimately, approaches to genetic prediction must acknowledge both the many similarities of human biology, but also the differences in history, cultural heritage, exposure, and behaviour that can lead to certain factors being of greater relevance for particular groups of individuals.

Methods

PGS Estimation using the LASSO

To construct PGSs from full genotype data, we use L1-penalised regression, also known as the LASSO[30]. The LASSO has previously been used in the context of genetic prediction (see for example refs. 12,31–34 and references therein) and is suitable for high-dimensional problems where one expects the number of non-zero predictors to be small relative to the number of total predictors. Although there exist other full genotype PGS methods such as linear mixed models (see e.g. ref. 58), we focus on the LASSO largely for its computational efficiency. We provide an analysis of predictive performance versus sample size, focusing on differences in predictive performance between European-ancestry individuals and non-European-ancestry individuals.

We first briefly recap the LASSO algorithm for constructing polygenic scores. Let n be the number of individuals in the training set. Denote $M \in \mathbb{R}^{n\times q}$ to be the matrix of q covariates, X to be the $n \times p$ genotype matrix, and $\mathbf{y}$ to be the n-vector of observed phenotype values. We assume a linear relationship between the phenotype and predictors,

$$\mathbf{y} = M\boldsymbol{\theta} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\theta}$ is a q-vector of covariate effects, $\boldsymbol{\beta}$ is a p-vector of variant effects, and $\boldsymbol{\epsilon}$ is an environmental noise term with mean 0.

The LASSO aims to minimise the objective function,

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) = ||y - M\boldsymbol{\theta} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{4}$$

with respect to $(\boldsymbol{\theta}, \boldsymbol{\beta})$, where λ is a regularisation parameter. The purpose of the $||\boldsymbol{\beta}||_1$ term is to penalise large values of $|\boldsymbol{\beta}|$ and thus encourage sparse solutions, that is, solutions with a relatively low number of non-zero coefficients. Note that the parameters $\boldsymbol{\theta}$ are not penalised. The choice of λ controls the degree of penalisation with higher values of λ encouraging smaller values of $\boldsymbol{\beta}$. To select λ, we applied an 80-20 training-validation split to the training set, generating a path of solutions for a range of values of λ using the training split, and then selecting the value of λ that maximises r^2 (for quantitative traits) or the Area Under the Curve (AUC; for binary traits) on the validation split.

We optimise (4) using the R packages `glmnet`[59] and `snpnet`[34] on simulated and real data respectively. The `snpnet` package is an extension of `glmnet` designed to interface directly with the PLINK software[60,61] to handle large-scale single nucleotide polymorphism datasets such as the genotype data in UK Biobank. Using the default package settings and following the guidance in[34], we did not standardise the genotypes before fitting.

Multiple-ancestry datasets. The above description assumes that the genetic effects $\boldsymbol{\beta}$ are the same for all individuals. While this assumption may be reasonable within a single ancestry group, it is unlikely to hold more generally. We consider a hypothetical model that allows genetic effects to vary among individuals. To this end, we introduce a low-dimensional latent variable z representing the ancestry of a given individual. For an individual i,

$$y_i = m_i^T\boldsymbol{\theta}(z_i) + x_i^T\boldsymbol{\beta}(z_i) + \epsilon_i, \tag{5}$$

where $(\boldsymbol{\theta}(\cdot), \boldsymbol{\beta}(\cdot))$ are now vector-valued functions parameterised by z. The only assumption we make about these functions is one of continuity, so that individuals with similar ancestry have similar covariate and genetic effects. Otherwise, we do not make any explicit statements regarding the form of the functions.

In this paper, we consider the special case with $J = 2$ ancestry groups, such that $z_i \in \{1,2\}$. We denote $(\boldsymbol{\theta}(z_i), \boldsymbol{\beta}(z_i)) = (\boldsymbol{\theta}^{(j)}, \boldsymbol{\beta}^{(j)})$ for an individual i in ancestry group j. Using superscripts to denote the ancestry group (for example, $y^{(j)}$ is the vector of observed phenotypes for group j),

$$\mathbf{y}^{(j)} = M^{(j)}\boldsymbol{\theta}^{(j)} + X^{(j)}\boldsymbol{\beta}^{(j)} + \boldsymbol{\epsilon}^{(j)}, \quad j = 1, 2. \tag{6}$$

Denoting n_j to be the number of individuals of ancestry j in the training set, we assume $n_1 > n_2$ to reflect a European-ancestry dominated training set. Applying the standard LASSO to such a training set, we would expect the estimate $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ to be closer to $(\theta^{(1)}, \boldsymbol{\beta}^{(1)})$ than to $(\boldsymbol{\theta}^{(2)}, \boldsymbol{\beta}^{(2)})$. All else being equal, this would result in better predictive performance for test individuals in group 1.

Importance reweighting. We return for a moment to the more general model in Equation (5). Suppose we wish to construct a polygenic score for a test set with ancestry distribution $\pi^{TE}(z)$ given a training set of observations $(y_i, m_i, x_i)_{i=1}^n$ from individuals with ancestry $z_1, \ldots, z_n$, how should one obtain a single estimate for $(\boldsymbol{\theta}, \boldsymbol{\beta})$? Our proposed approach is to use importance reweighting, a statistical technique commonly used in survey sampling[26]. Importance reweighting aims to account for distributional differences between the sampling population and the population of interest by assigning weights to each of the samples in the training data. Here, we assign each of the training individuals a weight w_i that quantifies the similarity of the ancestry variable z_i to the test set ancestries, with individuals who have a similar ancestry to the test set assigned a higher weight. More concretely, assuming an ancestry distribution $\pi^{TR}(\cdot)$ on the training set, we set

$$w_i = \frac{\pi^{TE}(z_i)}{\pi^{TR}(z_i)}. \tag{7}$$

Note that we do not have access to the distributions π^{TR}, π^{TE} so in practice we must approximate (7). Given weights w_i, to estimate $(\boldsymbol{\theta}, \boldsymbol{\beta})$, we minimise the following weighted LASSO objective function:

$$L_\lambda(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left(y_i - m_i^T\boldsymbol{\theta} + x_i^T\boldsymbol{\beta}\right)^2 + \lambda||\boldsymbol{\beta}||_1. \tag{8}$$

In this work, we consider only the special case with two ancestry groups (Eqn. (6)), specifying importance weights w_1, w_2 such that $w_2 \geq w_1$ with the aim of improving estimates for group 2, i.e. the underrepresented group. Specifically, we use weights of the form:

$$w_i = \left(\frac{n_1 + n_2}{n_i}\right)^\gamma, \tag{9}$$

where γ is a hyperparameter that controls the degree of reweighting. We normalised the weights so that $n_1 w_1 + n_2 w_2 = n_1 + n_2$. With these weights, we thus minimise the objective function,

$$L_\gamma(\boldsymbol{\theta}, \boldsymbol{\beta}) = w_1||\mathbf{y}^{(1)} - M^{(1)}\boldsymbol{\theta} - X^{(1)}\boldsymbol{\beta}||_2^2 + w_2||\mathbf{y}^{(2)} - M^{(2)}\boldsymbol{\theta} - X^{(2)}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{10}$$

with respect to $(\boldsymbol{\theta}, \boldsymbol{\beta})$.

Simulation study

To evaluate the effect of importance reweighting on PGS performance under various settings, we undertook a simulation study. We used the simulation engine `msprime`[27] in the standard library of population genetic simulation models `stdpopsim` v0.1.2[28] to generate African-ancestry and European-ancestry genotypes, following a similar simulation framework to that used by Martin et al.[7]. For each ancestry group, we generated a total of 200,000 genotypes for chromosome 20, based on a three-population 'out-of-Africa' demographic model[62], using a mean mutation rate of 1.29×10^{-8} and a recombination map of GRCh37. This recombination map is from the Phase II Hapmap project and is based on 3.1 million genotyped SNPs from 270 individuals across four populations (Nigeria, Beijing, Japan, northern and western Europe)[63]. Note that this particular simulation model also generates genotypes for East Asian individuals but we do not use these in our analysis. To reduce the computational burden, we used only the first 10% of the chromosome, applying a minor allele frequency (MAF) threshold to filter out any SNPs that had a MAF of <1% in each ancestry group, resulting in a total of 189,765 variants.

We simulated phenotypes from these genotypes assuming a normal linear model,

$$\mathbf{y}^{(j)} = X^{(j)}\boldsymbol{\beta}^{(j)} + \boldsymbol{\epsilon}^{(j)}, \quad j = 1, 2, \tag{11}$$

where $\epsilon_i^{(j)} \sim \mathcal{N}(0, \sigma^2)$. The noise variance parameter σ^2 controls the SNP heritability h_j^2 for each ancestry group and was chosen to yield an average SNP heritability of 0.3 across the groups. To reflect the sparsity of genetic effects, we randomly selected $p_0 = 100$ of the $p = 189,765$ variants to be causal. Motivated by Trochet et al.[64], we drew the causal effect sizes from a bivariate normal distribution to model the similarity of genetic effects between ancestries. Denoting $\mathcal{C} = \{k_1, \ldots, k_{100}\}$ to be the indices of the causal SNPs, and $\boldsymbol{\beta}_C^{(j)} = (\beta_{k_1}^{(j)}, \ldots, \beta_{k_{100}}^{(j)})$ we have the equation

$$\begin{pmatrix} \boldsymbol{\beta}_C^{(1)} \\ \boldsymbol{\beta}_C^{(2)} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \tag{12}$$

where ρ is the correlation between the ancestry-specific genetic effects. We set $\beta_i^{(1)} = \beta_i^{(2)} = 0$ for $i \notin C$. For each value of $\rho = 0.5, 0.6, \ldots, 1.0$, we repeated the above procedure 5 times to generate 5 quantitative traits.

To investigate the effect of sample size, for each trait we created five training sets with $n_{EUR} = 18{,}000$ randomly selected individuals of European ancestry and $n_{AFR} = 2000, 4500, 7714, 12{,}000, 18{,}000$ randomly selected individuals of African ancestry. For each of these training sets, we computed weights $w_{EUR}^{(\gamma)} = (10/9)^\gamma$, $w_{AFR}^{(\gamma)} = 10^\gamma$ and where γ is a hyperparameter controlling the degree of reweighting. To investigate the impact of reweighting, we varied $\gamma = 0, 0.1, \ldots, 1.5$. Note that $\gamma = 1$ corresponds to inverse proportion reweighting when $n_{AFR} = 2000$ and $n_{EUR} = 18{,}000$. We normalised the weights to ensure that their sum equalled n, in line with the unweighted case.

To assess predictive performance for each ancestry, we constructed an African-ancestry test set and a European-ancestry test set by randomly selecting 2000 individuals from each ancestry out of those not included in the training set. We used the proportion of variance explained, denoted r^2, as the measure of predictive performance,

$$r^2 = \frac{\widehat{\mathrm{Var}}(X\hat{\boldsymbol{\beta}})}{\widehat{\mathrm{Var}}(\mathbf{y})}, \tag{13}$$

where $\widehat{\mathrm{Var}}$ denotes the sample variance. Note that this is equivalent to the partial r^2 relative to an intercept-only covariate model. For $\rho = 0.5, 0.6, \ldots, 1.0$, we repeated the above process 5 times with different randomly selected individuals in the training and test sets to calculate a mean r^2 for each trait.

In supplementary studies to investigate the role of heritability and polygenicity, i.e. the number of causal SNPs, we repeated the above

analyses with the correlation between the ancestry-specific genetic effects fixed at $\rho = 0.8$, and varying the overall heritability $h_2 = 0.1, 0.3, 0.5$ and the number of causal SNPs $p_0 = 10, 100, 1000$ respectively.

UK Biobank

The UK Biobank is a large-scale cohort study with genomic and phenotypic data collected on ~500,000 individuals aged 40-69 at the time of recruitment[25]. We used inferred genetic ancestry labels made available by the Pan-UKBB initiative[29]. These labels were derived based on reference data from the 1000 Genomes Project[65] and the Human Genome Diversity Project (HGDP)[66]. Briefly, principal component analysis (PCA) was performed on unrelated individuals in the combined reference dataset, and a random forest classifier was trained on the first 6 principal components (PCs) using continental ancestry metadata. UK Biobank individuals were then projected onto this reference PC space and given initial ancestry assignments based on the random forest classifier. These assignments were then adjusted by removing outliers. Full details can be found on the Pan-UKBB website[29].

Phenotypes. We investigated a range of quantitative and binary traits available in our UK Biobank project (UKB Application Number 12788). To select these, we applied the following procedure. We first filtered out traits with estimated SNP heritability of <5% or a 'low' confidence rating for the SNP heritability estimate, as reported on the Neale Lab UKB SNP-Heritability Browser[67,68]. To ensure that the traits under investigation were not highly correlated with each other, we used genetic correlation estimates available on the Neale Lab UKBB Genetic Correlation Browser[69,70]. Specifically, working from the most heritable traits to the least heritable, we iteratively removed a trait if it had an estimated genetic correlation of >0.5 with any remaining trait of higher heritability. Of the remaining traits, we selected the top 10 most heritable quantitative traits: height, mean corpuscular volume (MCV), body mass index (BMI), eosinophill percentage, erythrocyte distribution width (erythrocyte), lymphocyte count, monocyte count, mean platelet volume (MPV), platelet crit, and high light scatter reticulocyte count (reticulocyte). We also selected the top 5 most heritable binary traits: asthma, female genital prolapse (FGP), atrial fibrillation (AFib), diverticular disease of the intestine (DDI), and hypothyroidism. For our African-ancestry analyses, we investigated all 15 of these traits, while for the other minority-ancestry analyses, we focused in on 4 traits: height, mean corpuscular volume, erythrocyte distribution width, and asthma.

Genotypes and covariates. We used imputed genotypes from UK Biobank, filtering out individuals and variants according to quality-control metrics used by the Pan-UKBB initiative[29]. Firstly, we removed individuals who were identified as displaying sex chromosome aneuploidy. We also removed individuals who were flagged as related by Pan-UKBB. Within each ancestry group, related individuals were identified using PC-Relate[71] with $k = 10$ and a minimum individual MAF of 5%. We filtered out variants that were not deemed to be of 'high quality' according to Pan-UKBB, retaining those with an INFO score of at least 0.8 and with an allele count of at least 20 per population. We also removed variants that had a MAF of <1% in both the European-ancestry group and the minority-ancestry group under analysis. To control for population structure, we included the following covariates in our model: age, sex, the first ten genetic principal components (PCs), and interactions between sex and the ten genetic PCs.

Construction of training sets. To assess the effect of sample size and composition on PGS performance, we used subsets of the above data as training sets, controlling the number of European-ancestry and minority-ancestry individuals. The remaining individuals were then used as a held-out test set. For each trait, we constructed three types of training sets using individuals with non-missing data for that particular trait. The first two types of training sets were single-ancestry datasets, one consisting solely of European-ancestry individuals and the other consisting solely of minority-ancestry individuals.

1. European-ancestry A random sample comprising 80% of the quality-controlled European-ancestry individuals.
2. Minority-ancestry A random sample comprising 80% of the respective quality-controlled minority-ancestry individuals.
3. Dual-ancestry This set consisted of both minority-ancestry individuals and European-ancestry individuals. The basic form of this training set was made up of the minority-ancestry training set described above, combined with European-ancestry individuals so that the proportion of minority-ancestry individuals was 10%. The European-ancestry individuals were matched to the minority-ancestry individuals on age and sex. We also considered variations of this training set by removing a proportion of either the European-ancestry individuals or the minority-ancestry individuals (see Results for more details).

For each dataset, we further removed variants with a MAF of <0.1% or a missing genotype call rate of <5%. Note that, as a result, the sets of variants generally differed by training set. For the single-ancestry datasets, we used the standard, unweighted LASSO to construct the PGS. For the dual-ancestry datasets, we used the weighted LASSO with $\gamma = 0, 0.2, 0.4, 0.6, 0.8, 1$. Note that $\gamma = 0$ corresponds to the unweighted LASSO.

Predictive accuracy. We evaluated predictive accuracy of each PGS by ancestry group using individuals that were not included in the corresponding training sets. To assess the genetic predictive accuracy of a PGS, we calculated the partial r^2 attributable to the PGS, relative to a covariate-only model, following Martin et al.[9]. Specifically, we fit the following nested linear models (or the equivalent logistic models for binary traits),

$$\mathbf{y} = M\boldsymbol{\theta}_1 + \hat{\mathbf{g}}\eta + \boldsymbol{\epsilon}_1 \tag{14}$$

$$\mathbf{y} = M\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}_2, \tag{15}$$

where M is the covariate matrix and $\hat{g}$ is the vector consisting of polygenic scores for each individual in the test set. We took the partial r^2 (or the Cox and Snell pseudo-r^2 for binary traits) between models (14) and (15) as our measure of predictive accuracy. To obtain more reliable estimates, we performed 5-fold cross-validation. That is, we repeated five times the process of training set construction and PGS estimation to obtain five estimates of partial r^2, and report the mean and the range of these estimates.

Reporting summary

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

Data availability

This research has been conducted using data from UK Biobank, a major biomedical database, under application 12788[25]. The UK Biobank data are available under restricted access; access can be obtained by researchers upon application—see https://www.ukbiobank.ac.uk/enable-your-research. The genetic correlation estimates from the Neale Lab UKBB Genetic Correlation Browser[69,70] are available at

https://ukbb-rg.hail.is/. The heritability estimates from the Neale Lab UKB SNP-Heritability Browser[67,68] are available at https://nealelab.github.io/UKBB_ldsc/downloads.html. The ancestry-specific heritability estimates and inferred genetic ancestry labels for UK Biobank individuals from the Pan-UKBB initiative[29] are available at are available at https://pan.ukbb.broadinstitute.org/downloads. The recombination map of GRCh37 is available as part of the stdpopsim python library—see https://popsim-consortium.github.io/stdpopsim-docs.

Code availability

The code used to generate these results is available at https://github.com/brieuclehmann/pgs_transferability/releases/tag/v0.1. To construct the PGS, we used the glmnet R package v4.1 (for simulated data), and a slight adaptation of the snpnet R package[34] v0.3.0 available at https://github.com/brieuclehmann/snpnet (for UK Biobank data). To generate simulated data, we used the msprime simulation engine via the stdpopsim v0.1.2 python library—see https://popsim-consortium.github.io/stdpopsim-docs/. To preprocess the UK Biobank genetic data, we made use of both PLINK v1.9 and PLINK v2.0, available at https://www.cog-genomics.org/plink/.

References

1. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
2. Torkamani, Ali, Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
3. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
4. Knowles, J. W. & Ashley, E. A. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* **15**, 1–7 (2018).
5. Maas, P. et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–1302 (2016).
6. Sharp, S. A. et al. Development and standardization of an improved Type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care* **42**, 200–207 (2019).
7. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
8. Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
9. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
10. Scutari, M., Mackay, Ian & Balding, D. Using genetic distance to infer the accuracy of genomic prediction. *PLoS Genet.* **12**, 1–19 (2016).
11. Wang, Y. et al. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
12. 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).
13. Shi, H. et al. Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am. J. Hum. Genet.* **106**, 805–817 (2020).
14. Carlson, C. S. et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: The PAGE Study. *PLoS Biol.* **11**, 1–11 (2013).
15. Brown, B. C., Ye, C., Price, A. L. & Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
16. Galinsky, K. J. et al. Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).
17. Franks, P. W., Pearson, E. & Florez, J. C. Gene-environment and gene-treatment interactions in type 2 diabetes: progress, pitfalls, and prospects. *Diabetes Care* **36**, 1413–1421 (2013).
18. Bentley, A. R. et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat. Genet.* **51**, 636–648 (2019).
19. H3 Africa Consortium. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
20. Grinde, K. E. et al. Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genet. Epidemiol.* **43**, 50–62 (2019).
21. Márquez-Luna, C. & Loh, Po-Ru Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
22. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
23. Ruan, Y. et al. Improving polygenic prediction in ancestrally diverse populations. *Nat. Genet.* **54**, 573–580 (2022).
24. 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).
25. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
26. Scheaffer, Richard L., Mendenhall III, William, Ott, R Lyman. and Gerow, Kenneth G. *Elementary survey sampling*. Cengage Learning, (2011).
27. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol.* **12**, 1–22 (2016).
28. Adrion, J. R. et al. A community-maintained standard library of population genetic models. *eLife* **9**, e54967 (2020).
29. The Pan-UKBB team. *Pan UKBB*. https://pan.ukbb.broadinstitute.org (2020).
30. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B (Methodol.)* **58**, 267–288 (1996).
31. Waldmann, P. et al. AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. *BMC Bioinforma.* **20**, 167 (2019).
32. Okser, S. et al. Regularized machine learning in the genetic prediction of complex traits. *PLOS Genet.* **10**, 1–9 (2014).
33. Privé, F., Aschard, H. & Blum, M. G. B. Efficient implementation of penalized regression for genetic risk prediction. *Genetics* **212**, 65–74 (2019).
34. Qian, J. et al. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* **16**, 1–31 (2020).
35. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, 1–20 (2018).
36. Zeng, Jian et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* **50**, 746–753 (2018).
37. Schoech, A. P. et al. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).
38. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
39. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
40. Fatumo, S. et al. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).

41. Lloyd-Jones, L. R. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
42. Mak, T., Porsch, R., Choi, S., Zhou, X. & Sham, P. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
43. Vilhjálmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
44. Albi nana, C. et al. Leveraging both individual-level genetic data and GWAS summary statistics increases polygenic prediction. *Am. J. Hum. Genet.* **108**, 1001–1011 (2021).
45. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
46. McManus, K. F. et al. Population genetic analysis of the darc locus (duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLoS Genet.* **13**, 1–27 (2017).
47. Chen, I. Y. et al. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
48. Vidyasagar, D. Global notes: the 10/90 gap disparities in global health research. *J. Perinatol.* **26**, 55–56 (2006).
49. Hofstra, Bas et al. The diversity–innovation paradox in science. *Proc. Natl Acad. Sci.* **117**, 9284–9291 (2020).
50. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
51. Peterson, R. E. et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
52. Mostafavi, H. et al. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife* **9**(Jan.), 48376 (2020).
53. Mathieson, I. & Scally, A. What is ancestry? *PLoS Genet.* **16**, 1–6 (2020).
54. Mersha, T. B. & Abebe, T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Hum. Genomics* **9**, 1 (2015).
55. Foster, M. W. & Sharp, R. R. Race, ethnicity, and genomics: Social classifications as proxies of biological heterogeneity. *Genome Res.* **12**, 844–850 (2002).
56. Belbin, G. M. et al. Toward a fine-scale population health monitoring system. *Cell* **184**, 2068–2083.e11 (2021).
57. Lewis, AnnaC. F. et al. Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).
58. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genet.* **9**, 1–14 (2013).
59. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
60. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
61. Chang, C & Shaun P. *PLINK 1.90 beta* www.cog-genomics.org/plink/1.9/ (2023).
62. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, 1–11 (2009).
63. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
64. Trochet, Holly et al. Bayesian meta-analysis across genome-wide association studies of diverse phenotypes. *Genet. Epidemiol.* **43**, 532–547 (2019).
65. Auton, Adam et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
66. Cann, H. M. et al. A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
67. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
68. The Neale Lab. *UK Biobank Heritability Browser*. https://nealelab.github.io/UKBB_ldsc/ (2019).
69. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
70. The Neale Lab. *UK Biobank Genetic Correlation Browser*. https://ukbb-rg.hail.is/ (2019).
71. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
72. Zhou, Wei et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

Acknowledgements

This research was funded in whole, or in part, by the UKRI [EP/R018561/1; UK Engineering and Physical Sciences Research Council (EPSRC) Bayes4Health programme] (B.L., C.H., G.M.) and the Wellcome Trust [100956/Z/13/Z] (G.M.). The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. B.L. also gratefully acknowledges funding from Jesus College, Oxford. C.H. was also supported by The Alan Turing Institute, Health Data Research UK, the Medical Research Council UK, and AI for Science and Government UK Research and Innovation (UKRI). G.M. was also supported by the Li Ka Shing Foundation. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author contributions

B.L., G.M., and C.H. designed the study. B.L. carried out the analyses. M.M., G.M., and C.H. provided critical feedback and framing. All authors contributed to writing and editing of the manuscript.

Competing interests

G.M. is a director of and shareholder in Genomics PLC, and is a partner in Peptide Groove LLP. M.M. is a Programme Lead for the Diverse Data initiative at Genomics England Ltd. B.L. and C.H. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-38930-7.

Correspondence and requests for materials should be addressed to Brieuc Lehmann.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.