LEVERAGING LOCAL GENETIC INFORMATION IN HIGH-DIMENSIONAL BAYESIAN REGRESSION: METHODS AND COMPUTATION TOOLS

By

Alexa S. Lupi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics – Doctor of Philosophy

2024

ABSTRACT

I present three projects that propose computationally efficient Bayesian methods and applications for analyzing high-dimensional genetic data, focusing on incorporating local SNP information, such as linkage disequilibrium, to elucidate genetic variability across the genome. Genome-wide information offers valuable insights, but its vast scale presents significant statistical, computational, and interpretation challenges. Focusing on local genomic segments can help address these challenges by providing a more refined approach to understanding genetic variation, particularly across different ancestry groups.

In Chapter 1, I propose an approach to map the contribution of short chromosome segments to the genetic correlation between traits. While genome-wide genetic correlations between traits offer an overall estimate for comorbid traits, local regions with opposing directional genetic correlations are masked, making it challenging to untangle the strength of the relationship overall. Chapter 1 addresses this limitation by estimating local genetic correlations. Hyperuricemia/gout and chronic kidney disease are comorbid conditions for which the biological roots of the comorbidity remain unknown. Utilizing a novel approach, I disentangled the shared genetic regions contributing to both conditions. The results presented in this chapter validate several previously suggested pleiotropic loci and discovered new ones, with about a third showing genetic correlation estimates opposite to the overall correlation.

Chapter 2 focuses on estimating the portability of local polygenic scores in cross-ancestry prediction accuracy. The vast majority of genetic data comes from individuals of European ancestry. As a result, many investigators attempt cross-ancestry prediction, utilizing European data to predict the risk of disease/traits among underrepresented non-European ancestries. In most cases, cross-ancestry prediction remains more accurate than within-ancestry predictions due to

limitations imposed by non-European sample sizes, but it is still low. This shortcoming is largely due to differences in allele frequencies and linkage disequilibrium patterns between different ancestry groups, as well as genetic-by-environmental interactions involving environmental exposures that are not independent of ancestry. In this study, I propose a method, MC-ANOVA, to estimate the relative accuracy loss in cross-ancestry prediction across ancestries due to local linkage disequilibrium and allele frequency differences. I implemented the proposed algorithm and developed maps of the relative accuracy of cross-ancestry prediction for four non-European ancestry groups. Furthermore, I developed an interactive R Shiny app that can be used to visualize the results obtained in each portability map. My findings revealed significant variability in the portability of local PGS across genomic regions, reflecting varying degrees of genetic similarity between ancestries across regions. This study highlights the potential for improving cross-ancestry predictions by taking local genetic differences into account.

The advent of big data has had a remarkable impact on PGS prediction accuracy. Sample size affects both the power to detect significant associations between SNPs and phenotypes and the accuracy of SNP effects estimates. For homogenous populations, PGS prediction accuracy grows monotonically with sample size. However, when using multi-ancestry data, the relative proportion of each ancestry group can greatly impact prediction accuracy. Therefore, in Chapter 3, using data from individuals of European ancestry from the UK Biobank and African ancestry from All of Us, I investigate how sample size and the relative proportion of each ancestry group influence PGS prediction accuracy. This study sheds light on the relative benefits of increasing within- and across-ancestry sample sizes in cross-ancestry genetic predictions through empirical results, ultimately highlighting the importance of prioritizing the collection of non-European ancestry data.

TABLE OF CONTENTS

INTRODUCTION

In recent years, statistical genetics has obtained unprecedented access to vast datasets such as the UK Biobank, with near half a million participants with genotypes (at millions of single-nucleotide polymorphisms [SNPs]) and thousands of phenotypes and disease records. The ever-increasing sample sizes in genetic data have significantly improved the statistical power of genome-wide association studies (GWAS), leading to the publication of thousands of results[1]. However, this increase in data is accompanied by an increase in statistical and computational challenges.

While advancements in statistical methods have allowed for the evaluation of complex genome-wide models, as sample sizes grow it becomes increasingly less efficient and feasible to analyze hundreds of thousands of SNPs using standard techniques. Common approaches, such as single-SNP methods, fail to incorporate linkage disequilibrium (LD) that exists between flanking variants. Additionally, due to variation across the genome, models attempting to use whole-genome information can mask important differences between chromosome segments[2]. In this dissertation, I propose methods to estimate important genetic parameters for short chromosome segments.

In Chapter 1, I propose an approach to map the contribution of short chromosome segments to the correlation between traits. Using this methodology, and data from the UK Biobank, I report estimates of the (local) genetic correlation between serum urate and estimated glomerular filtration rate. The results presented in Chapter 1 validate several previously suggested pleiotropic loci and discovered new ones, with about a third showing genetic correlation estimates opposite to the overall correlation.

The prediction accuracy for European (EUR)-ancestry individuals has improved with increased statistical power from larger sample sizes[3,4]. However, the vast overrepresentation of

EUR-ancestry in GWAS datasets (approximately 80%[5]) leads to poor cross-ancestry prediction accuracy. This is particularly true for more distant ancestry groups, such as African (AF)[5–16]. The poor portability of EUR-derived PGS in cross-ancestry prediction has been primarily attributed to differences in allele frequencies and LD, among other factors[8,9,17].

I hypothesize that, owing to varying levels of LD and allele frequency differences between ancestry groups, the portability of local PGS varies substantially over the genome, with some regions having high portability of SNP effects between ancestries and others exhibiting very poor potability in cross-ancestry prediction. Therefore, in Chapter 2, I propose a methodology to map the portability of local PGS between ancestry groups. The methodology uses a Monte Carlo approach to map both within-ancestry loss of accuracy (due to imperfect LD between markers and causal loci) and the loss of accuracy in cross-ancestry prediction attributable to differences in allele frequencies and LD between ancestry groups. I used the proposed methodology, and data from the UK Biobank to generate maps of the relative accuracy of local PGS in cross-ancestry prediction for several non-EUR ancestry groups.

Finally, in Chapter 3, building on the investigations in cross-ancestry PGS prediction accuracy, I investigate the impact of sample size and of the proportion of data from different ancestry groups on PGS prediction accuracy. In this study, I used data from individuals of EUR ancestry from the UK Biobank (n~250,000)[18] and AF ancestry data from All of Us (n~50,000)[19]. The results emphasize the importance of investing in the collection of non-EUR data.

REFERENCES

1. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023 Jan 6;51(D1):D977–85.

2. Shi H, Mancuso N, Spendlove S, Pasaniuc B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. The American Journal of Human Genetics. 2017 Nov;101(5):737–51.

3. Lello L, Avery SG, Tellier L, Vazquez AI, de los Campos G, Hsu SDH. Accurate Genomic Prediction of Human Height. Genetics. 2018;210(2):477–97.

4. Kim H, Grueneberg A, Vazquez AI, Hsu S, de los Campos G. Will Big Data Close the Missing Heritability Gap? Genetics. 2017;207(3):1135–45.

5. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.

6. Dikilitas O, Schaid DJ, Kosel ML, Carroll RJ, Chute CG, Denny JA, et al. Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. Am J Hum Genet. 2020 May 7;106(5):707–16.

7. Scutari M, Mackay I, Balding D. Using Genetic Distance to Infer the Accuracy of Genomic Prediction. Hickey JM, editor. PLoS Genet. 2016 Sep 2;12(9):e1006288.

8. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020 Jul 31;11(1):3865.

9. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, 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. 2022 Jan 6;109(1):12–23.

10. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. The American Journal of Human Genetics. 2015 Oct;97(4):576–92.

11. Belsky DW, Moffitt TE, Sugden K, Williams B, Houts R, McCarthy J, et al. Development and evaluation of a genetic risk score for obesity. Biodemography Soc Biol. 2013;59(1):85–100.

12. Domingue BW, Belsky D, Conley D, Harris KM, Boardman JD. Polygenic Influence on Educational Attainment: New evidence from The National Longitudinal Study of Adolescent to Adult Health. AERA Open. 2015;1(3):1–13.

13. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018 Jul 23;50(8):1112–21.

14. Vassos E, Di Forti M, Coleman J, Iyegbe C, Prata D, Euesden J, et al. An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis. Biol Psychiatry. 2017 Mar 15;81(6):470–7.

15. Li Z, Chen J, Yu H, He L, Xu Y, Zhang D, et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nat Genet. 2017 Nov;49(11):1576–83.

16. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. The American Journal of Human Genetics. 2017 Apr;100(4):635–49.

17. Lupi AS, Vazquez AI, de los Campos G. Mapping the relative accuracy of cross-ancestry prediction. Nat Commun. 2024; *accepted 2024*.

18. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

19. The All of Us Research Program Investigators. The “All of Us” Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.

CHAPTER 1: Local genetic covariance between serum urate and kidney function estimated with Bayesian multitrait models

This chapter is from a published manuscript:

Alexa S. Lupi, Nicholas A. Sumpter, Megan P. Leask, Justin O'Sullivan, Tayaza Fadason, Gustavo de los Campos, Tony R. Merriman, Richard J. Reynolds, Ana I. Vazquez. Local genetic covariance between serum urate and kidney function estimated with Bayesian multitrait models. *G3 Genes|Genomes|Genetics*, Volume 12, Issue 9, September 2022, jkac158, https://doi.org/10.1093/g3journal/jkac158

Abstract

Hyperuricemia (serum urate >6.8 mg/dl) is associated with several cardiometabolic and renal diseases, such as gout and chronic kidney disease. Previous studies have examined the shared genetic basis of chronic kidney disease and hyperuricemia in humans either using single-variant tests or estimating whole-genome genetic correlations between the traits. Individual variants typically explain a small fraction of the genetic correlation between traits, thus the ability to map pleiotropic loci is lacking power for available sample sizes. Alternatively, whole-genome estimates of genetic correlation indicate a moderate correlation between these traits. While useful to explain the comorbidity of these traits, whole-genome genetic correlation estimates do not shed light on what regions may be implicated in the shared genetic basis of traits. Therefore, to fill the gap between these two approaches, we used local Bayesian multitrait models to estimate the genetic covariance between a marker for chronic kidney disease (estimated glomerular filtration rate) and serum urate in specific genomic regions. We identified 134 overlapping linkage disequilibrium windows with statistically significant covariance estimates, 49 of which had positive directionalities, and 85 negative directionalities, the latter being consistent with that of the overall genetic covariance. The 134 significant windows condensed to 64 genetically distinct shared loci which validate 17 previously identified shared loci with consistent directionality and revealed 22 novel pleiotropic genes. Finally, to examine potential biological mechanisms for these shared loci, we have identified a subset of the genomic windows that are associated with gene expression using colocalization analyses. The regions identified by our local Bayesian multitrait model approach may help explain the association between chronic kidney disease and hyperuricemia.

Introduction

Chronic kidney disease (CKD) carries significant global health and economic burden[1,2]. CKD stages three to five manifest as decreased renal function and are defined by elevated serum creatinine (sCr) or estimated glomerular filtration rate (eGFR) <60 mL/min/1.73m^2. Hyperuricemia is defined by serum urate (sU) concentration >6.8 mg/dL and is contributed to by deteriorating renal function[3]. Hyperuricemia has several comorbidities associated with it, including CKD and gout[3–5]. Among people with hyperuricemia, there is a higher prevalence of CKD, and among patients with CKD, sU concentrations are higher[6,7].

Genome-wide analyses have demonstrated that the association observed between eGFR and serum urate has a genetic basis. Tin et al. carried out a large-sample trans-ethnic genome-wide association study (GWAS) of sU and, through cross-trait linkage disequilibrium (LD) score regression, obtained an estimate of overall genetic correlation between eGFR and sU of -0.26 (standard error of 0.04)[8]. This was one of the largest negative correlations with sU out of 748 traits analyzed[8]. Reynolds et al., using two large family-based datasets and Bayesian whole-genome regressions, obtained global genetic correlations between sCr (which has a direct inverse relationship to eGFR, hence the directionality difference between the estimates) and sU of 0.20 (95% credibility region (CR): 0.07, 0.33) in one dataset and 0.25 (95% CR: 0.07, 0.41) in the other[9]. While these estimates contribute to dissecting biological causes of the observed comorbidities, the shared pleiotropic genomic regions and underlying biological mechanisms are only reliably discovered by estimating local genetic covariances[10].

GWAS of sU and eGFR have identified numerous loci associated with each phenotype separately. A recent study comparing large GWAS of these traits identified 36 shared loci[11]. However, the GWAS methods used to detect the shared signals are based on the marginal

association of individual single-nucleotide polymorphisms (SNPs) with phenotypes, thus not accounting for LD between SNPs. Our method improves over post-analysis of GWAS summary statistics by estimating neighboring SNP effects concomitantly. Incorporating local LD to estimate genetic effects in a tightly segregating chromosomal segment has been previously suggested to account for the correlation between SNPs[12–14]. Additionally, our methodology implements a multi-trait model so we obtain direct genetic covariance estimates.

In this study, we aimed to characterize the common genetic basis for CKD (eGFR) and hyperuricemia (sU levels) by identifying pleiotropic genomic regions. To achieve this goal, we identified the local regions contributing to genetic variances and covariances across the whole genome[14]. We used Bayesian multi-trait models to estimate the genetic (co)variances. SNP effects were estimated in large DNA regions and genetic variances and covariances were calculated from the posterior means per LD window. We identified 64 unique local genetic regions with significant local genetic covariance, including previously implicated and novel shared loci.

Materials and Methods

Participants

This study was based on 333,542 Caucasian participants from the UK Biobank. Participants missing serum urate or serum creatinine for both of their two visits were excluded from the analysis. We excluded close relatives with relatedness ≥ 0.1, estimated using the R package BGData[15] (see details in the Supplementary Methods).

Genotypes and phenotypes

The UK Biobank used the custom UK Biobank Axiom™ Array by Affymetrix to genotype study participants[16]. Quality control involved removing SNPs that had a minor allele frequency less than 1% or a missing call rate greater than 5%, resulting in 607,490 autosomal

chromosomes (1-22) SNPs[17].

Serum urate and sCr data were obtained from the first visit. For the small number of participants (0.28%) that did not have phenotype data of interest collected at the first visit, we retrieved data from the second visit. sCr was used to define eGFR and details on this can be found in the Supplementary Methods. For both eGFR and sU, we took a *log* transformation to normalize their distributions and preadjusted by age, sex, and the first five SNP-derived principal components using ordinary least squares.

Local Bayesian multi-trait models

We estimated local (co)variances by fitting Bayesian models to chromosomal segments with a non-overlapping core of 1,000 contiguous SNPs (between 3-4 Mbp depending on the region). We included two overlapping flanking regions each consisting of 250 SNPs to each side of the core. The SNPs in the flanking regions were included to account for the effects of SNPs that were outside of the core region but possibly in LD with SNPs in the core segment. Whole genome regressions have been used to fit several markers concomitantly (e.g., Vazquez et al.[18]). However, biobank data imposes computational restrictions due to its large dimensions. In the context of a single trait, local Bayesian conditional regressions have been employed to deal with the computational burden (Funkhouser et al.[14]). In their study, the authors indagated sex differences in genetic effects in single-trait models. Here, we utilized the idea of conditional regressions in large chunks of DNA with flanking regions in the context of a multi-trait Bayesian model. This provides posterior estimates of variances and covariances between traits to find pleiotropic regions. The linear model used had the form $\mathbf{Y}= \mathbf{1}\boldsymbol{\mu}' + \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$, where $\mathbf{Y}_{n \text{ x } 2}$ is a matrix containing the pre-adjusted phenotypes, $\boldsymbol{\mu}_{2 \text{ x } 1}$ is a vector of trait-specific intercepts, $\mathbf{X}_{n \text{ x } 1500}$ is a SNP-genotype matrix (1,000 core SNPs plus 250 flanking SNPs to each side), $\boldsymbol{\beta}_{1500 \text{ x } 2}$ is a matrix

of SNP effects, and $\mathbf{E}_{n \times 2}$ is a matrix of error terms. The error terms were assumed to be IID multivariate normal with a mean of zero and covariance $\text{Var}(\boldsymbol{\varepsilon}_i)=\mathbf{R}_{2 \times 2}$, where $\boldsymbol{\varepsilon}_i$ is the i^{th} row of **E**. We used IID priors with a point of mass at zero and a bivariate Gaussian slab with a mean of zero and (co)variance matrix $\boldsymbol{\Sigma}_{2 \times 2}$. The extent of shrinkage and variable selection was influenced by three groups of parameters: **R**, **Σ**, and the prior proportion of non-zero effects, **π**. For a two-trait model, $\boldsymbol{\pi}=\{\pi_1, \pi_2\}$ and represents the prior probability of non-zero effects for traits 1 and 2 (sU and eGFR), respectively. We treated the {**R**, **Σ**, **π**} parameters as unknown and we assigned Inverse-Wishart priors for the (co)variance matrices and Beta priors for the prior probability of non-zero effects.

We used the Multitrait function from the BGLR R package available in the R CRAN[19] to generate 5,000 samples from the posterior distribution for each chromosomal segment. We filtered the samples of the SNP effects collected using a burn-in of 250 SNPs and a thinning interval of 10, thus retaining 475 samples for further inference.

Defining local LD-based windows

After we obtained the model estimates, for each core segment SNP we defined an LD window that contained correlated, neighboring SNPs with an overlapping sliding technique[13,14]. Within each LD window, we collected the corresponding estimated effects and computed (co)variance estimates (described below). For each seed SNP x_{ij} (i=1,...,n individuals and j=1,...,p core segment SNPs) coming from the core segment of SNPs, we sequentially identified SNPs in both directions (x_{ij*}) surrounding the seed SNP and included them in window j if $\text{Corr}(x_{ij}, x_{ij*}) \geq 0.1$. In a simplified example, if SNP x_{ij} had an adequate pairwise correlation with 2 SNPs to the left, and 1 SNP to the right, the window for that SNP would be defined as the set of SNPs: $\{x_{ij-2}, x_{ij-1}, x_{ij}, x_{ij+1}\}$. That is, $\text{Corr}(x_{ij}, x_{ij-1}) \geq 0.1$ and $\text{Corr}(x_{ij}, x_{ij-2}) \geq 0.1$ and $\text{Corr}(x_{ij}, x_{ij+1}) \geq 0.1$. Our

definition of an LD sliding window also involved an allowance for one SNP in the sequential process to not meet this correlation criterion, to allow for a brief loss of LD or minor mapping errors, and the SNP was still included in the LD window. In the previous example, if Corr(x_{ij}, $x_{ij\text{-}1}$) < 0.1, and Corr(x_{ij}, $x_{ij\text{-}2}$) ≥ 0.1, then the set would still include both $x_{ij\text{-}2}$ and $x_{ij\text{-}1}$. The LD window ends when two SNPs sequentially did not meet the criteria described above. The LD windows could include flanking buffer SNPs, but buffer SNPs were never used to define an LD window.

Local (co)variances

For each LD window, we computed the local variances for traits 1 and 2 and the local and covariances using $V_{w1s} = \mathrm{Var}(\mathbf{X}_{\mathrm{w}}\boldsymbol{\beta}_{\mathrm{w1s}})$, $V_{w2s} = \mathrm{Var}(\mathbf{X}_{\mathrm{w}}\boldsymbol{\beta}_{\mathrm{w2s}})$, and $\mathrm{Cov}_{\mathrm{ws}} = \mathrm{Cov}(\mathbf{X}_{\mathrm{w}}\boldsymbol{\beta}_{\mathrm{w1s}}, \mathbf{X}_{\mathrm{w}}\boldsymbol{\beta}_{\mathrm{w2s}})$. Here, $\mathbf{X}_{\mathrm{w}}$ is the matrix containing the genotypes of the SNPs in the w^{th} window and $\boldsymbol{\beta}_{\mathrm{w1s}}$ and $\boldsymbol{\beta}_{\mathrm{w2s}}$ are the samples of effects of those SNPs for traits 1 and 2 collected at the s^{th} iteration of the sampler. This generated samples from the posterior distribution of the local (co)variances, which we used to produce posterior mean estimates (by averaging across the samples from the posterior distribution), estimate posterior standard deviations, and obtain 95% posterior CRs. As discussed in Lehermeier et al.[20], this approach accounts for the contribution of local LD to genetic (co)variances and, by averaging over samples from the posterior distribution, for uncertainty about SNP effects.

Gene expression/eQTL analysis

A colocalization analysis was performed between GWAS significant markers for sU and sCr and the publicly available eQTL data from GTEx V8[21]. The R package COLOC was used, which implements a Bayesian test that analyses a single genomic region and identifies LD patterns in that locus using SNP summary statistics and the associated minor allele frequencies. The lead variant for both sCr and sU was used at each significant covariance window with a

surrounding 500 kb buffer in the GTEx database. The Contextualizing Developmental SNPs using 3D Information algorithm[22,23] was modified to identify long-distance regulatory relationships for the lead sU and sCr variants at each significant covariance window within a 500 kb region. eQTL data for variants +/- 500 kb of the lead variant were also extracted from GTEx and then COLOC was used to assess if the significant *cis*- and *trans*-eQTL identified were colocalized with sCr and sU signals. An eQTL was determined to be colocalized if the COLOC H4 (posterior probability of colocalization (PPC)) was at least 0.5 for both traits and at least 0.8 for one of the two traits, according to Giambartolomei et al.[21].

Validation

We performed a validation analysis with the related Caucasian UK Biobank cohort, consisting of 57,370 subjects not missing sU or eGFR phenotypes. The genotyping array used for this cohort is the same as that used for the discovery analysis cohort. The validation analysis repeated the estimation procedures described above and the sliding LD windows used were identical to those used in the discovery set.

Results

This study was based on 333,542 distantly related white participants, of whom 53.7% were female with an average age of 56.9 ± 8.0 years old. The average sCr level was 0.8 ± 0.2 mg/dL (the average ± standard error), average eGFR was 144.2 ± 56.0 ml/min/1.73 m^2, and the average sU level was 5.2 ± 1.3 mg/dL. Two (2.0) percent of the individuals had an ICD10 diagnosis or self-diagnosis of gout, 12.4% had hyperuricemia, 0.5% had CKD, and 0.3% had hyperuricemia and CKD.

We analyzed the markers (sU and eGFR) using a sequence of Bayesian multi-trait models where the markers were regressed on contiguous SNPs in a large chromosomal segment (core)

plus overlapping flanking buffers. We collected the samples from the posterior distribution of effects for each core segment and used these samples to estimate the local variances for each marker (Figure 1) and the local covariances between the markers (Figure 2). The (co)variances were estimated within 511,828 overlapping LD windows (small, non-independent contiguous chromosomal regions).

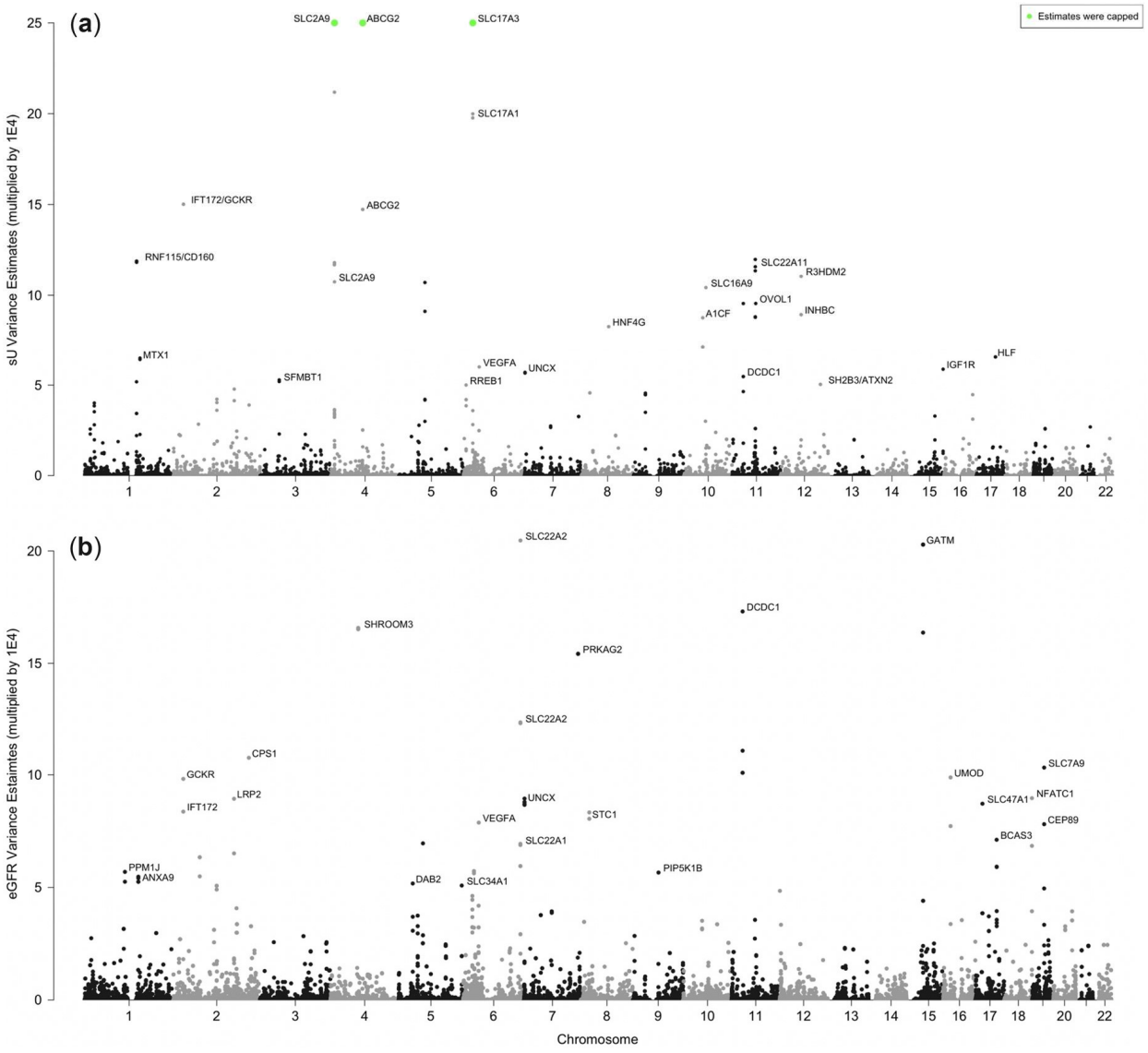


Figure 1: The variance estimates of overlapping LD windows. a) Variance estimates multiplied by 1E4 for sU concentrations and (b) for eGFR.

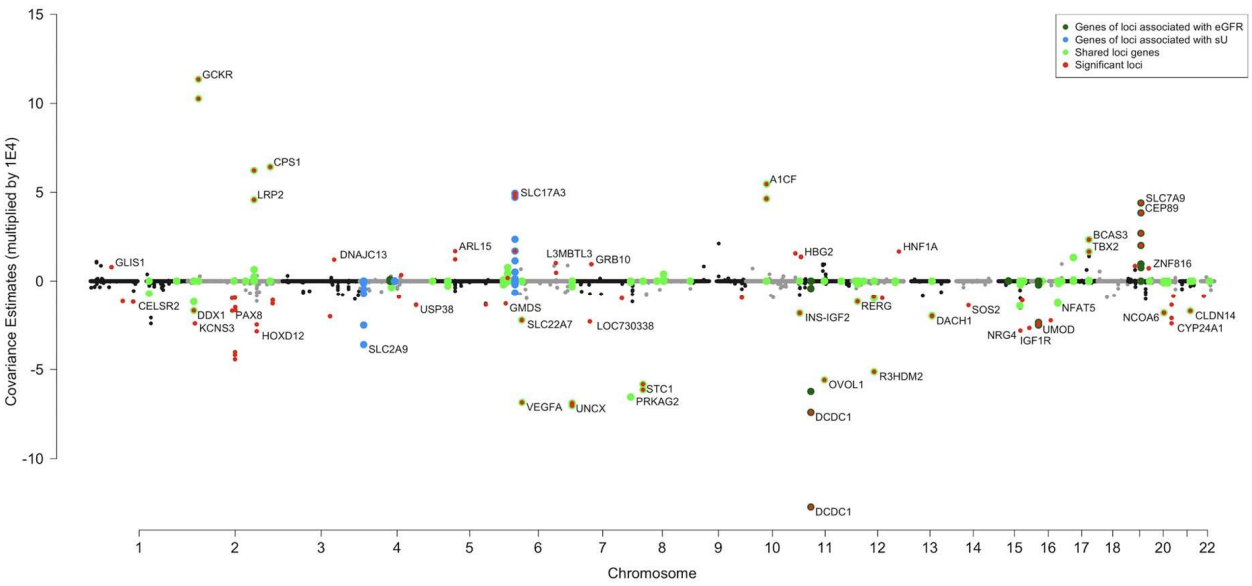


Figure 2: The covariance estimates of overlapping LD windows. Windows are selectively annotated with the gene name of the mid-point SNP of that window. Windows that contained SNPs in loci associated with known eGFR genes are highlighted in dark green, windows that contained SNPs in genes associated with sU are highlighted in blue, and windows that contained SNPs in genes associated with both sU and eGFR (from comparing GWAS, Leask et al., 2020[11]) are highlighted in bright green. Windows significant for genetic covariance are highlighted in red. The covariance estimates were multiplied by 1E4.

We found 134 LD windows with covariance estimates that had a 95% CR excluding zero (Figure 2; Table A1). The number of SNPs in the significant LD windows ranged from one to 56, and the median SNPs per window was 6.0 (22 kbp on average, excluding 12 single-SNP windows). Interestingly, although the global correlation between sU and eGFR is negative[8,9], 49 of the 134 significant windows showed positive genetic covariance directionality, and the remaining 85 were negative.

The 134 significant LD windows often included the same variants and mapped to identical GWAS loci, so we collapsed the 134 windows to 64 unique loci that possessed genetic covariance signal between eGFR and sU (Table A2 and Supplementary Methods). The top 25 distinct loci implicated by the significant windows in terms of covariance magnitude are listed in Table 1. A

graphical representation of the top significant loci is presented in Figure 3.

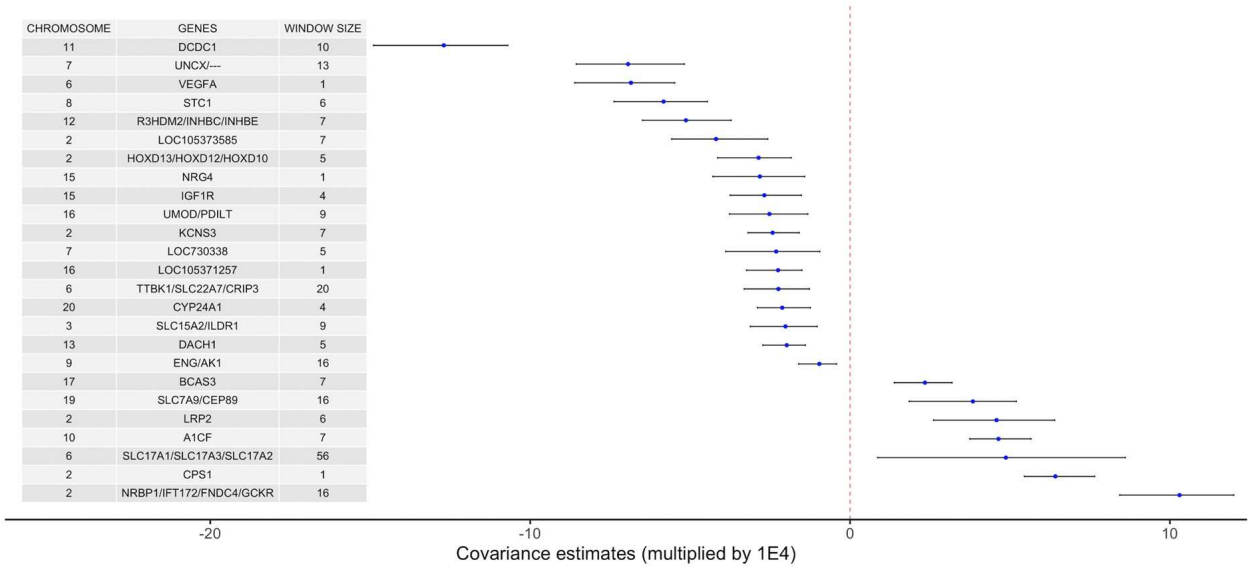


Figure 3: The top 25 shared loci and their covariance estimates with corresponding 95% CRs. The top 25 distinct loci from LD genomic regions with CRs not including zero. The window size indicates the number of SNPs in each window. The covariance estimates and CRs were multiplied by 1E4.

Table 1: The top 25 magnitude genomic windows significant for covariance between sU and eGFR with their chromosome, annotated gene name, number of SNPs and first and last SNP names, estimated covariance [95% CR], and colocalized genes.

Chromosome	Annotated Gene Name	Number of SNPs in the Window and First to Last SNP	Estimated Covariance [95% CR][a]	Colocalized Genes
2	*CPS1*	1 rs1047891	6.42 [5.45, 7.65]	
2	*LRP2*	6 rs41268683-rs2075252	4.58 [2.61, 6.4]	
2	*NRBP1/IFT172/FNDC4/GCKR*	16 Affx-19857019-rs1260333	10.3 [8.43, 12]	*NRBP1*
6	*SLC17A1/SLC17A3/SLC17A2*	56 rs1165196-rs9467632	4.87 [.863, 8.61]	
10	*A1CF*	7 rs12413118-rs61856594	4.64 [3.74, 5.66]	*A1CF*
17	*BCAS3*	7 rs9904048-rs9895661	2.34 [1.38, 3.19]	*CRHBP*, *SH3GL2*
19	*SLC7A9/CEP89*	16 rs78676942-rs11668957	3.84 [1.85, 5.2]	*SLC7A9*, *CLDND2*
2	*LOC105373585*	7 rs11122800-rs35932591	-4.19 [-5.58, -2.57]	
2	*HOXD13/HOXD12/HOXD10*	5 rs847153-rs711818	-2.86 [-4.14, -1.84]	
2	*KCNS3*	7 rs9789415-rs11688124	-2.42 [-3.19, -1.59]	
3	*SLC15A2/ILDR1*	9 rs2049330-rs6438689	-2.02 [-3.12, -1.03]	*SLC15A2*, *CD86*
6	*VEGFA*	1 rs881858	-6.85 [-8.61, -5.48]	*SETD1A*
6	*TTBK1/SLC22A7/CRIP3*	20 rs2651206-rs2242416	-2.24 [-3.31, -1.27]	*SETD1A*
7	*UNCX*	13 rs6950388-rs1880301	-6.94 [-8.56, -5.18]	*PALM2*, *PSMD11*
7	*LOC730338*	5 rs700752-rs12537178	-2.31 [-3.89, -.944]	
8	*STC1*	6 rs62502212-rs1705690	-5.83 [-7.38, -4.46]	*RP11-38H17.1*
11	*OVOL1*	7 rs4014195-rs36008241	-5.59 [-8.13, -3.29]	*PCNX3*, *MAP3K11*, *SCYL1*, *RP-11-770G2.2*, *OVOL1*, *KRT8P26*
11	*DCDC1*	10 rs963837-rs10767873	-12.7 [-14.9, -10.7]	
12	*R3HDM2/INHBC/INHBE*	7 rs73115999-rs507562	-5.13 [-6.49, -3.72]	*KMT2A*, *R3HDM2*, *SFXN5*
13	*DACH1*	5 rs7981995-rs626277	-1.98 [-2.73, -1.39]	

Table 1 (cont'd)

15	*NRG4*	1 rs8024155	-2.82 [-4.29, -1.42]	*MAN2C1*, *PARD3*
15	*IGF1R*	4 rs907808-rs12437561	-2.68 [-3.75, -1.52]	*IGF1R, NRCAM, TRAPPC10*
16	*UMOD/PDILT*	9 rs1123670-rs12917707	-2.52 [-3.77, -1.32]	*ACSM1*, *DNAH3*
16	*LOC105371257*	1 rs12927956	-2.25 [-3.24, -1.5]	
20	*CYP24A1*	4 rs4809954-rs2616278	-2.12 [-2.9, -1.24]	

[a] Estimates and CRs were multiplied by 1E4 for readability.

Gene expression/eQTL analysis

We used COLOC[21] and expression data from The Genotype Tissue Expression (GTEx) project (v8)[24] to identify candidate causal genes at significant local genetic covariance windows between sU and eGFR. Twenty-six of the 64 distinct significant shared loci (41.6%) were shown to modify the expression of candidate causal genes colocalized with the covariance signals (Table A3). Of note are *TRIM6* and *L3MBTL3* in *cis*, which are genes that have a significant covariance signal and a colocalized eQTL that is expressed in the kidney.

Validation

In the related white UK Biobank validation cohort twelve LD windows were significant for genetic covariance between sU and eGFR (Table A1). All of the twelve significant windows were also significant in the main analysis with consistent directionality. The 12 windows condensed to five distinct loci (Table A2), meaning five out the 64 significant distinct loci from the main analysis were also significant in this validation. The sample size of the related cohort is 82.8% smaller (n=57,370) than the unrelated cohort used in the discovery set (n=333,542), so our validation analysis was comparatively underpowered to the main analysis.

Discussion

The goal of this study was to infer the shared genetic architecture of sU (causal for gout), and eGFR (a marker for CKD). Our results highlight genes that may be involved in the observed relationship between the traits. In this study, we estimated local genetic (co)variances between sU and eGFR and identified regions with pleiotropy. This study was based on the large-scale UK Biobank and formal statistical inference from local Bayesian multi-trait models. Our results demonstrated that genetic covariance between eGFR and sU was widespread across the genome. Our method identified 64 distinct LD windows with shared genetic effects between eGFR and sU, the majority of which had negative genetic covariance estimates. We identified 22 distinct novel shared loci, to our knowledge, with significant local genetic covariance for sU and eGFR, including *MMP11*/*SMARCB1*, *ADH1B*, *MIP*/*GLS2*, *ENG*/*AK1*, *EPB41L5*, *KIAA1199*, *CELSR2*, *SOS2*, *KCNS3*, *TET2*, *SMLR1*/*EPB41L2*, *GLIS1*, *KIAA1683*/*JUND*, and *METTL10*/*FAM175B*. Furthermore, 14 distinct loci identified were previously only known to be associated with only one of the two traits, demonstrating that the set of loci contributing to both traits is substantially larger than previously thought. These loci are partially responsible for the comorbidity between hyperuricemia/gout and CKD.

One advantage of the local method that we present here is that it facilitates the identification of genomic windows with opposite signs to the overall negative genetic correlation between eGFR and sU. Out of the significant shared loci, about two-thirds showed negative local genetic covariance estimates. This is consistent with the overall genetic covariance directionality[8,9], indicating that they either contribute to worsening kidney function (decreasing eGFR or increasing sCr) and increasing sU, or vice versa. Interestingly, there were 21 distinct significant shared loci with positive local genetic covariance estimates (about one-third). Positive

covariance indicates that the genomic region either contributes to increasing sU and improved kidney function or decreasing sU and worsening kidney function. Two of the loci with a significant positive signal, *GCKR* and *CPS1*, are mainly expressed in the liver and one, *LRP2*, is mainly expressed in the kidney[24]. One novel shared locus identified in this study consisted of the genes *SLC17A1*, *SLC17A3*, and *SLC17A2*. This large window in chromosome six (56 SNPs, Table 1) had a strong, positive significant covariance signal and *SLC17A1* and *SLC17A3* are urate transporters both linked to gout[25]. The opposite signs of locus-specific genetic covariances are indicative of distinct physiological processes governing the phenotypic expression of urate and eGFR. The loci with positive covariance in particular are excellent candidates for discovering functional mechanisms that simultaneously increase sU and improve kidney function.

Urate transporters *SLC2A9* and *ABCG2* have the largest GWAS effect sizes for sU, accounting for 4-5% of the variance in sU[8,26–29]. However, no windows in *SLC2A9* or *ABCG2* had a 95% CR for local genetic covariance that did not include zero. Our results demonstrate that windows in both *SLC2A9* and *ABCG2* loci are associated with just sU levels but are not pleiotropic regions for sU and eGFR. A similar phenomenon is observed with the eGFR gene *SHROOM3*. That is, none of the windows containing SNPs in *SHROOM3* were significant for local genetic covariance. This exemplifies that the loci driving the genetic correlation between these two traits are not necessarily the leading GWAS hits.

Previous research investigating pleiotropic genetic loci between serum urate and eGFR has implicated loci as shared if signals of association obtained from marginal single-marker regressions (e.g., GWAS) for both traits are colocalized[11]. Leask et al.[11] recently compared overlapping loci between two large GWAS, one of sU and the other kidney function[8,30], and found 36 independent colocalized loci. Our results validate 20 of these 36 loci, and all but three

loci (*DACH1*, *CPS1*, and *INS-IGF2*) had covariance directionality that matched the directionality of effects found by Leask *et al*.[11].

Our covariance approach may have direct implications for assessing causal relationships between exposures using Mendelian randomization (MR). Pleiotropic genetic variants violate assumptions of univariate MR, however, they are useful in multivariable MR that can simultaneously assess the causal effects of multiple risk factors on an outcome[31]. For example, genetic variants from *SLC2A9* and *ABCG2* may be valid instrumental variables to use in MR to test for a causal effect of sU on CKD, however, the loci listed in Table A1 would not. In fact, *SLC22A11* has previously been identified as a pleiotropic variant that may improve kidney function through its activity in raising urate levels[28]. MR has previously been used to show that serum urate is not causal of CKD[32], however, Jordan et al. noted significant pleiotropy in the genetic variants used in their study, which they attempted to counter using MR techniques robust to pleiotropy. Of the 26 SNPs used by Jordan et al., rs1260326 (*GCKR*) and rs17050272 (*LINC01101*) were identified by us as shared, and rs1165151 and rs3741414 were located within one of our significant pleiotropic regions but were not in our genotyping platform.

Our eQTL analysis of the windows significant for local genetic covariance uncovered numerous genes of interest, such as *SLC7A9*, which encodes a solute transporter largely expressed in the small intestine, *A1CF*, which encodes a protein involved in apolipoprotein B synthesis in the liver, and *TRIM6*, which encodes an E3 ubiquitin ligase involved in interferon gamma signaling and innate immune response with high expression levels in the kidney[24]. The genes uncovered from the eQTL analysis will be particularly interesting for future study, as they will likely aid our understanding of the relationship between kidney function and sU.

Through our approach of obtaining local genetic (co)variance estimates from Bayesian

multi-trait models in very large datasets, we have uncovered twenty-two novel shared genetic regions for sU and eGFR. The approach presented in this paper was applied in the context of sU and eGFR, but it could be applied to any pair of traits. While our discovery set sample size is excellent, we lack a dataset of a similar size for the validation. Some regions were validated but not all.

The local shared genomic regions we have uncovered in this study can provide insight into the relationship between hyperuricemia/gout and CKD, elucidating the biological mechanisms underlying the traits. This will help further understanding of the genetic basis of hyperuricemia/gout and CKD.

Data Availability

All data used are secondary and are held in public repositories. This study utilized deidentified data from the UK Biobank where genotype and phenotype data are available to researchers upon registration. The protocol and consent were approved by the UK Biobank's Research Ethics Committee and were conducted under the application number "15326." For eQTL analysis, *cis*- and *trans*-eQTL data were downloaded from the GTEx V8 portal (Carithers and Moore 2015). Supplemental material is available at *G3* online. UK Biobank: https://www.ukbiobank.ac.uk/.

REFERENCES

1. Bikbov, B. et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet 395, 709–733 (2020).

2. Hill, N. R. et al. Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis. PLoS ONE 11, e0158765 (2016).

3. Sun, M. et al. Untangling the complex relationships between incident gout risk, serum urate, and its comorbidities. Arthritis Res Ther 20, 90 (2018).

4. Singh, G., Lingala, B. & Mithal, A. Gout and hyperuricaemia in the USA: prevalence and trends. Rheumatology (Oxford) 58, 2177–2180 (2019).

5. Clarson, L. E. et al. Increased risk of vascular disease associated with gout: a retrospective, matched cohort study in the UK clinical practice research datalink. Ann Rheum Dis 74, 642–647 (2015).

6. Jing, J. et al. Genetics of serum urate concentrations and gout in a high-risk population, patients with chronic kidney disease. Sci Rep 8, 13184 (2018).

7. Zhu, Y., Pandya, B. J. & Choi, H. K. Comorbidities of Gout and Hyperuricemia in the US General Population: NHANES 2007-2008. The American Journal of Medicine 125, 679-687.e1 (2012).

8. Tin, A. et al. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. Nat Genet 51, 1459–1474 (2019).

9. Reynolds, R. J. et al. Genetic correlations between traits associated with hyperuricemia, gout, and comorbidities. Eur J Hum Genet (2021) doi:10.1038/s41431-021-00830-z.

10. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local Genetic Correlation Gives Insights into the Shared Genetic Architecture of Complex Traits. The American Journal of Human Genetics 101, 737–751 (2017).

11. Leask, M. P. et al. The Shared Genetic Basis of Hyperuricemia, Gout, and Kidney Function. Seminars in Nephrology 40, 586–599 (2020).

12. Vilhjálmsson, B. J. et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. The American Journal of Human Genetics 97, 576–592 (2015).

13. Fernando, R., Toosi, A., Wolc, A., Garrick, D. & Dekkers, J. Application of Whole-Genome Prediction Methods for Genome-Wide Association Studies: A Bayesian Approach. JABES 22, 172–193 (2017).

14. Funkhouser, S. A., Vazquez, A. I., Steibel, J. P., Ernst, C. W. & los Campos, G. de. Deciphering Sex-Specific Genetic Architectures Using Local Bayesian Regressions. Genetics 215, 231–241 (2020).

15. Grueneberg, A. & de los Campos, G. BGData - A Suite of R Packages for Genomic Analysis with Big Data. G3 9, 1377–1383 (2019).

16. Affymetrix. Genetic data: Detailed genetic data on half a million people. http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/ (2021).

17. 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).

18. Vazquez, A. I. et al. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. Genetics 192, 1493–1502 (2012).

19. Pérez, P. & de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198, 483–495 (2014).

20. Lehermeier, C., de Los Campos, G., Wimmer, V. & Schön, C.-C. Genomic variance estimates: With or without disequilibrium covariances? J Anim Breed Genet 134, 232–241 (2017).

21. Giambartolomei, C. et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet 10, e1004383 (2014).

22. Fadason, T., Schierding, W., Lumley, T. & O'Sullivan, J. M. Chromatin interactions and expression quantitative trait loci reveal genetic drivers of multimorbidities. Nat Commun 9, 5198 (2018).

23. Genome3d/codes3d-v2. Genome3d (2019).

24. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. Biopreservation and Biobanking 13, 307–308 (2015).

25. Reimer, R. J. SLC17: a functionally diverse family of organic anion transporters. Mol Aspects Med 34, 350–359 (2013).

26. Johnson, R. J. et al. Hyperuricemia, Acute and Chronic Kidney Disease, Hypertension, and Cardiovascular Disease: Report of a Scientific Workshop Organized by the National Kidney Foundation. Am. J. Kidney Dis. 71, 851–865 (2018).

27. Major, T. J., Dalbeth, N., Stahl, E. A. & Merriman, T. R. An update on the genetics of hyperuricaemia and gout. Nat Rev Rheumatol 14, 341–353 (2018).

28. Hughes, K., Flynn, T., de Zoysa, J., Dalbeth, N. & Merriman, T. R. Mendelian randomization analysis associates increased serum urate, due to genetic variation in uric acid transporters, with improved renal function. Kidney Int. 85, 344–351 (2014).

29. Yang, Q. et al. Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. Circ Cardiovasc Genet 3, 523–530 (2010).

30. Wuttke, M. & Köttgen, A. Insights into kidney diseases from genome-wide association studies. Nat Rev Nephrol 12, 549–562 (2016).

31. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am J Epidemiol 181, 251–260 (2015).

32. Jordan, D. M. et al. No causal effects of serum urate levels on the risk of chronic kidney disease: A Mendelian randomization study. PLoS Med. 16, e1002725 (2019).

33. Levey, A. S. et al. A new equation to estimate glomerular filtration rate. Ann. Intern. Med. 150, 604–612 (2009).

34. Coresh, J. et al. Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate. American Journal of Kidney Diseases 39, 920–929 (2002).

APPENDIX A: Chapter 1

Supplementary Tables A1-A3 for Chapter 1 can be found with the publication: https://academic.oup.com/g3journal/article/12/9/jkac158/6649732#supplementary-data.

Supplementary Methods for Lupi et al. 2022

Identification of distantly related samples

We used the R package BGData[15] to compute the expected proportion of allele sharing among UK Biobank individuals with the additive genomic relationship matrix $\boldsymbol{G}$, $\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{tr(\boldsymbol{ZZ'})/n}$, where $\boldsymbol{Z}$ is a matrix of centered genotypes. That is, $\boldsymbol{Z}_{ij} = x_{ij} - 2p_j$ where x_{ij} is the number of copies of the reference allele at the j^{th} loci of the i^{th} individual and p_j is the frequency of the reference allele of the j^{th} loci. In a homogeneous sample, g_{ij} (where $i \neq j$) can be considered as an estimate of the relatedness between subjects i and j. If $g_{ij} \geq 0.1$ they were excluded from the sample.

Phenotypes

eGFR is an indicator of renal function and was used to ascertain CKD. In this study, we defined eGFR using the abbreviated Modification of Diet in Renal Disease (MDRD) equation, which uses fewer variables than others yet performs just as well[33], with a modification to include a calibration factor to correct for the variability of sCr measures across laboratories and time[34]:

$eGFR = 186.3 \times (sCr - 0.24)^{-1.154} \times Age^{-0.203} \times (0.742 \text{ if Female})$.

Defining distinct loci

We condensed our 134 significant windows to 64 distinct, non-overlapping regions. To determine which significant window would represent each region, we first checked if a window's base pair position overlapped with that of a neighboring window. If the windows overlapped, we kept whichever window had the most SNPs. If the number of SNPs in the windows were equal,

we kept the first of the two. This iterative process ended once there were no overlapping neighboring significant windows.

CHAPTER 2: Mapping the relative accuracy of cross-ancestry prediction

This chapter is from a manuscript accepted for publication:

Alexa S. Lupi, Ana I. Vazquez, Gustavo de los Campos. Mapping the Relative Accuracy of Cross-Ancestry Prediction. *Accepted for publication in Nature Communications, 09/11/2024*.

Abstract

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.

Introduction

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].

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:

$$y_i = \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_i \qquad [1]$$

where y_i ($i=1,...,n$ 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):

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i \qquad [2]$$

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:

$$\beta_j = \frac{\mathrm{Cov}(x_{ij}, z_{ij})}{\mathrm{Var}(x_{ij})}\alpha_j \quad [3]$$

where $\mathrm{Var}(x_{ij})$ 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:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathrm{X}}^{-1}\boldsymbol{\Sigma}_{\mathrm{XZ}}\boldsymbol{\alpha} \quad [4]$$

where $\boldsymbol{\Sigma}_{\mathrm{X}}$ is the covariance matrix of marker genotypes and $\boldsymbol{\Sigma}_{\mathrm{XZ}}$ 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]):

$$R^2 = \mathrm{Corr}(\mathbf{x}_{\mathrm{i}}'\boldsymbol{\beta}, \mathbf{z}_{\mathrm{i}}'\boldsymbol{\alpha})^2 = (\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathrm{ZX}}\boldsymbol{\Sigma}_{\mathrm{X}}^{-1}\boldsymbol{\Sigma}_{\mathrm{XZ}}\boldsymbol{\alpha})/(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathrm{z}}\boldsymbol{\alpha})\,. \quad [5]$$

Under perfect LD between markers and QTLs (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 QTLs, 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 casual 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:

$$R_{1\to 2}^2 = \mathrm{Corr}\left(\mathbf{x}_{\mathrm{i}_2}'\boldsymbol{\beta}_1, \mathbf{z}_{\mathrm{i}_2}'\boldsymbol{\alpha}\right)^2 \quad [6]$$

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. (2020)[7], we define the RA of a PGS as:

$$\text{RA} = \frac{R^2_{1\to 2}}{R^2_{1\to 1}} \qquad [7]$$

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 (QTLs), $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ 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 (Figure 4) that, for a given chromosome segment, estimates the distribution of these R-squared values by computing R-squared values 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.

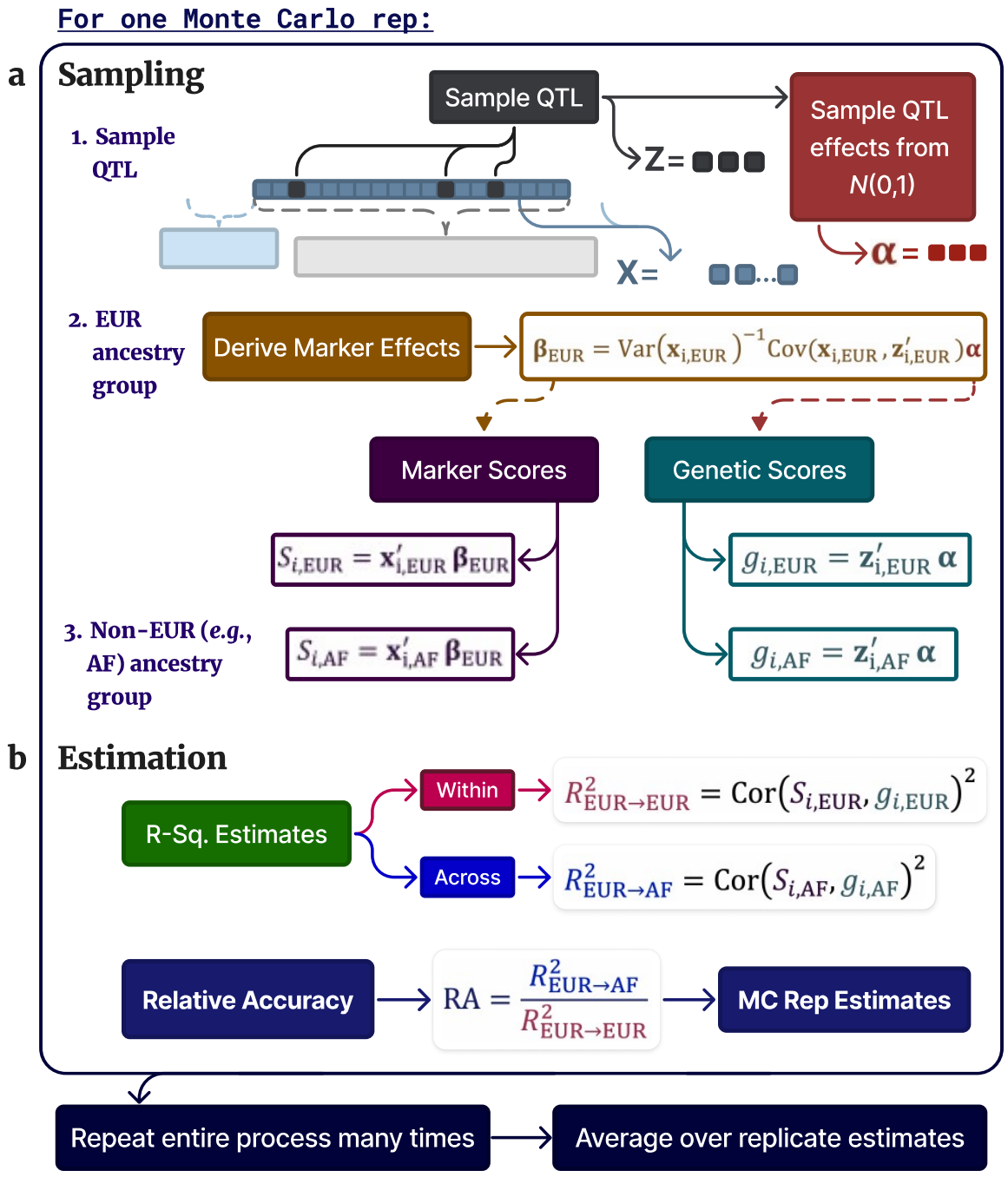


Figure 4: A representation of the MC-ANOVA algorithm. MC-ANOVA uses genetic data from two or more ancestry groups (here, to illustrate we consider European [EUR] and African [AF] ancestry) to estimate the proportion of variance at causal loci explained by EUR-derived marker effects in testing data from 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 ($g_{i_*} = \mathbf{z}'_{\mathbf{i}_*}\boldsymbol{\alpha}$, * = EUR or AF) holds in

Figure 4 (cont'd)

both ancestry groups. Within a short chromosome segment, for one Monte Carlo replicate (MC rep), we sample quantitative trait locus (QTL) (e.g., three) positions ($\mathbf{z_{i_*}}$ for $i = 1,...,n$) at random. 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 $g_{i_*} = \mathbf{z}'_{\mathbf{i}_*}\boldsymbol{\alpha}$. Marker effects are derived from the EUR ancestry group using $\boldsymbol{\beta}_{\mathbf{EUR}} = \mathrm{Var}\left(\mathbf{x}_{\mathbf{i}_{\mathbf{EUR}}}\right)^{-1}\mathrm{Cov}\left(\mathbf{x}_{\mathbf{i}_{\mathbf{EUR}}}, \mathbf{z}'_{\mathbf{i}_{\mathbf{EUR}}}\right)\boldsymbol{\alpha}$, where Var is the variance and Cov is the covariance, and these effects are used to obtain local marker scores for both ancestry groups ($S_{i_*} = \mathbf{x}'_{\mathbf{i}_*}\boldsymbol{\beta}_{\mathbf{EUR}}$). 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. 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.

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 ~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 (Figure 4) 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 2 and Figure B1). 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.

Table 2: 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 ($R^2_{1\to2}$)*	Relative Accuracy ($R^2_{1\to2}/R^2_{1\to1}$)	Standard Error of RA***	Variance in RA Across Segments***
European (EUR)	230,000	---	0.648**	1.000	---	---
African (AF)	3,083	0.120	0.182	0.268	0.016	0.033
Caribbean (CR)	3,343	0.102	0.228	0.340	0.017	0.030
East Asian (EA)	1,329	0.095	0.379	0.564	0.030	0.043
South Asian (SA)	7,919	0.022	0.506	0.771	0.017	0.016

* Subscript 1 always indicates an EUR training or testing set; 2 indicates non-EUR testing; ** $R^2_{1\to1}$; *** Median

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 2). 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 2 and Figure B1). 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 (Figure B2 and Table B1). The variance of RA between segments was slightly smaller when the map was produced with the HapMap variants (Table 2 and Table B1). 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 (Table B2), 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] ($\text{LOA } \% = \frac{1-\text{predicte RA}}{1-\text{empirica RA}} \times 100$, where the predicted RA is defined in [7]).

For most traits, the empirical RA estimates were smaller than the predicted RA (Figure 5 for UK Biobank arrays and Figure B3 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 (Figure 5). 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].

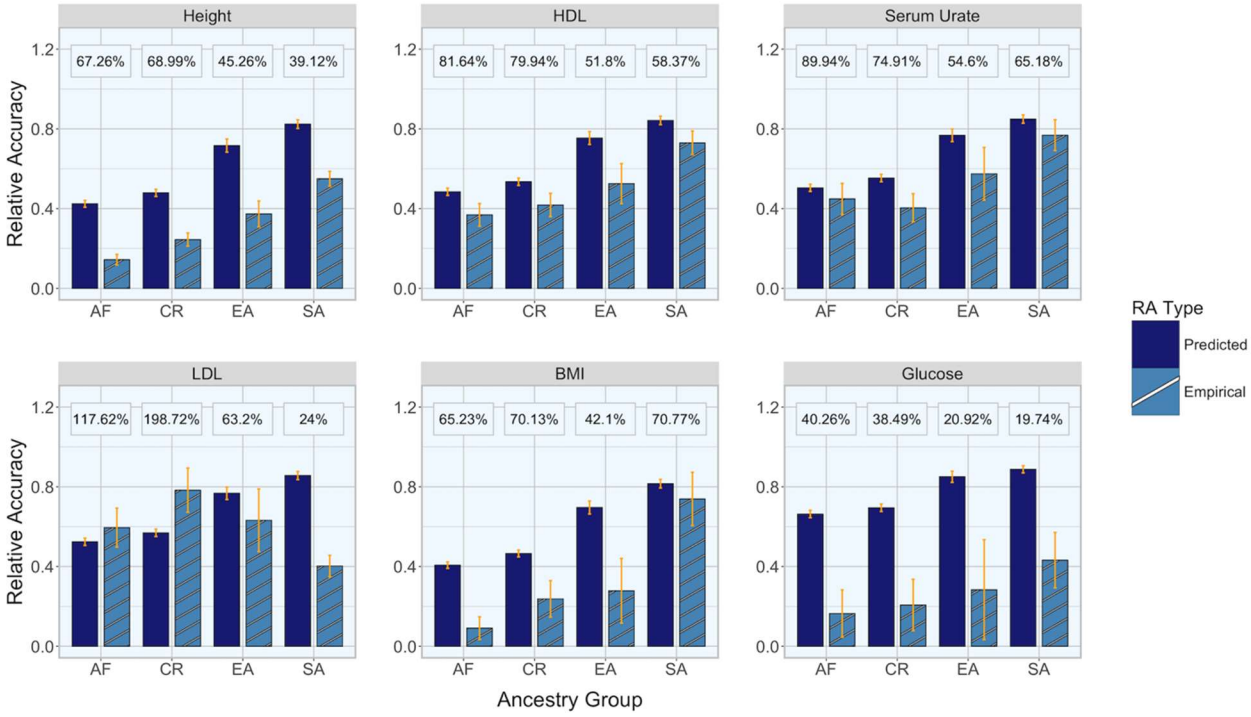


Figure 5: Predicted relative accuracy versus empirical RA with UK Biobank array SNPs. MC-ANOVA predicted relative accuracy (RA) versus empirical RA of European (EUR)-derived polygenic scores when used to predict phenotypes of individuals of non-EUR ancestry (AF, CR, EA, and SA denote African, Caribbean, 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 Table B2. These results are based on SNPs from the UK Biobank arrays; see Figure B3 for results obtained using HapMap SNPs.

We compared our PGS-predicted RA and LOA estimates with those reported by Wang et al., 2020[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 (Figure B4), 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 (Figure 6 and Figure B5 [UK Biobank arrays]; Figure B6 and Figure B7 [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 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.

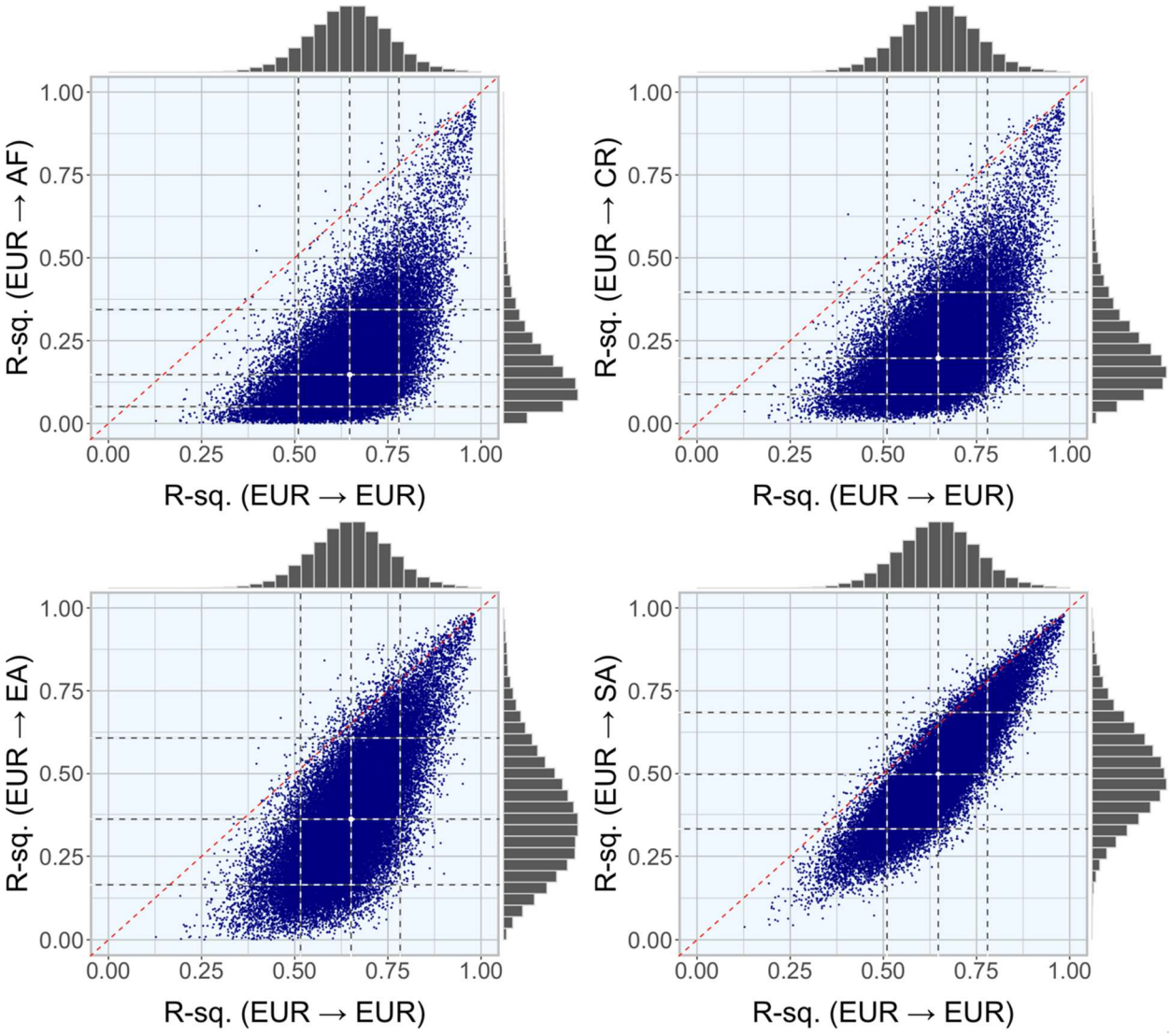


Figure 6: Within- and cross-ancestry R-squared distributions based on UK Biobank array SNPs. Distribution of the cross-ancestry R-squared (R-sq.) versus the within-European (EUR) R-squared for the African (AF), Caribbean (CR), East Asian (EA), and South Asian (SA) ancestry groups obtained when using SNPs from the UK Biobank arrays (see Figure B6 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 10^{th}, 50^{th}, and 90^{th} 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 Figure B6 for results based

Figure 6 (cont'd)
on HapMap SNPs.

The estimates presented in Figure 6 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 Figure B8.

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

The results shown in Figure 6 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 3 for AF; Table B3 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.

Table 3: Estimated relative accuracy (RA) of the SNP segments across the genome grouped by their estimated portability in terms of cross-ancestry R-squared ($R^2_{1\to2}$ 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	$R^2_{1\to2}$ Range	Number of SNPs	Average $R^2_{1\to1}$	Average $R^2_{1\to2}$	Average RA $(R^2_{1\to2}/R^2_{1\to1})$
African (AF)	High	(0.8,1]	(0.26,0.97]	122,135	0.751	0.400	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 Table B3 for other ancestry groups.)

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 (Figure 7 for AF; Figure B9 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 Figure 7). Similar patterns were observed for the other traits (and ancestry groups; Figure B9). 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 (Figure B10). 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.

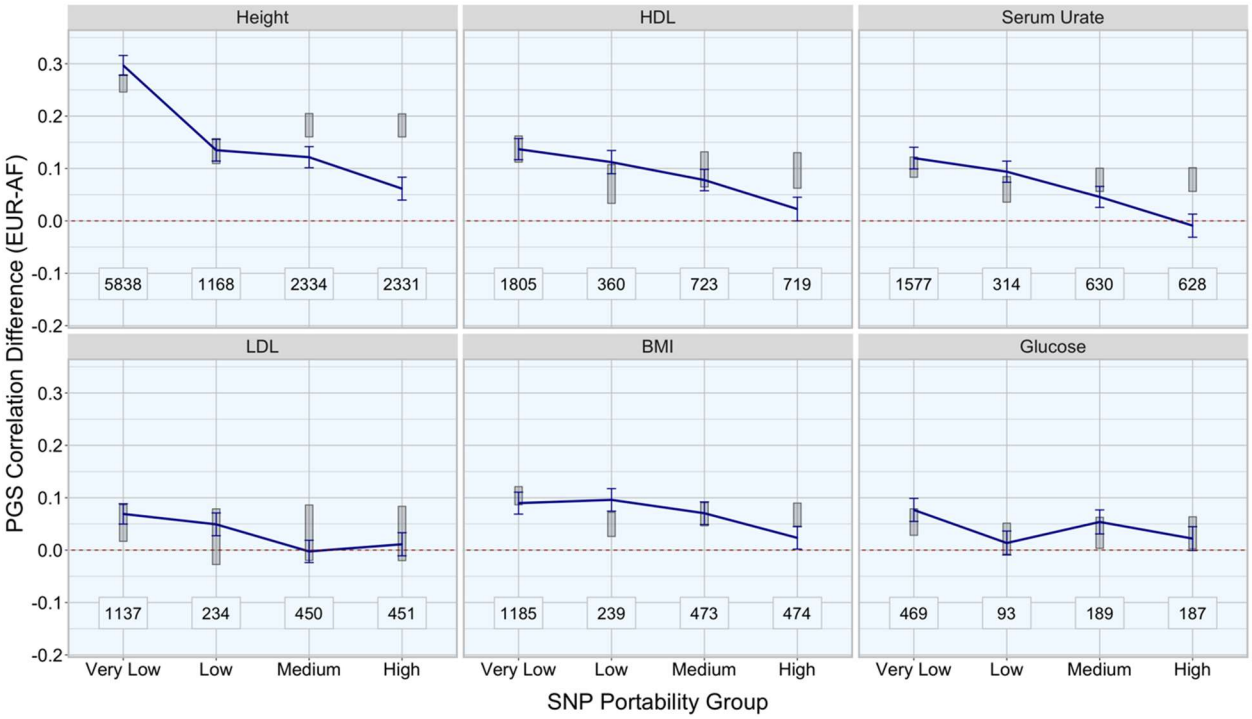


Figure 7: **The difference between polygenic score prediction correlation by SNP portability group based on UK Biobank array SNPs.** The vertical axis represents the difference between the within- and cross-ancestry polygenic score prediction correlations of European (EUR) derived polygenic scores (PGS) for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability ($R^2_{1\to2}$ groupings, Table 3) 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 Table B2. See Figure B9 for results for other

Figure 7 (cont'd)
ancestry groups (Caribbean, East Asian, and South Asian) and Figure B11 for results based on HapMap SNPs.

Using HapMap SNPs did not notably improve PGS local portability over using the called genotypes set (Figure B11). 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 with the map based on SNPs from the UK Biobank arrays (Figure 7, and Figures B9 and B11). 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 Figure 7, we performed a similar analysis to that presented in Figure 4, Figure B9, and Figure B11 classifying SNPs into portability groups using Fst[24] and Wang et al.'s RA method[7] (Figure B12). 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 (Figure B12). 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 (Figure 7).

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; Table B4) 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 (Figure 8) 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 (Figure B13), 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.

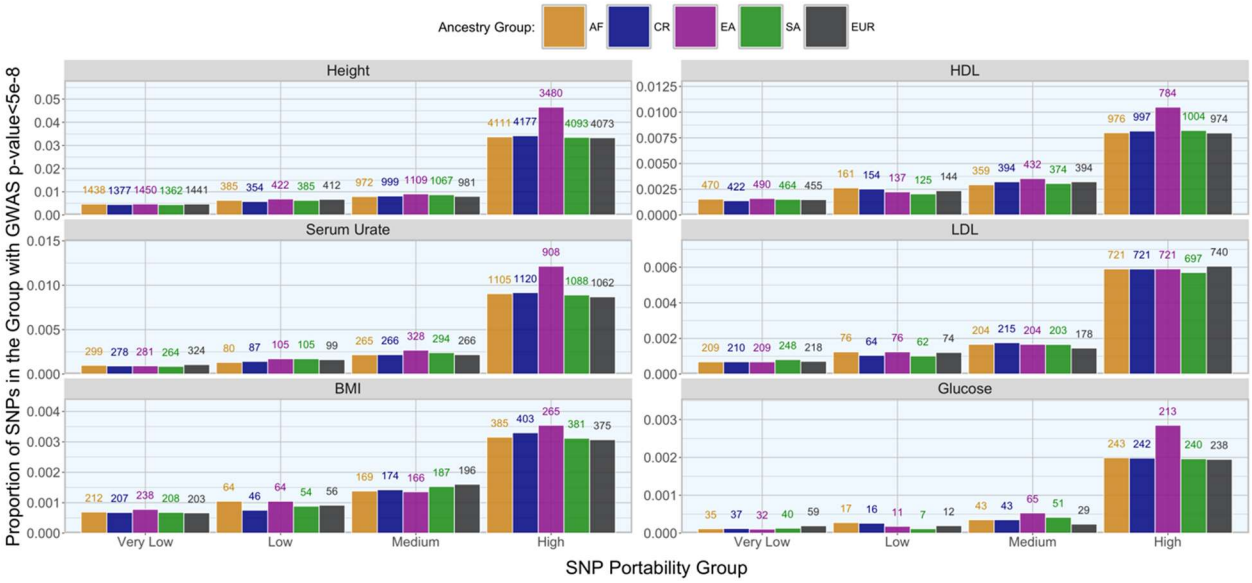


Figure 8: **The proportion of UK Biobank array SNPs that were significantly associated with a trait for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability.** The y-axes give the proportion of SNPs for which a European (EUR)-based genome-wide association study (GWAS) p value (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 (African [AF], Caribbean [CR], East Asian [EA], and South Asian [SA]), the grouping was based on the within-ancestry R-squared [5]. The number of SNPs is noted above each bar and is based on SNPs from the UK Biobank arrays.

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 Biobank, using EUR-derived effects as initial values. Our preliminary results (Table B8) 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 (Table B8).

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 9,628 European American (AEA) and 3,130 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% (Figure B14a), 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 (Figure B14b).

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.'s 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 (Figure B12).

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 (Figure 5), 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 (Figure 7). 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 (Figures B15a, B15b, and B16, 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 (Figures B15c and B15d).

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 (Figure 5) than the HapMap-based map (Figure B3). 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 Figure 7 and Figure B9 with Figure B11). 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 (Figure B5), 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 (Tables B5-B7)

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 (TL) 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-Europeans 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 6,698. Additionally, UK Biobank testing sets included individuals of African ([AF], n=3,083), Caribbean ([CR], n=3,343), East Asian ([EA], n=1,329), and South Asian ([SA], n=7,919) ancestry (Table 2). Ancestral groups were defined by the UK Biobank self-reported Ethnic background (Data-Field 21000[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 9,628 and an African American (AAA) testing set of 3,130 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 ($R^2_{1\to1} = \mathrm{Corr}\left(\mathbf{x}'_{i_1}\boldsymbol{\beta}_1, \mathbf{z}'_{i_1}\boldsymbol{\alpha}\right)^2$ [5]) and AF ($R^2_{1\to2} = \mathrm{Corr}\left(\mathbf{x}'_{i_2}\boldsymbol{\beta}_1, \mathbf{z}'_{i_2}\boldsymbol{\alpha}\right)^2$ [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, Figure 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 Figure B16, 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 $g_{i_1} = \mathbf{z}'_{\mathrm{i}_1}\boldsymbol{\alpha}$ and $g_{i_2} = \mathbf{z}'_{\mathrm{i}_2}\boldsymbol{\alpha}$. Marker effects in ancestry group 1 were computed as $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}'_1\mathbf{X}_1 + \mathbf{I}k)^{-1}\mathbf{X}'_1\mathbf{Z}_1\boldsymbol{\alpha}$, where k = 1e-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 ($\widehat{\boldsymbol{\beta}}_1$) are almost identical to the true population effects ($\boldsymbol{\beta}_1$) because the response used to derive $\widehat{\boldsymbol{\beta}}_1$ ($g_{i_1} = \mathbf{z}'_{\mathrm{i}_1}\boldsymbol{\alpha}$) 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: $R^2_{1\to1} = \mathrm{Corr}\left(\mathbf{x}'_{\mathrm{i}_1}\boldsymbol{\beta}_1, \mathbf{z}'_{\mathrm{i}_1}\boldsymbol{\alpha}\right)^2$ [5] and $R^2_{1\to2} = \mathrm{Corr}\left(\mathbf{x}'_{\mathrm{i}_2}\boldsymbol{\beta}_1, \mathbf{z}'_{\mathrm{i}_2}\boldsymbol{\alpha}\right)^2$ [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 Figure 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 QTLs sampled for a given segment (one, two, three, four, five, and six QTLs 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 (Table B2): 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 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 (Table B2), we conducted a GWAS in the training set described above – distantly related individuals of EUR ancestry (n=230,000) from the UK Biobank (Table B4). 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 Figure 8, 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 $\hat{y}_{i_*} = \mathbf{x}'_{\mathbf{i}_*}\hat{\mathbf{b}}_{\mathbf{1}}$, 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 i^{th} 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: $\mathrm{LOA}\ \% = \frac{1-\mathrm{predicte\ \ RA}}{1-\mathrm{empirica\ \ RA}} \times 100$.

Standard error estimates. We obtained approximate standard error estimates for the PGS correlation coefficients, $\mathrm{Corr}(\hat{y}_{i_*}, y_{i_*})$, using $\sqrt{\frac{1-\mathrm{Corr}(\hat{y}_{i_*}, y_{i_*})^2}{n_*-2}}$, 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 $\mathrm{SE}(\text{empirical RA}) = \sqrt{(\text{empirical RA})^2 * \left(\frac{4\left(1-\mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2\right)}{n_1 * \mathrm{Corr}(\hat{y}_{i_1}, y_{i_1})^2} + \frac{4\left(1-\mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2\right)}{n_2 * \mathrm{Corr}(\hat{y}_{i_2}, y_{i_2})^2}\right)}$. 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 $R^2_{1\to2}$ [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 Table B4, 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 Figure 7 and Figures B9, B11, B12, and B14b, we permuted the grouping labels over 2,000 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., 2020[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 q^{th} window as the average Fst of all core SNPs in that segment, where the Fst for a single SNP is:

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

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., 2020[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:

$$\left(\frac{\overline{r_{1,j}r_{2,j}}\sqrt{\frac{p_{2,j}(1-p_{2,j})}{p_{1,j}(1-p_{1,j})}}}{r_{1,j}^2}\right)^2 \text{x}\frac{p_{1,j}(1-p_{1,j})}{p_{2,j}(1-p_{2,j})}. \quad [9]$$

Here, $p_{*,j}$ is the allele frequency for the j^{th} SNP, and $r_{*,j}$ is the mean correlation between the j^{th} 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 $\hat{y}_{i_*} = \mathbf{x}'_{\mathrm{i}_*}\hat{\mathbf{b}}_1$, 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 $R^2_{1\to2}$ [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 TL 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 ($\widehat{\boldsymbol{\beta}}_1$ derived from D1 of, e.g., European ancestry) as the initial value. Thus, GD-ES produces a sequence of estimates, $\{\widetilde{\boldsymbol{\beta}}_{2(0)}, \widetilde{\boldsymbol{\beta}}_{2(1)}, \ldots, \widetilde{\boldsymbol{\beta}}_{2(s)}\}$, starting with $\widetilde{\boldsymbol{\beta}}_{2(0)} = \widehat{\boldsymbol{\beta}}_1$ (pure cross-ancestry prediction) and moving toward the solution that one would obtain only using D2 ($\widehat{\boldsymbol{\beta}}_2$) after *s* iterations. Early stopping of the GD algorithm renders estimates that are a compromise between $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ and have been shown to improve cross-ancestry PGS prediction compared to using either a purely external ($\widehat{\boldsymbol{\beta}}_1$) or a purely internal ($\widehat{\boldsymbol{\beta}}_2$) 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 $\beta_{2j}^{\text{new}} = \beta_{2j}^{\text{current}} - \text{LR} \times \text{dL}/\text{d}\beta_{2\text{j}}$, 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 $\text{dL}/\text{d}\beta_{2\text{j}}$ is the gradient of the loss function with respect to the j^{th} 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: $\text{LR}_\text{j} = 0.01 \times e^{-3R^2_{1\rightarrow 2,j}}$, where $R^2_{1\rightarrow 2,j}$ 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 $R^2_{1\rightarrow 2,j}$ will have a higher learning rate, thus moving further away from the EUR-derived estimated effect.

For a EUR ancestry group effect, $\widehat{\boldsymbol{\beta}}_1$, 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 Table B8. 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 (Table B8), 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 2 and Figure B1 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.

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.

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.

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 Statement

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 Statement

No authors have any competing interests to declare.

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. *Biodemography 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 Bioinformatics* **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. Theoretical Population Biology **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).

APPENDIX B: Chapter 2

Supplementary Materials

Mapping the relative accuracy of cross-ancestry prediction

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

[1] Department of Epidemiology and Biostatistics, Michigan State University (MSU), East Lansing, Michigan 48824, United States.

[2] Institute for Quantitative Health Science and Engineering, Systems Biology, MSU.

[3] Department of Statistics and Probability, MSU.

Supplementary Methods

Derivation of the Within- and cross-ancestry R-squared parameters

In this note we present a step-by-step derivation of the within- and cross-ancestry R-squared parameters. The note expands what is presented in the main text and shows that these parameters (and functions thereof such as the relative accuracy) are functions of (i) allele frequencies (which impact the variance of genotypes at each locus), (ii) linkage-disequilibrium patterns (which impacts the correlation of genotypes at markers and causal loci), and (iii) the effects of causal alleles (which in MC-ANOVA we integrate out by averaging over MC replicates).

Under the framework described in the Results section, for ancestry group 1 the causal model (expression [1] in the main text) is:

$$y_{i_1} = \mathbf{z}'_{\mathbf{i}_1}\boldsymbol{\alpha} + \varepsilon_{i_1} \qquad \text{(SE1)}$$

where y_{i_1} is the phenotype of the i^{th} individual (from ancestry group 1), $\mathbf{z}_{\mathbf{i}_1}$ is the vector of QTL genotypes from the same individual, and $\boldsymbol{\alpha}$ is the vector of QTL effects. To isolate the effects of LD and allele frequency differences between ancestry groups in cross-ancestry prediction, MC-ANOVA assumes that the same causal model holds in both ancestries and, hence, $\boldsymbol{\alpha}$ does not have a population subscript (and S1 is $y_{i_2} = \mathbf{z}'_{\mathbf{i}_2}\boldsymbol{\alpha} + \varepsilon_{i_2}$ for ancestry group 2).

The instrumental model (expression [2] in the main text) for ancestry group 1 can be written as:

$$y_{i_1} = \mathbf{x}'_{\mathbf{i}_1}\boldsymbol{\beta}_1 + \varepsilon_{i_1} \qquad \text{(SE2)}$$

where $\mathbf{x}_{\mathbf{i}_1}$ is the vector of markers/SNPs for an individual *i* from ancestry group 1.

Assuming the causal model of SE1 and that the errors in SE2 are uncorrelated with markers, marker effects in population 1 are

$$\boldsymbol{\beta}_1 = \mathrm{Var}\left(\mathbf{x}_{i_1}\right)^{-1}\mathrm{Cov}\left(\mathbf{x}_{i_1}, \mathbf{z}'_{i_1}\right)\boldsymbol{\alpha} = \boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha} \qquad \text{(SE3)}$$

where $\boldsymbol{\Sigma}_{X_1}$ is the covariance matrix among markers and $\boldsymbol{\Sigma}_{X_1Z_1}$ is the covariance matrix between the markers and QTL, both in ancestry group 1; these matrices are functions of allele frequencies and LD patterns in population 1.

Therefore, the squared correlation between the true genetic values ($\mathbf{z}'_{i_1}\boldsymbol{\alpha}$) and the marker-predicted genetic scores ($\mathbf{x}'_{i_1}\boldsymbol{\beta}_1$) in population 1 (expression [5] in the main text) is:

$$\mathrm{Corr}\left(\mathbf{x}'_{i_1}\boldsymbol{\beta}_1, \mathbf{z}'_{i_1}\boldsymbol{\alpha}\right)^2 = \frac{\left[\boldsymbol{\beta}_1'\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}\right]^2}{\left[\boldsymbol{\beta}_1'\boldsymbol{\Sigma}_{X_1}\boldsymbol{\beta}_1\right]\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1}\boldsymbol{\alpha}\right]}. \qquad \text{(SE4)}$$

Using $\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}$ (SE3) in SE4 we get:

$$\mathrm{Corr}\left(\mathbf{x}'_{i_1}\boldsymbol{\beta}_1, \mathbf{z}'_{i_1}\boldsymbol{\alpha}\right)^2 = \frac{\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}\right]^2}{\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}\right]\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1}\boldsymbol{\alpha}\right]} = \frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1}\boldsymbol{\alpha}}. \qquad \text{(SE5)}$$

Conceptually, this can be elucidated if we consider the QTL ($\mathbf{Z}_1$) and the markers ($\mathbf{X}_1$) to have some multivariate distribution with covariance $\boldsymbol{\Sigma}_1 = \begin{bmatrix} \boldsymbol{\Sigma}_{Z_1} & \boldsymbol{\Sigma}_{Z_1X_1} \\ \boldsymbol{\Sigma}_{X_1Z_1} & \boldsymbol{\Sigma}_{X_1} \end{bmatrix}$. Then the conditional covariance of the QTL given the markers, $\mathrm{Cov}(\mathbf{Z}_1|\mathbf{X}_1)$, is known to be the Schur complement of $\boldsymbol{\Sigma}_{X_1}$, which is $\boldsymbol{\Sigma}_{Z_1} - \boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}$. Thus, the term $\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}$ captures the variance and covariance from QTL explained by regression on markers.

Therefore, we define the within-ancestry R-squared as the squared correlation in (SE5):

$$R^2_{1\to1} = \mathrm{Corr}\left(\mathbf{x}'_{i_1}\boldsymbol{\beta}_1, \mathbf{z}'_{i_1}\boldsymbol{\alpha}\right)^2 = \boxed{\frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}_{X_1}^{-1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1}\boldsymbol{\alpha}}}. \qquad \text{(SE6)}$$

This is equivalent to what is shown in expression [5] in the main text (including ancestry group indices here).

To derive the cross-ancestry correlation, we define the following (co)variance matrices for ancestry group 2:

$$\mathrm{Var}\left(\mathbf{x}_{i_2}\right) = \boldsymbol{\Sigma}_{X_2},\ \mathrm{Cov}\left(\mathbf{x}_{i_2}, \mathbf{z}'_{i_2}\right) = \boldsymbol{\Sigma}_{X_2Z_2},\ \text{and}\ \mathrm{Var}\left(\mathbf{z}_{i_2}\right) = \boldsymbol{\Sigma}_{Z_2}\ . \tag{SE7}$$

Using the marker effects from ancestry group 1 to predict genetic scores in ancestry group 2 ($\mathbf{z}'_{i_2}\boldsymbol{\alpha}$), the cross-ancestry R-squared is:

$$\mathrm{Corr}\left(\mathbf{x}'_{i_2}\boldsymbol{\beta}_1, \mathbf{z}'_{i_2}\boldsymbol{\alpha}\right)^2 = \frac{\left[\boldsymbol{\beta}'_1\mathrm{Cov}\left(\mathbf{x}_{i_2}, \mathbf{z}'_{i_2}\right)\boldsymbol{\alpha}\right]^2}{\mathrm{Var}\left(\mathbf{x}'_{i_2}\boldsymbol{\beta}_1\right)\mathrm{Var}\left(\mathbf{z}'_{i_2}\boldsymbol{\alpha}\right)} = \frac{\left[\boldsymbol{\beta}'_1\boldsymbol{\Sigma}_{X_2Z_2}\boldsymbol{\alpha}\right]^2}{\left[\boldsymbol{\beta}'_1\boldsymbol{\Sigma}_{X_2}\boldsymbol{\beta}_1\right]\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_2}\boldsymbol{\alpha}\right]}\ . \tag{SE8}$$

Replacing $\boldsymbol{\beta}_1$ with the right-hand side of (SE3) we get:

$$R^2_{1\to 2} = \mathrm{Corr}\left(\mathbf{x}'_{i_2}\boldsymbol{\beta}_1, \mathbf{z}'_{i_2}\boldsymbol{\alpha}\right)^2 = \frac{\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}^{-1}_{X_1}\boldsymbol{\Sigma}_{X_2Z_2}\boldsymbol{\alpha}\right]^2}{\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_1X_1}\boldsymbol{\Sigma}^{-1}_{X_1}\boldsymbol{\Sigma}_{X_2}\boldsymbol{\Sigma}^{-1}_{X_1}\boldsymbol{\Sigma}_{X_1Z_1}\boldsymbol{\alpha}\right]\left[\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{Z_2}\boldsymbol{\alpha}\right]}\ . \tag{SE9}$$

It is interesting to compare the quadratic forms involved in the within- and cross-ancestry R-squared parameters (expressions SE6 and SE9). If the variances of genotypes at individual loci and the LD patterns are the same in both ancestry groups (i.e., if $\boldsymbol{\Sigma}_{Z_1} = \boldsymbol{\Sigma}_{Z_2}$, $\boldsymbol{\Sigma}_{X_1} = \boldsymbol{\Sigma}_{X_2}$, and $\boldsymbol{\Sigma}_{X_1Z_1} = \boldsymbol{\Sigma}_{X_2Z_2}$), the two R-squared values are identical.

The variance of MC-ANOVA predicted relative accuracies

Following Wang et al.[1], the variance of a ratio is approximately:

$$\mathrm{Var}(x/y) \approx \left(\frac{E(x)}{E(y)}\right)^2 \left[\frac{\mathrm{Var}(x)}{E(x)^2} + \frac{\mathrm{Var}(y)}{E(y)^2} - 2\left(\frac{\mathrm{Cov}(x,y)}{E(x)E(y)}\right)\right]. \quad \text{(SE10)}$$

For notation purposes, let relative the accuracy (RA, [7]) be denoted as:

$$\mathrm{RA} = \frac{R^2_{1\to 2}}{R^2_{1\to 1}} = \frac{R^2_2}{R^2_1}. \quad \text{(SE11)}$$

Plugging (SE11) into the general formula in (SE10), we obtain:

$$\mathrm{Var}\left(\frac{R^2_2}{R^2_1}\right) \approx \left(\frac{E(R^2_2)}{E(R^2_1)}\right)^2 \left[\frac{\mathrm{Var}(R^2_2)}{E(R^2_2)^2} + \frac{\mathrm{Var}(R^2_1)}{E(R^2_1)^2} - 2\left(\frac{\mathrm{Cov}(R^2_2,R^2_1)}{E(R^2_2)E(R^2_1)}\right)\right]. \quad \text{(SE12)}$$

Replacing the expected value with the MC-ANOVA estimate, $E(R^2_*) = R^2_*$ (* = 1, 2), and assuming $\mathrm{Cov}(R^2_2, R^2_1) = 0$ since the ancestry cohorts are independent of one another we get:

$$\mathrm{Var}\left(\frac{R^2_2}{R^2_1}\right) \approx \left(\frac{R^2_2}{R^2_1}\right)^2 \left[\frac{\mathrm{Var}(R^2_2)}{R^4_2} + \frac{\mathrm{Var}(R^2_1)}{R^4_1}\right]. \quad \text{(SE13)}$$

For the $\mathrm{Var}(R^2_*)$, we must consider two sources of uncertainty, the sampling variance of the estimator (resulting from the use of a finite sample size) and the Monte Carlo error; therefore

$$\mathrm{Var}(R^2_*) = \mathrm{MC\_variance}(R^2_*) + \left(\frac{4}{n_*}\right) R^2_*(1 - R^2_*), \quad \text{(SE14)}$$

where the MC_variance component is the variance of the estimate over the 300 Monte Carlo replications, and the sample variance component is from the same Taylor series-based derivation as used in the empirical RA variance approximation (see 'Standard error estimates' in Methods). Thus:

$$\mathrm{Var}\left(\frac{R^2_2}{R^2_1}\right) \approx \left(\frac{R^2_2}{R^2_1}\right)^2 \left[\frac{\mathrm{MC\_variance}(R^2_2) + \left(\frac{4}{n_2}\right)R^2_2(1-R^2_2)}{R^4_2} + \frac{\mathrm{MC\_variance}(R^2_1) + \left(\frac{4}{n_1}\right)R^2_1(1-R^2_1)}{R^4_1}\right]. \quad \text{(SE15)}$$

The standard error bars presented in the portability maps are the square root of (SE15).

Supplementary Data

Supplementary Figures

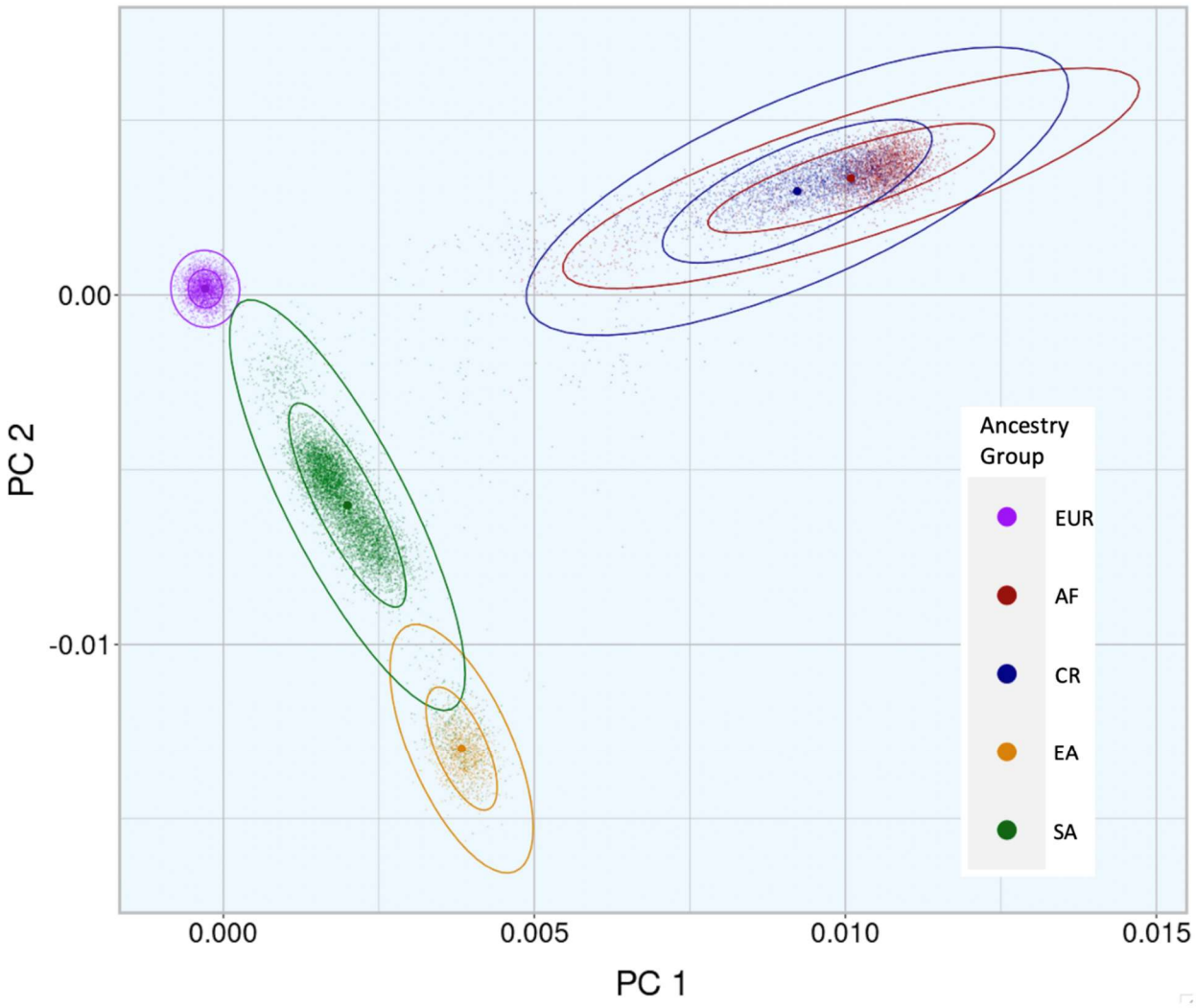


Figure B1: Loadings in the first two SNP-derived principal components (PC) colored by ancestry. The inner and outer ellipses represent the 68th and 99th percentile of the PC loadings of each ancestry (European [EUR], African [AF], Caribbean [CR], East Asian [EA], and South Asian [SA]). A random sample of 3,000 EUR ancestry individuals were selected for plotting.

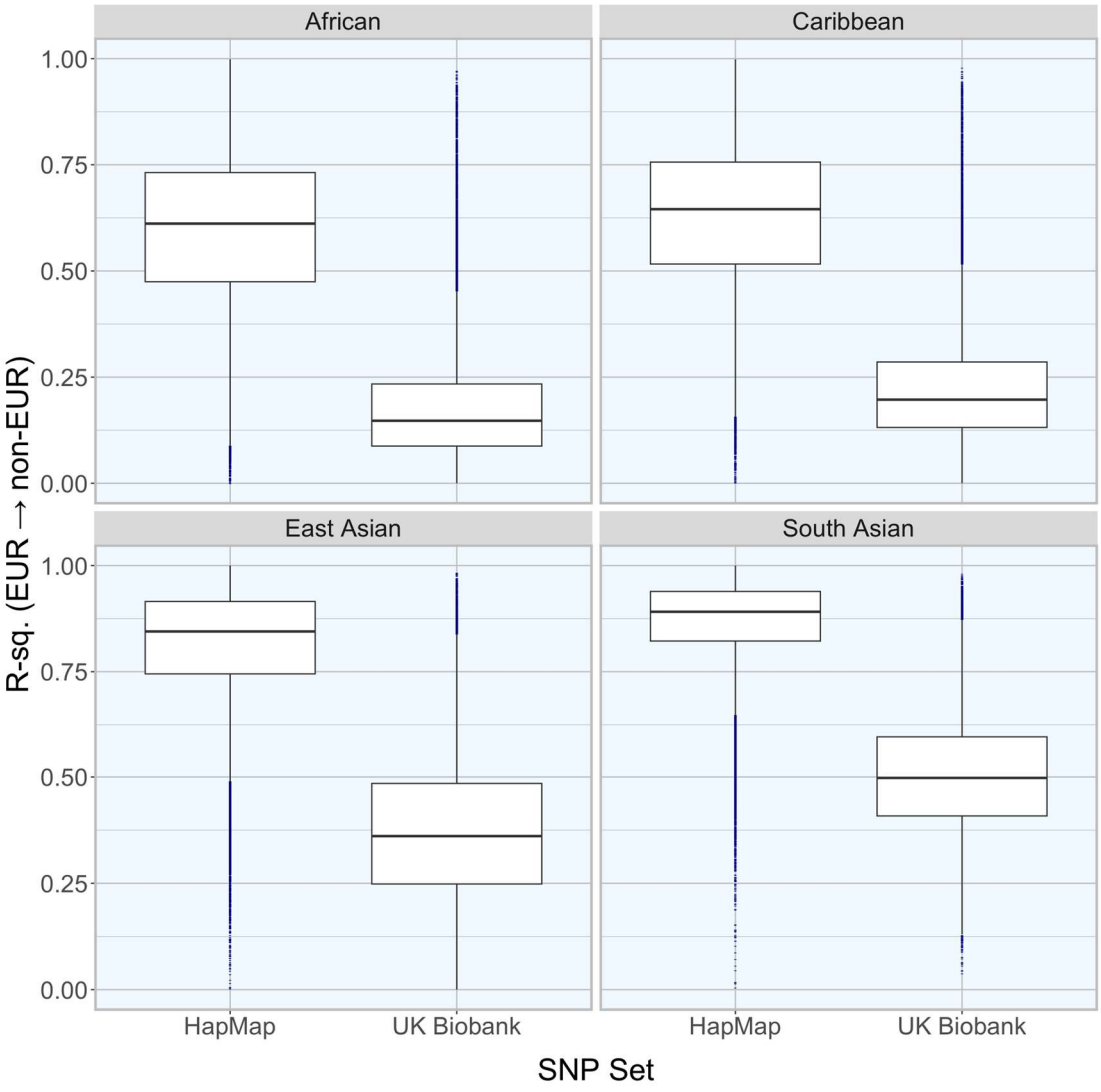


Figure B2: Comparing the UK Biobank and HapMap SNP set estimates. The cross-ancestry (European [EUR] to non-EUR) R-squared [6] distributions for each SNP set (HapMap variants compared to UK Biobank arrays) and each ancestry group (African, Caribbean, East Asian, and South Asian). Each panel displays a different non-EUR ancestry group. The bottom line of each box represents the first quartile, the next line is the median, and the top line is the third quartile. Verticle lines extend from the first (third) quartiles to the minimum (maximum) and outliers are represented by blue points.

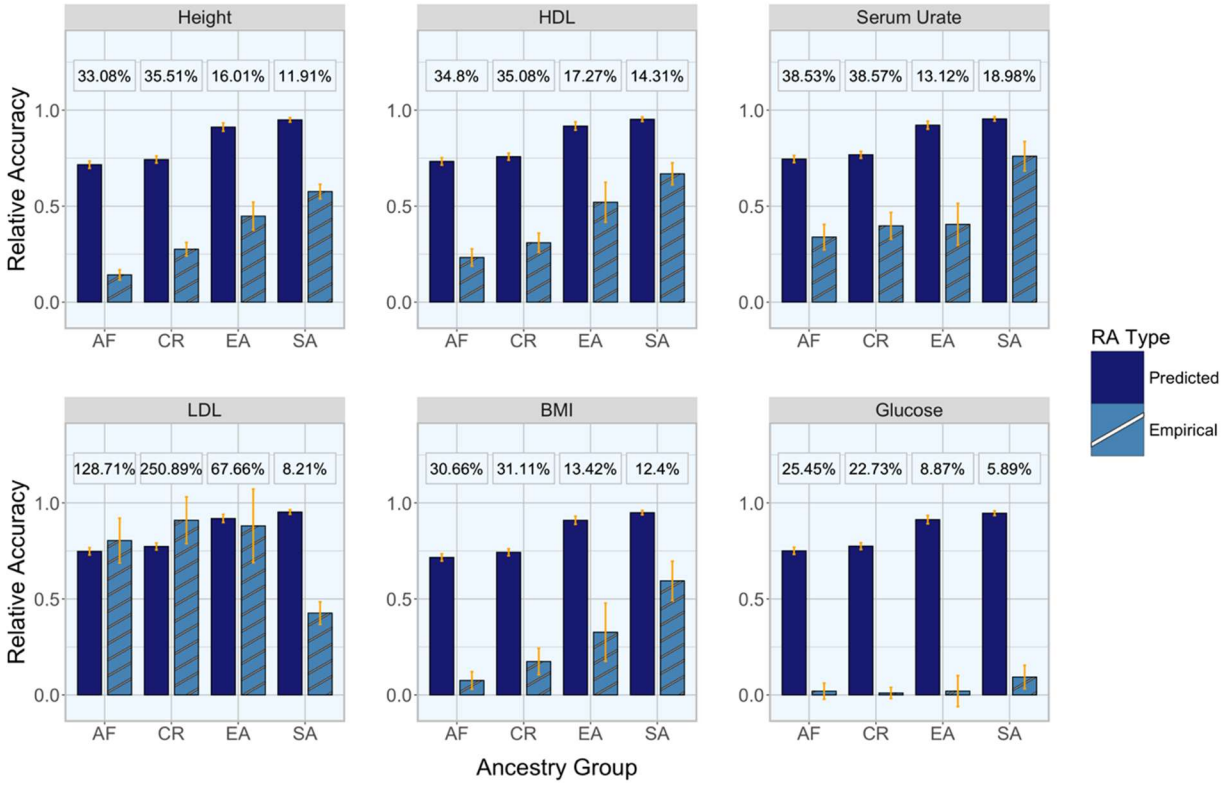
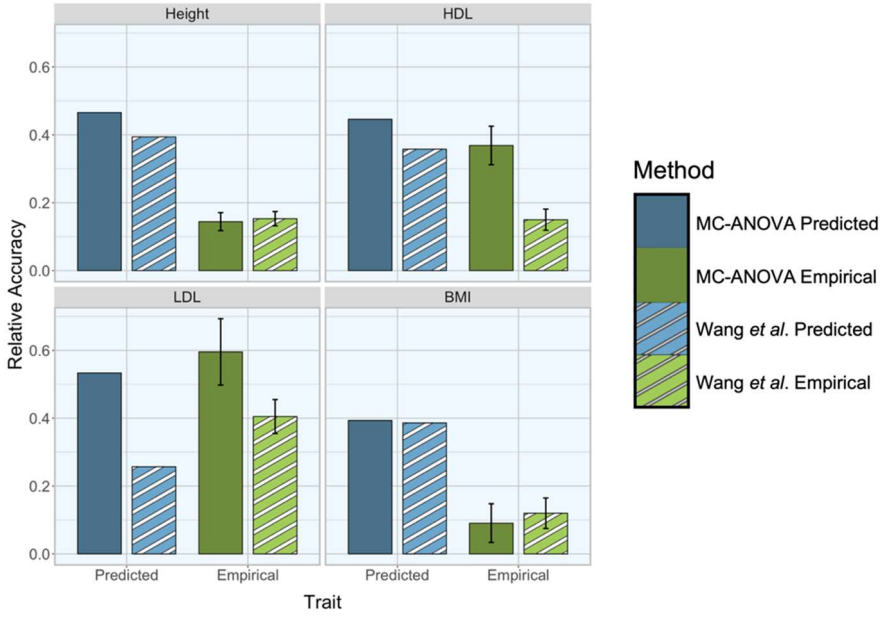


Figure B3: MC-ANOVA predicted relative accuracy (RA) versus empirical RA using SNPs from the HapMap variants. Predicted compared to empirical RA of European (EUR)-derived polygenic scores when used to predict phenotypes of individuals of non-EUR ancestry (AF, CR, EA, and SA denote African, Caribbean, East Asian, and South Asian ancestry). Each panel displays a different phenotype. 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 is shown for each mean RA estimate (derivation details for predicted RA 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 Table B2. See Figure 5 for results based on SNPs from the UK Biobank arrays.

a: Relative accuracy.



b: Loss of accuracy.

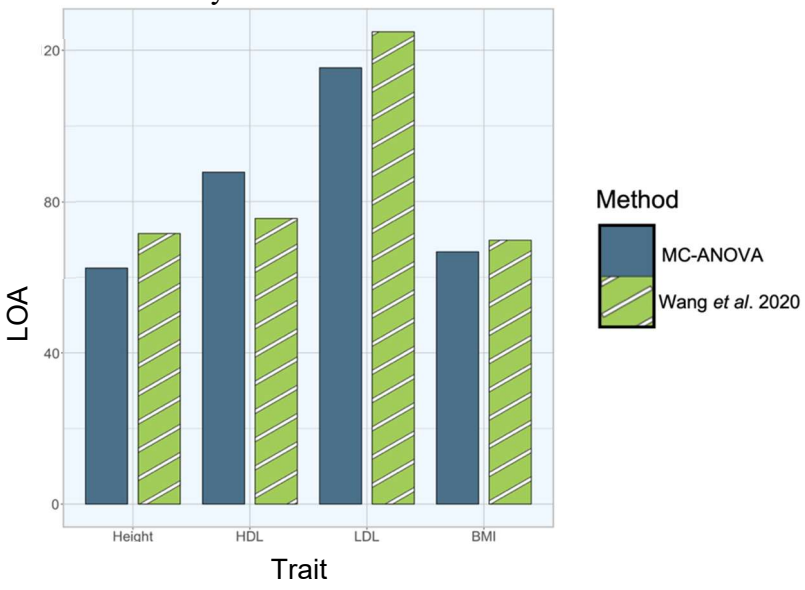
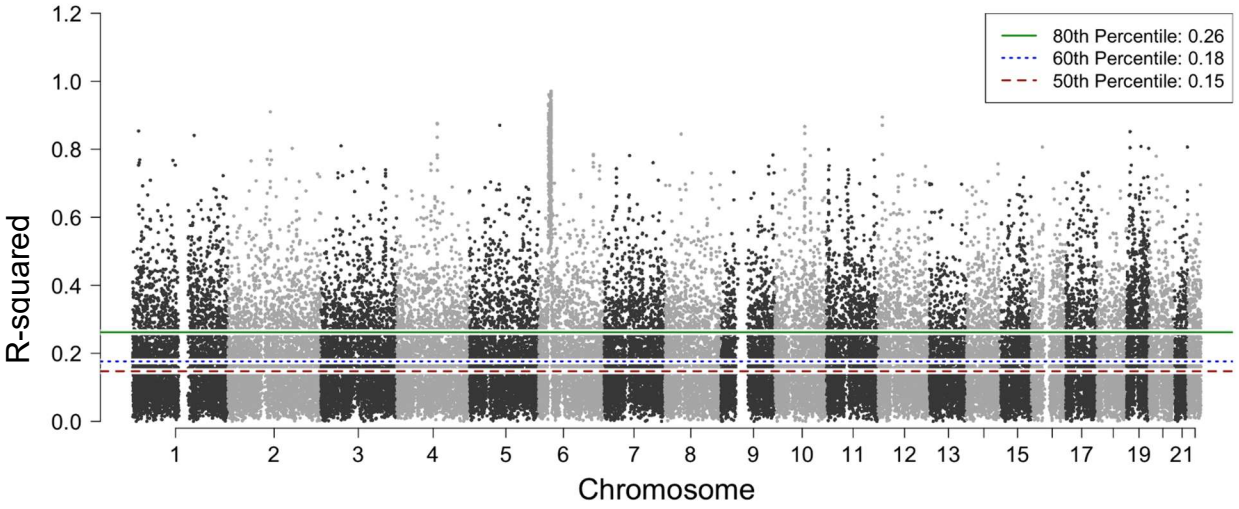


Figure B4: Polygenic score relative accuracy and loss of accuracy by method. (**a**) Predicted and empirical relative accuracy (RA) for four traits (height, high-density lipoprotein [HDL], low-density lipoprotein [LDL], and body mass index [BMI]) in the African (AF) ancestry group (compared to European). (**b**) Loss of accuracy (LOA) explained by genome differentiation for four traits in the AF group by method: MC-ANOVA (using SNPs from the UK Biobank arrays) and the values reported in Wang et al. 2020[1]. The sample sizes used to derive the standard errors for MC-ANOVA mean RA are in Table B2.

a: African ancestry group.



b: Caribbean ancestry group.

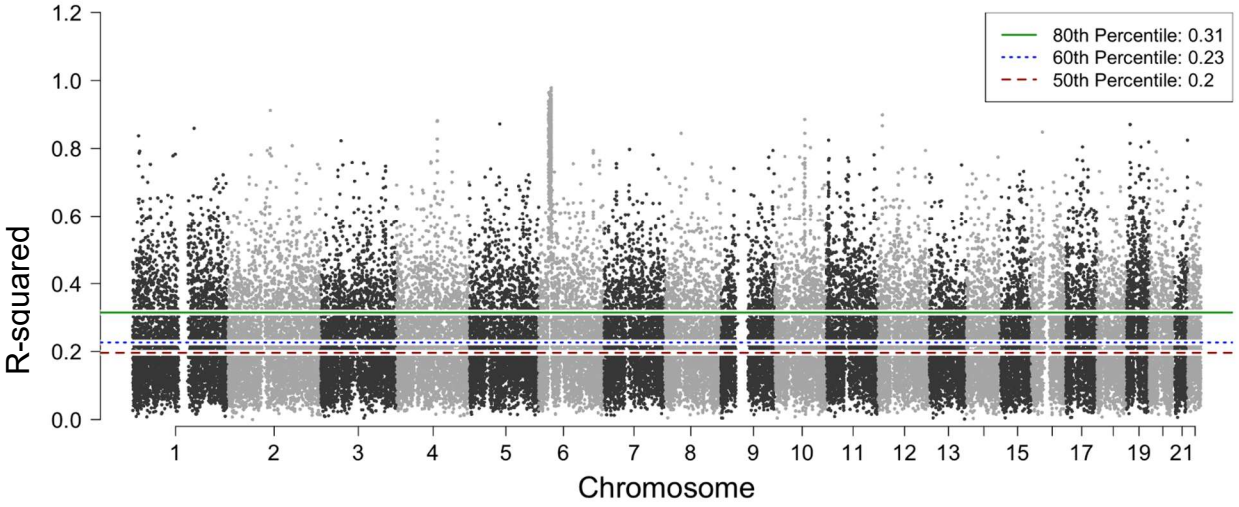
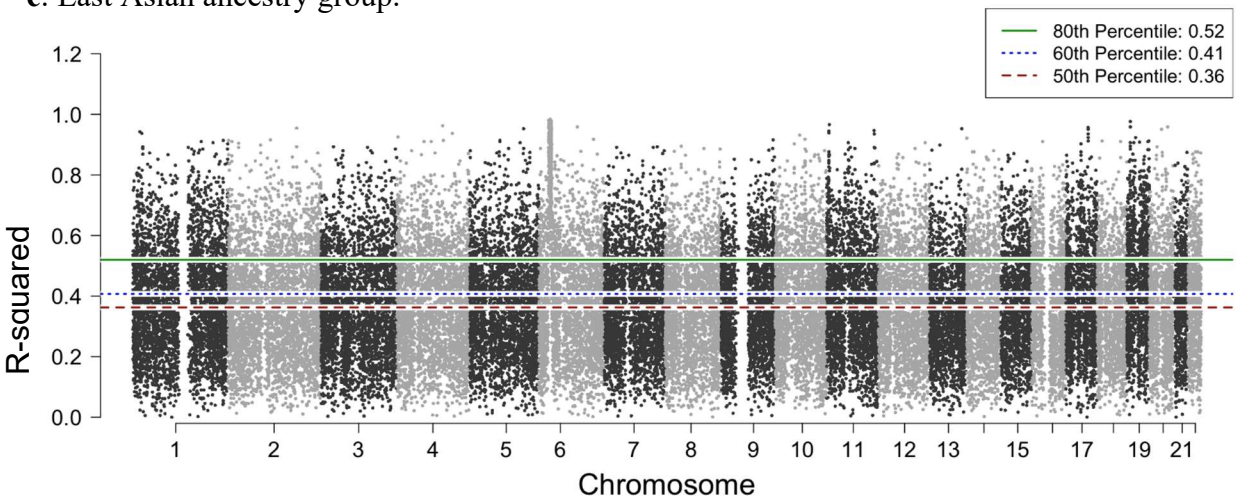


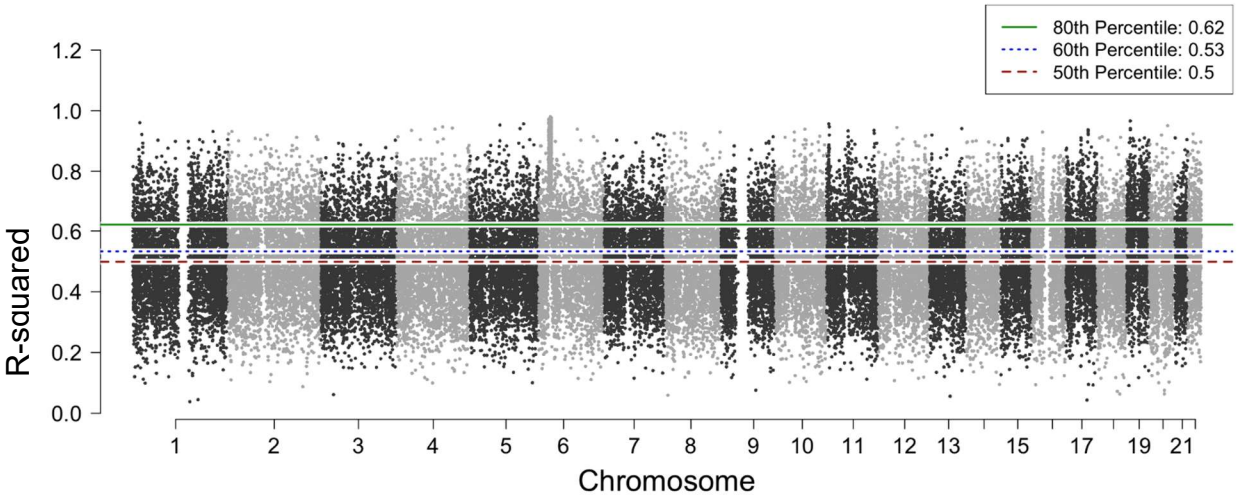
Figure B5: Cross-ancestry MC-ANOVA predicted R-squared by chromosome and position based on SNPs from the UK Biobank arrays. Each dot represents the estimated $R^2_{1\to 2}$ [6] for a chromosome segment (with an average length of 45 Kbp) by the ancestry group of the testing data: AF=African (**a**), CR=Caribbean (**b**), EA=East Asian (**c**), and SA=South Asian (**d**). Ancestry group 1 is European (EUR). The green line is the 80th percentile value of $R^2_{1\to 2}$, the blue short dashed line is the 60th percentile, and the red long dashed line is the 50th percentile. See Figure B7 for results based on HapMap SNPs.

Figure B5 (cont'd)

c: East Asian ancestry group.



d: South Asian ancestry group.



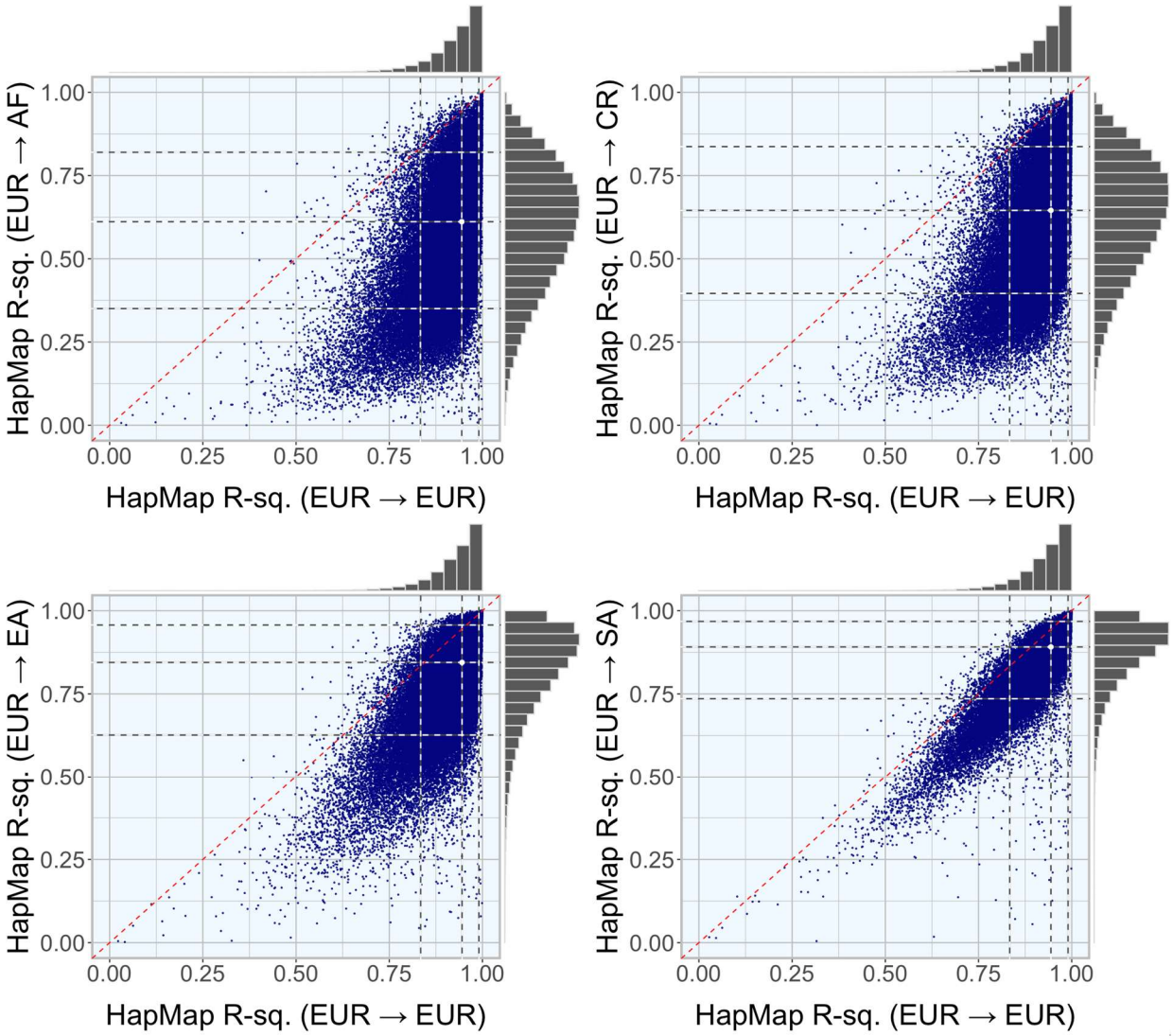
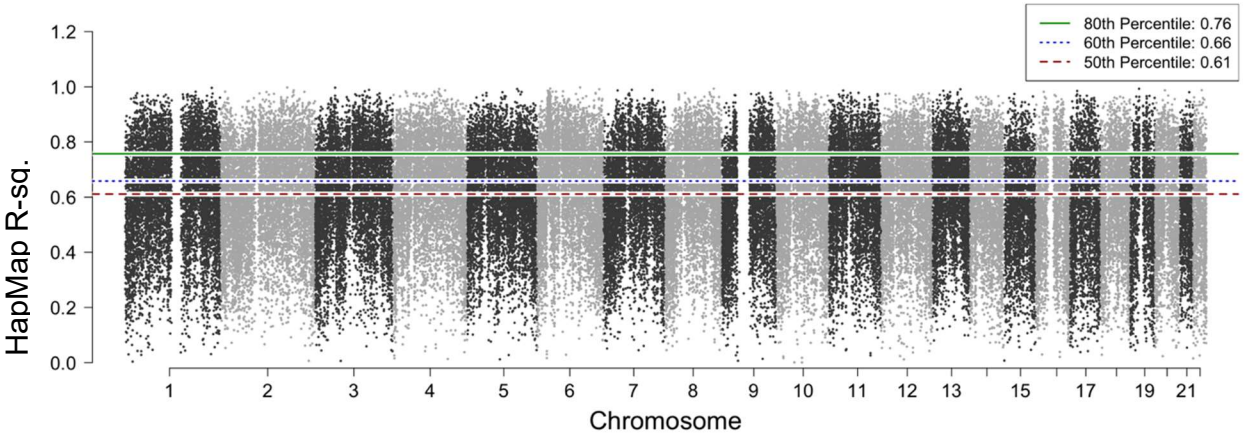


Figure B6: Within- and cross-ancestry R-squared distributions based on HapMap SNPs. Distribution of the cross-ancestry R-squared (R-sq.) versus the within-European (EUR) R-squared for the African, Caribbean, East Asian, and South Asian (AF, CR, EA, and SA, respectively) ancestry groups. Each panel displays a different non-EUR ancestry group. Each point represents a small chromosome segment (23 Kbp on average). 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 Figure 6 for results based on SNPs from the UK Biobank arrays.

a: HapMap SNPs African ancestry group.



b: HapMap SNPs Caribbean ancestry group.

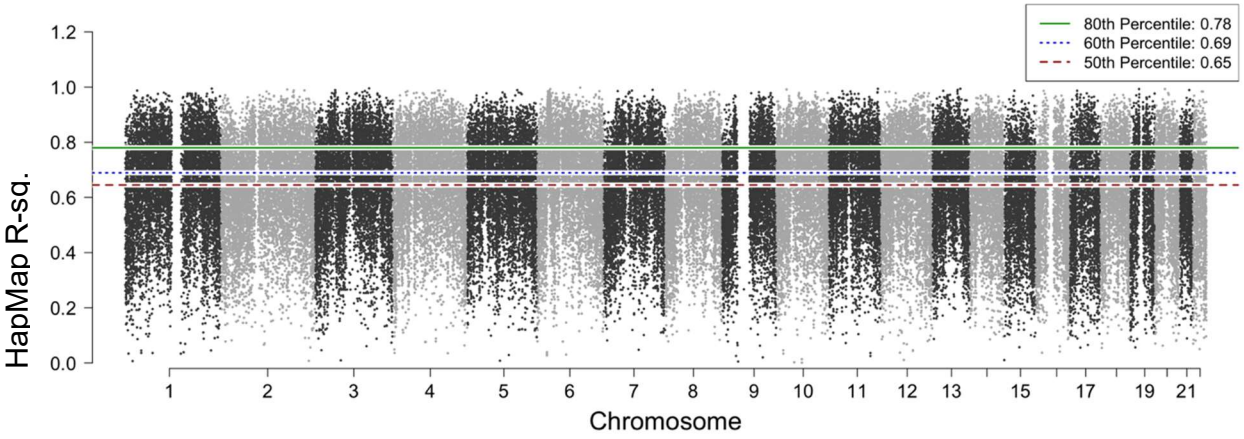
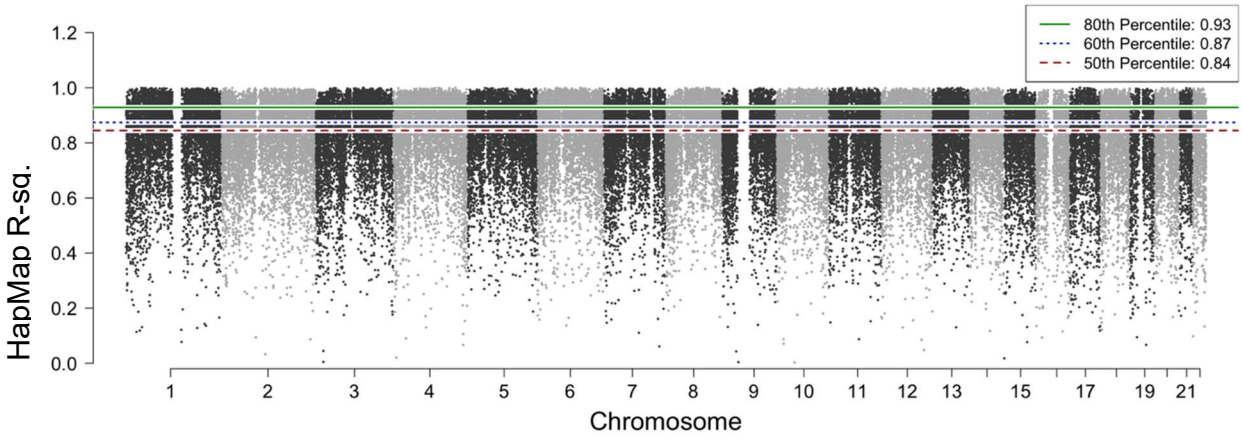


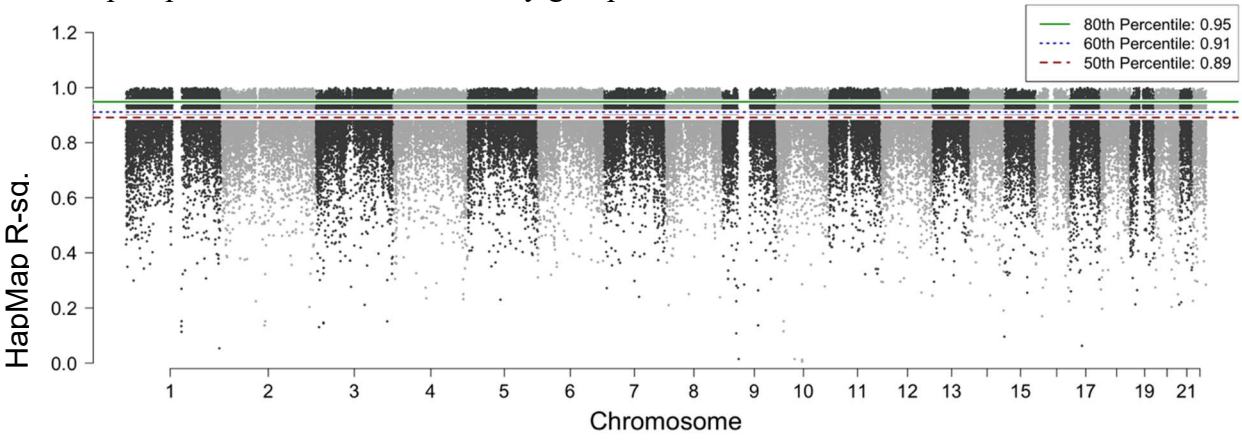
Figure B7: Cross-ancestry MC-ANOVA predicted R-squared (R-sq.) by chromosome and position based on SNPs from the HapMap variants. Each dot represents the estimated $R^2_{1\to2}$ [6] for a chromosome segment from the HapMap SNP set (with an average length of 23 Kbp) by the ancestry group of the testing data: AF=African (**a**), CR=Caribbean (**b**), EA=East Asian (**c**), and SA=South Asian (**d**). Ancestry group 1 is European (EUR). The green line is the 80th percentile value of $R^2_{1\to2}$, the blue short dashed line is the 60th percentile, and the red long dashed line is the 50th percentile. See Figure B5 for results based on SNPs from the UK Biobank arrays.

Figure B7 (cont'd)

c: HapMap SNPs East Asian ancestry group.



d: HapMap SNPs South Asian ancestry group.



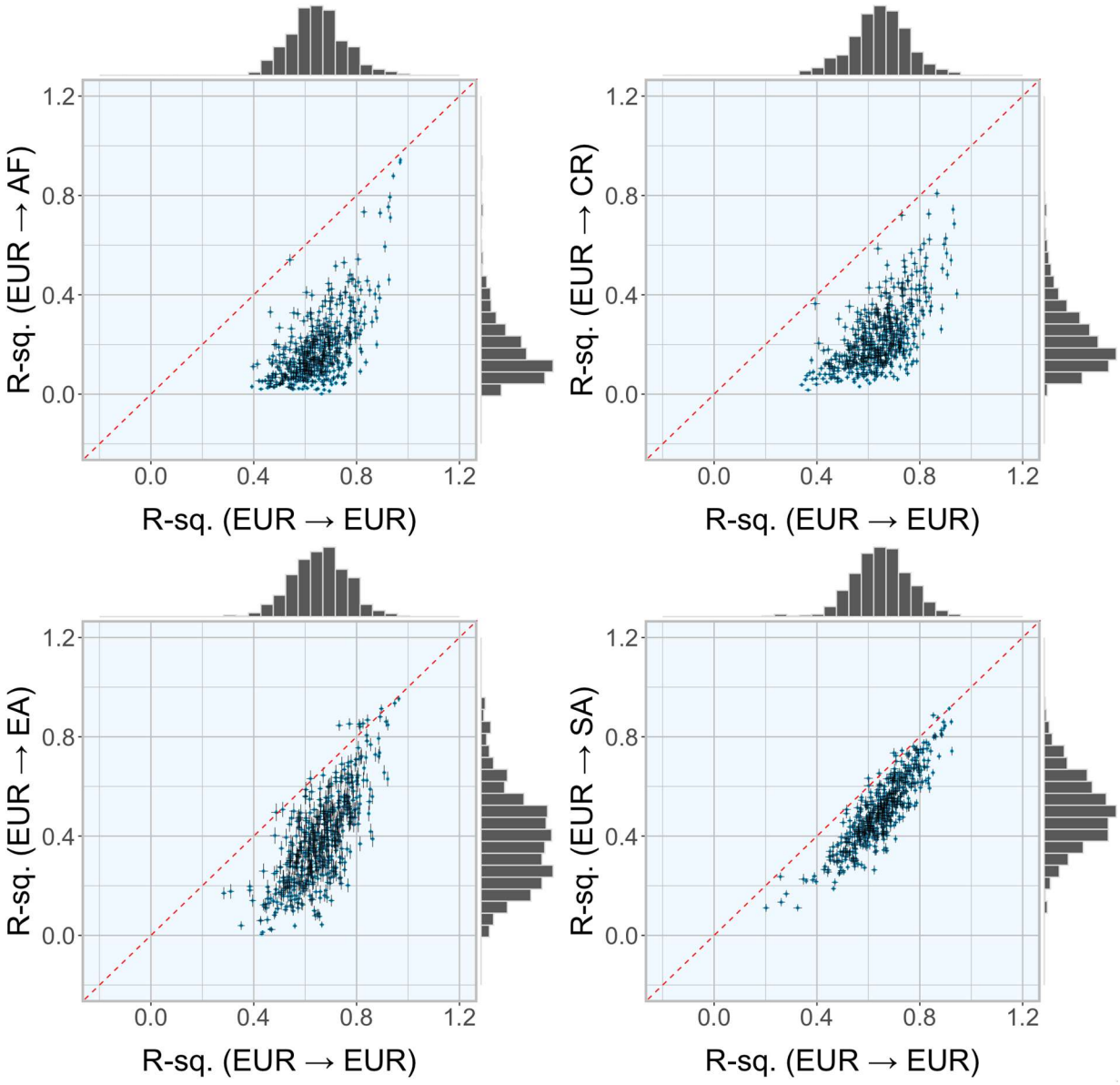
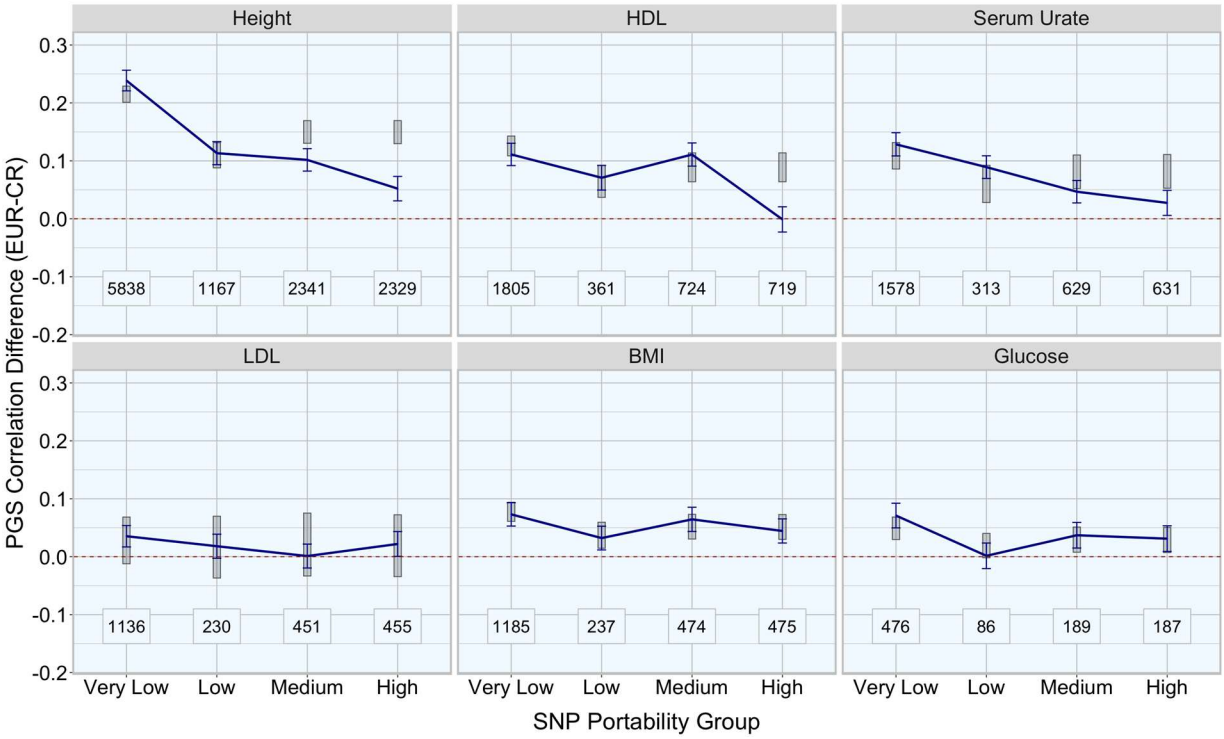


Figure B8: Distribution of the cross-ancestry R-squared (R-sq.) versus the within-European (EUR) R-squared for the African (AF), Caribbean (CR), East Asian (EA), and South Asian (SA) ancestry groups. Each point represents a small chromosome segment (45 Kbp) from the UK Biobank arrays. Each panel displays a different non-European (EUR) ancestry group. Five hundred segments were randomly sampled and plotted for each ancestry group. The standard error bars for each R-squared (R-sq.) point estimate are shown with the cross bars. The sample sizes used to derive the standard errors are in Table B2.

a: Caribbean ancestry group.



b: East Asian ancestry group.

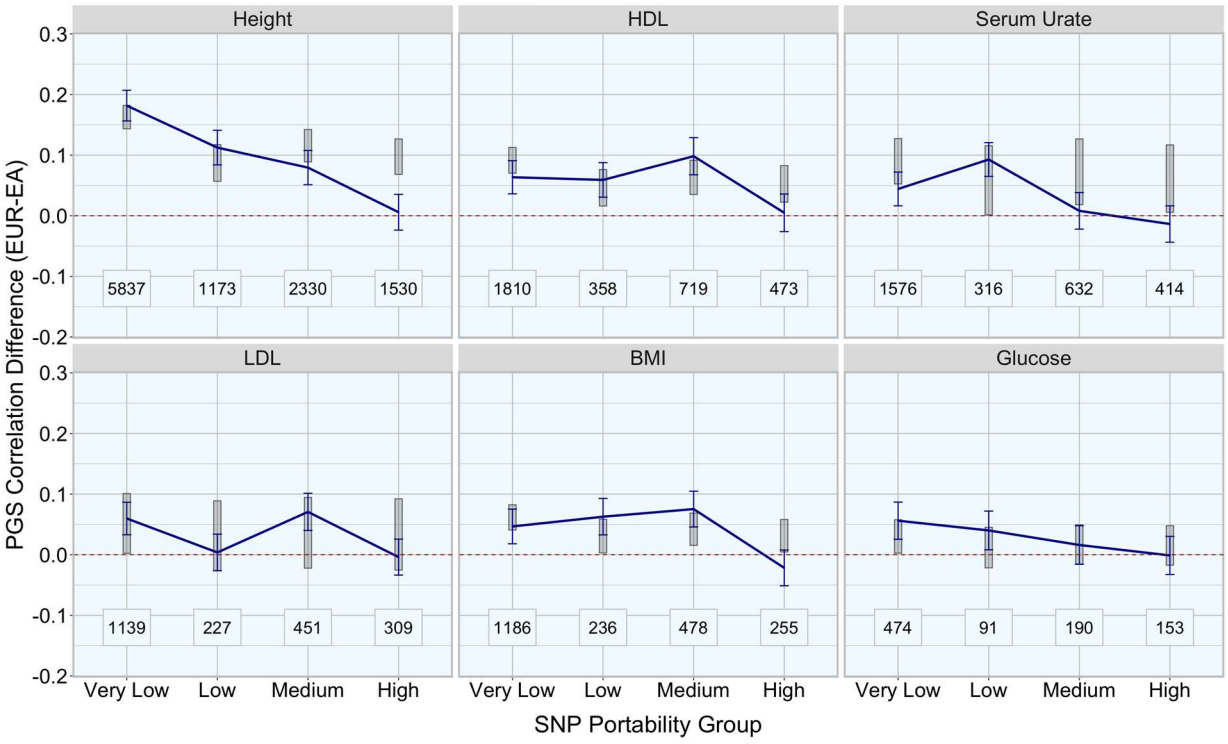


Figure B9: Difference between within- and cross-ancestry polygenic score prediction

Figure B9 (cont'd)

correlation of European (EUR) derived polygenic scores by ancestry and SNP portability group based on SNPs from the UK Biobank arrays. The vertical axis represents the difference between the within- and cross-ancestry polygenic score prediction correlation for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability ($R^2_{1\to2}$ groupings, Table 3) by trait (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose) and ancestry group (CR=Caribbean [**a**], EA=East Asian [**b**], and SA=South Asian [**c**]). A positive difference in PGS prediction correlation indicates that the PGS of the SNP set had a higher prediction correlation in EURs (within-ancestry prediction) than in individuals of CR, EA, or SA (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) 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 Table B2.

Figure B9 (cont'd)

c: South Asian ancestry group.

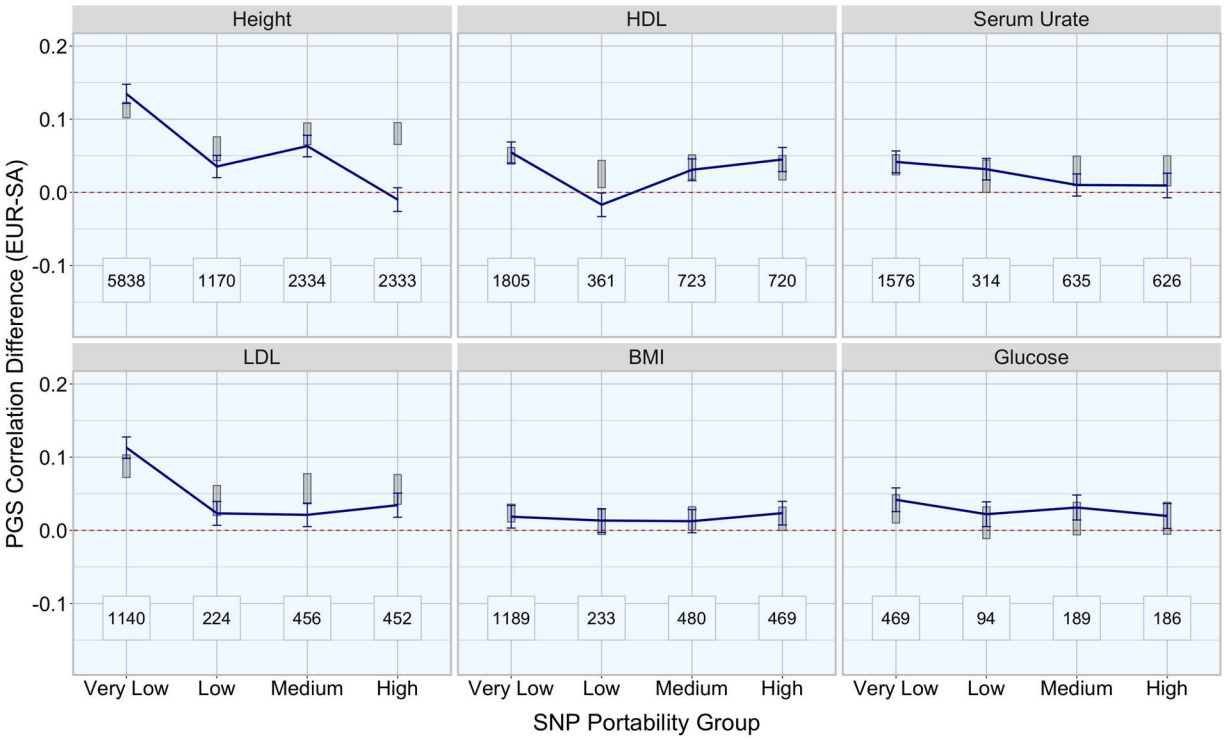
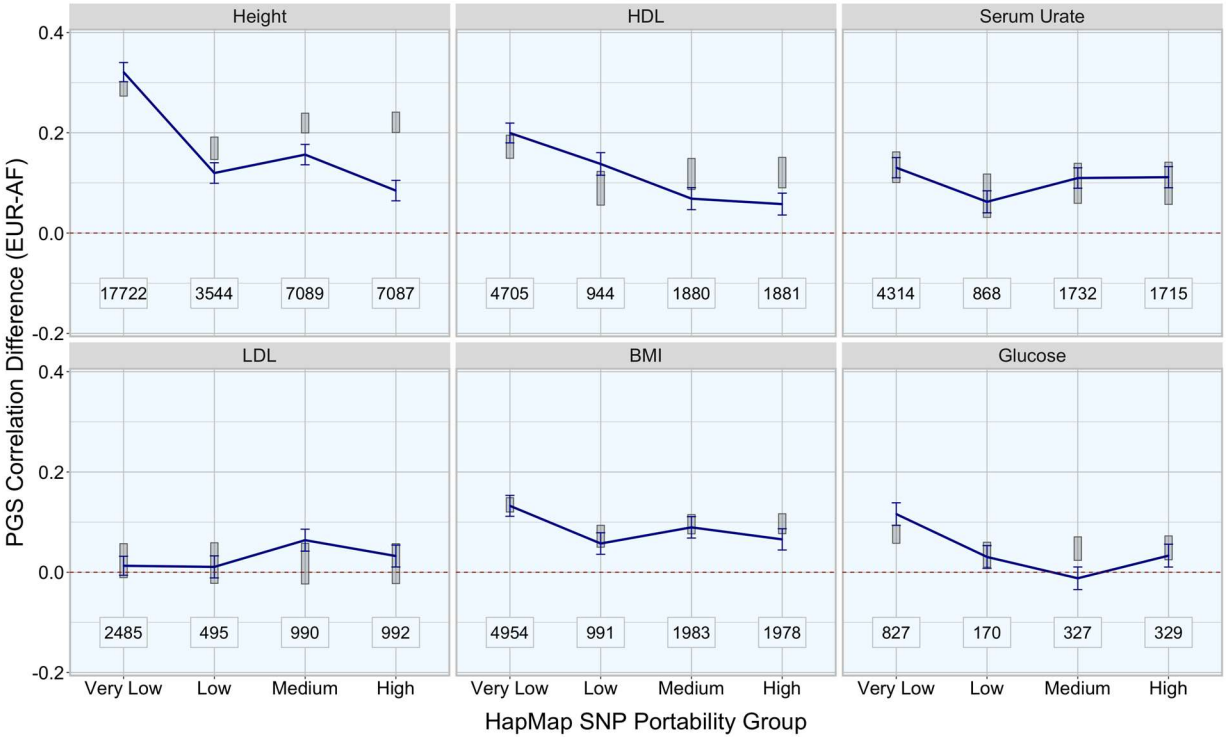


Figure B10: Predicted and empirical relative accuracies (RA) by SNP portability group by trait and ancestry group based on SNPs from the UK Biobank arrays. MC-ANOVA predicted relative accuracy (RA) and empirical RA of European (EUR)-derived polygenic scores when used to predict phenotypes of individuals of non-EUR ancestry (AF, CR, EA, and SA denote African, Caribbean, East Asian, and South Asian ancestry) by SNP portability group for six traits (height, high-density lipoprotein [HDL], serum urate [SU], low-density lipoprotein [LDL], body mass index [BMI], and glucose). Each panel displays a different phenotype-ancestry group combination. The loss of accuracy (LOA, %) attributable to genome differentiation is shown on top of each bar set. A standard error bar is shown for each mean RA estimate (derivation details are in the Methods). The sample sizes used to derive the standard errors are in Table B2.

a: African ancestry group (HapMap SNPs).



b: Caribbean ancestry group (HapMap SNPs).

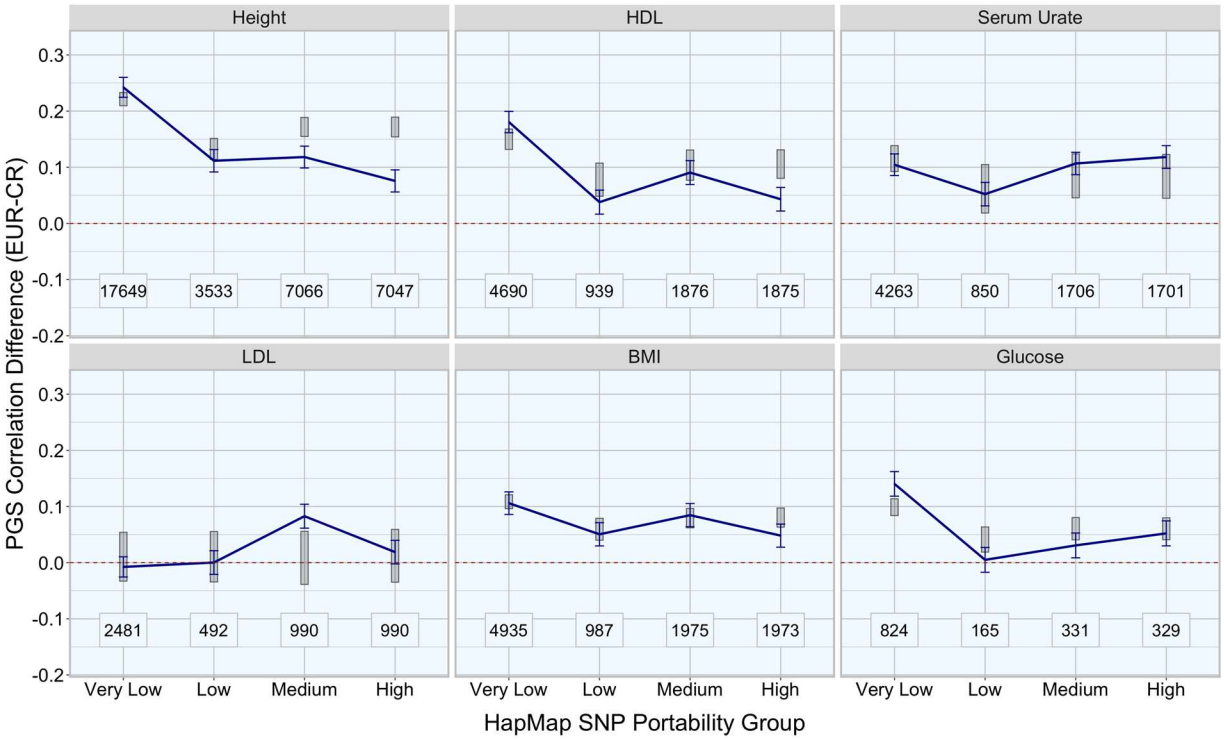


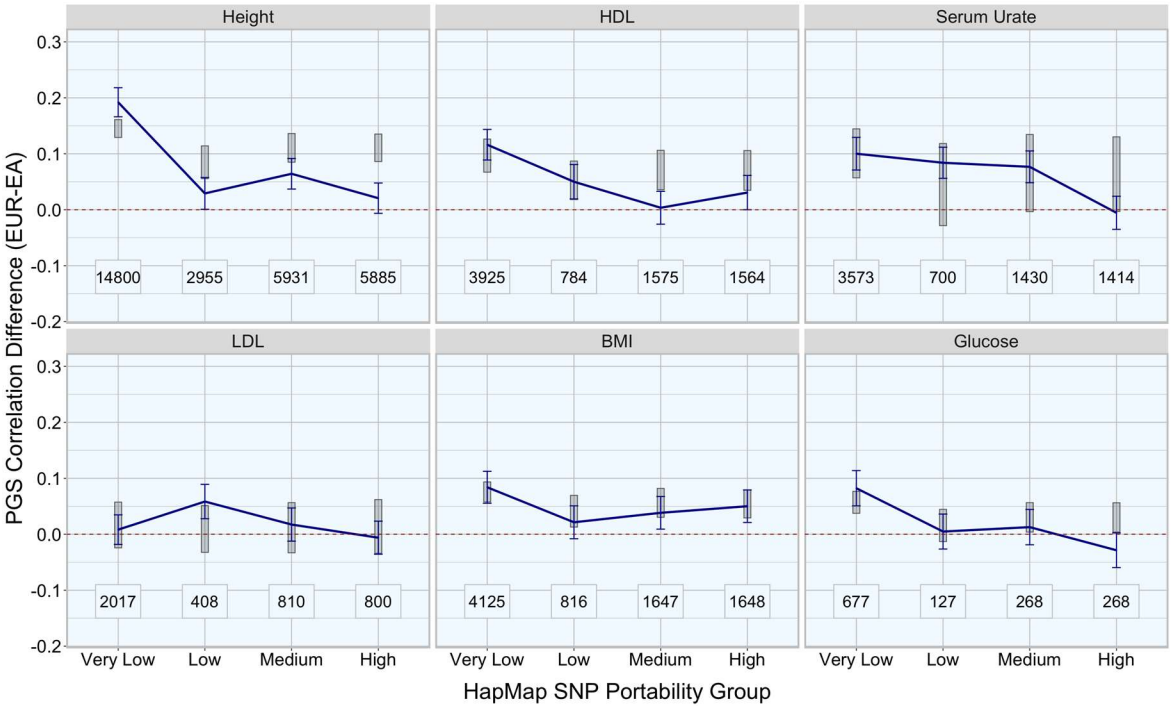
Figure B11: Validation plots for the HapMap SNP set. The difference between polygenic

Figure B11 (cont'd)

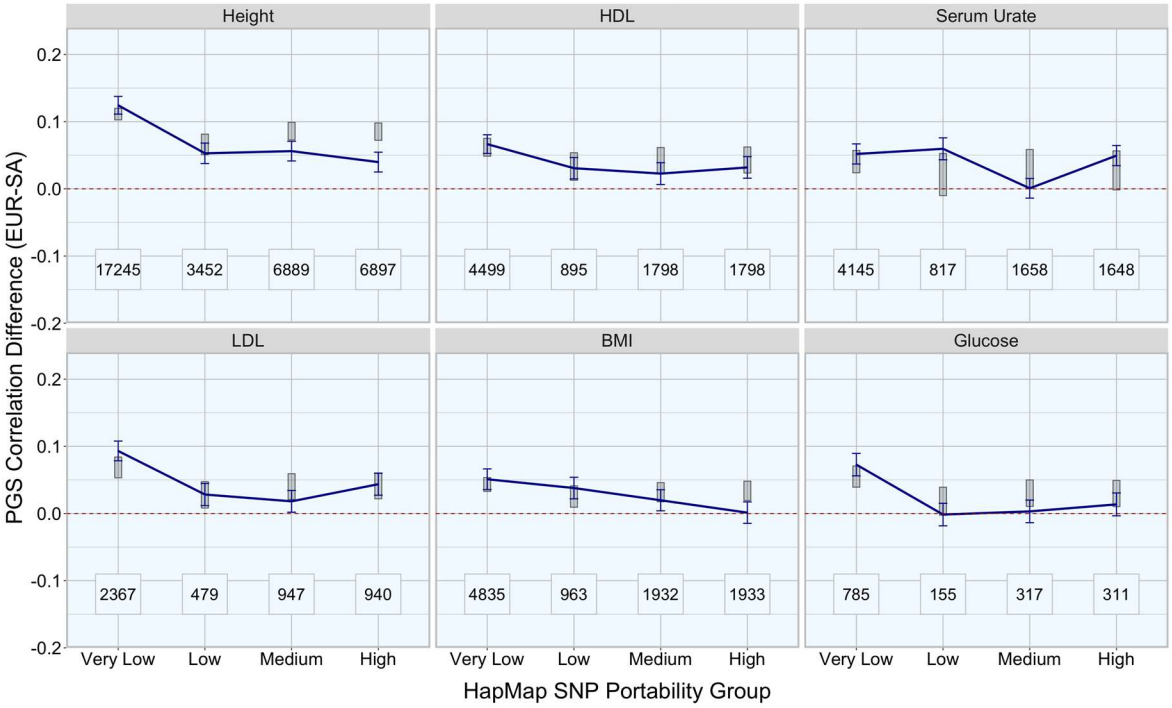
score prediction correlation by HapMap SNP portability group. The vertical axis represents the difference between the within- and cross-ancestry polygenic score prediction correlations of European (EUR) derived polygenic scores for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability ($R^2_{1\rightarrow 2}$ groupings) by trait (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose) and ancestry group (AF=African [**a**], CR=Caribbean [**b**], EA=East Asian [**c**], and SA=South Asian [**d**]). A positive difference in PGS prediction correlation indicates that the PGS of the SNP set had a higher prediction correlation in EURs (within-ancestry prediction) than in individuals of 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) 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 Table B2. See Figure 7 and Figure B9 for results based on SNPs from the UK Biobank array

Figure B11 (cont'd)

c: East Asian ancestry group (HapMap SNPs).



d: South Asian ancestry group (HapMap SNPs).



a: Fst[2] compared to MC-ANOVA.

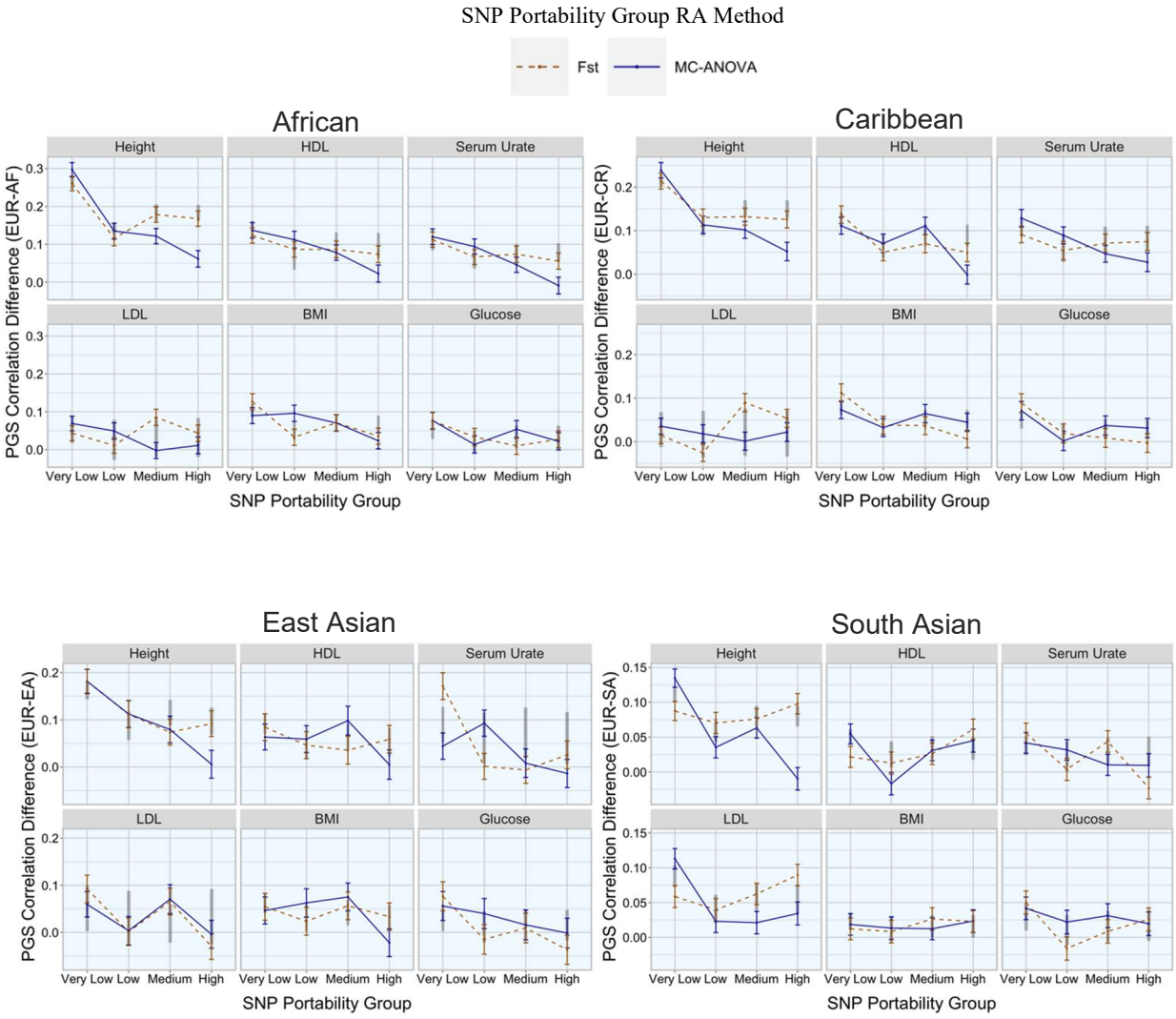
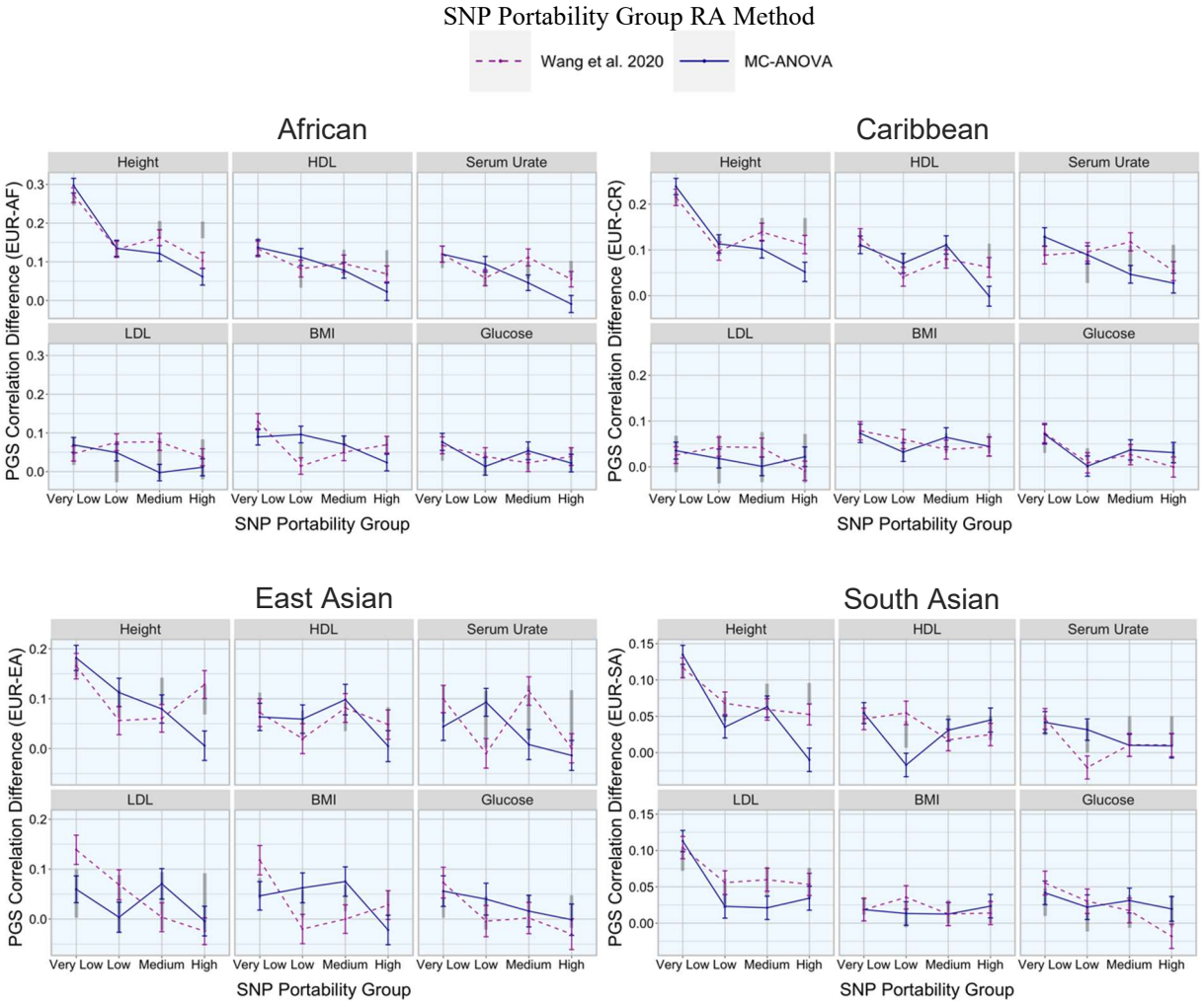


Figure B12: Difference between within- and cross-ancestry polygenic score prediction correlation of European (EUR) derived polygenic scores by ancestry and SNP portability groups based on different methods (using SNPs from the UK Biobank arrays). The vertical axis represents the difference between the within- and cross-ancestry polygenic score prediction correlation for SNP groups with Very Low, Low, Medium, and High predicted portability determined from different methods (Fst[2] vs. MC-ANOVA [**a**] and Wang et al. RA[1] vs. MC-ANOVA [**b**]) by trait (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose) and ancestry group (AF=African, CR=Caribbean, EA=East Asian, and SA=South Asian). A positive difference in PGS prediction

Figure B12 (cont'd)

correlation indicates that the PGS of the SNP set had a higher prediction correlation in EURs (within-ancestry prediction) than in individuals of AF, CR, EA, or SA (cross-ancestry prediction) ancestry. Within (**a**) and (**b**), the panels are first grouped by ancestry group and then by trait. 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) 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 Table B2.

b: Wang et al.'s[1] RA compared to MC-ANOVA.



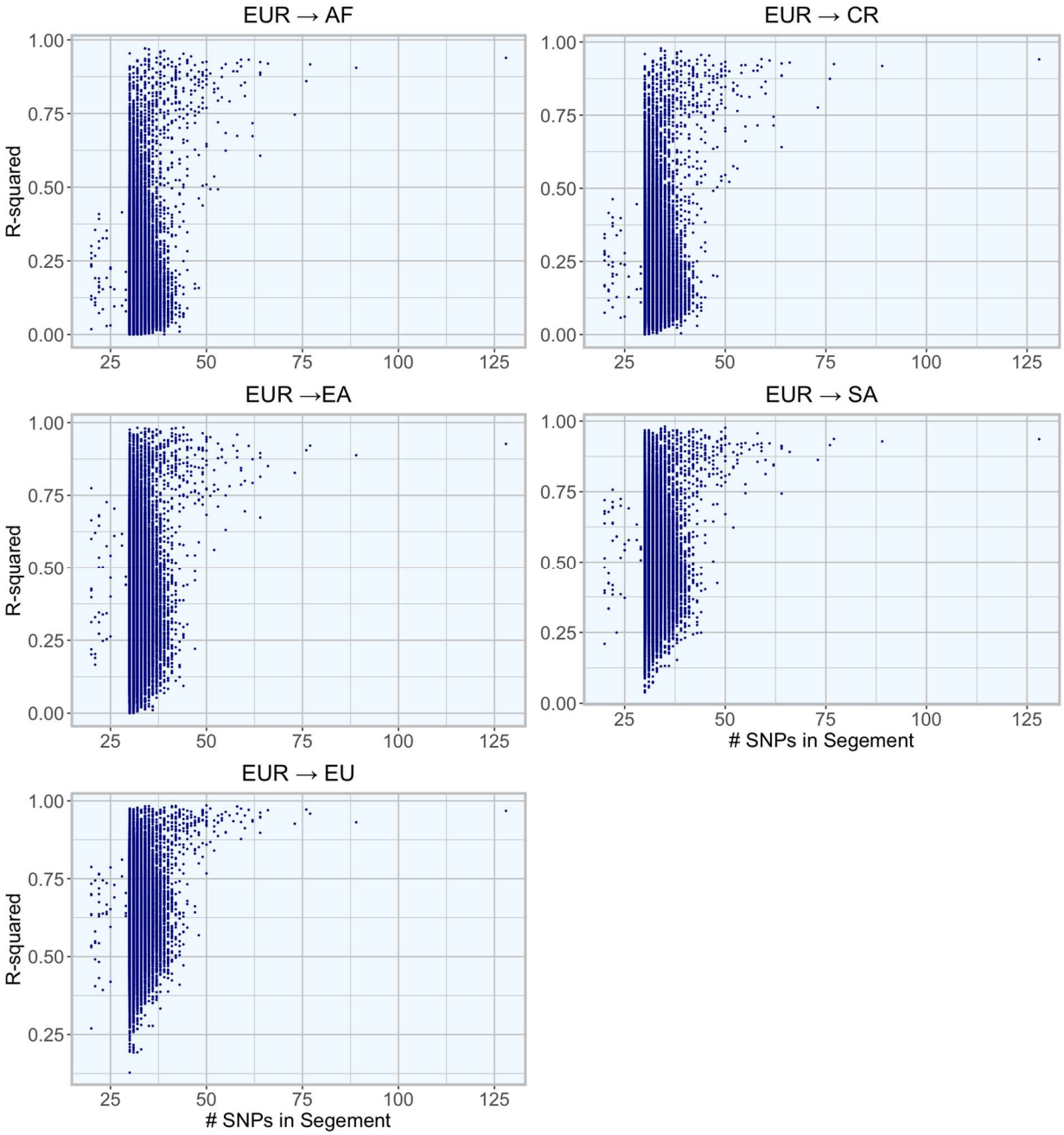


Figure B13: Predicted cross-ancestry R-squared [6] by number of SNPs in the chromosome segment (including the core and the SNPs in the flaking regions) by ancestry of the testing data (AF=African, CR=Caribbean, EA=East Asian, and SA=South Asian). Each panel displays a different ancestry group. The results are based on SNPs from the UK Biobank arrays. For European (EUR) the plot displays the within-ancestry R-squared parameter [5].

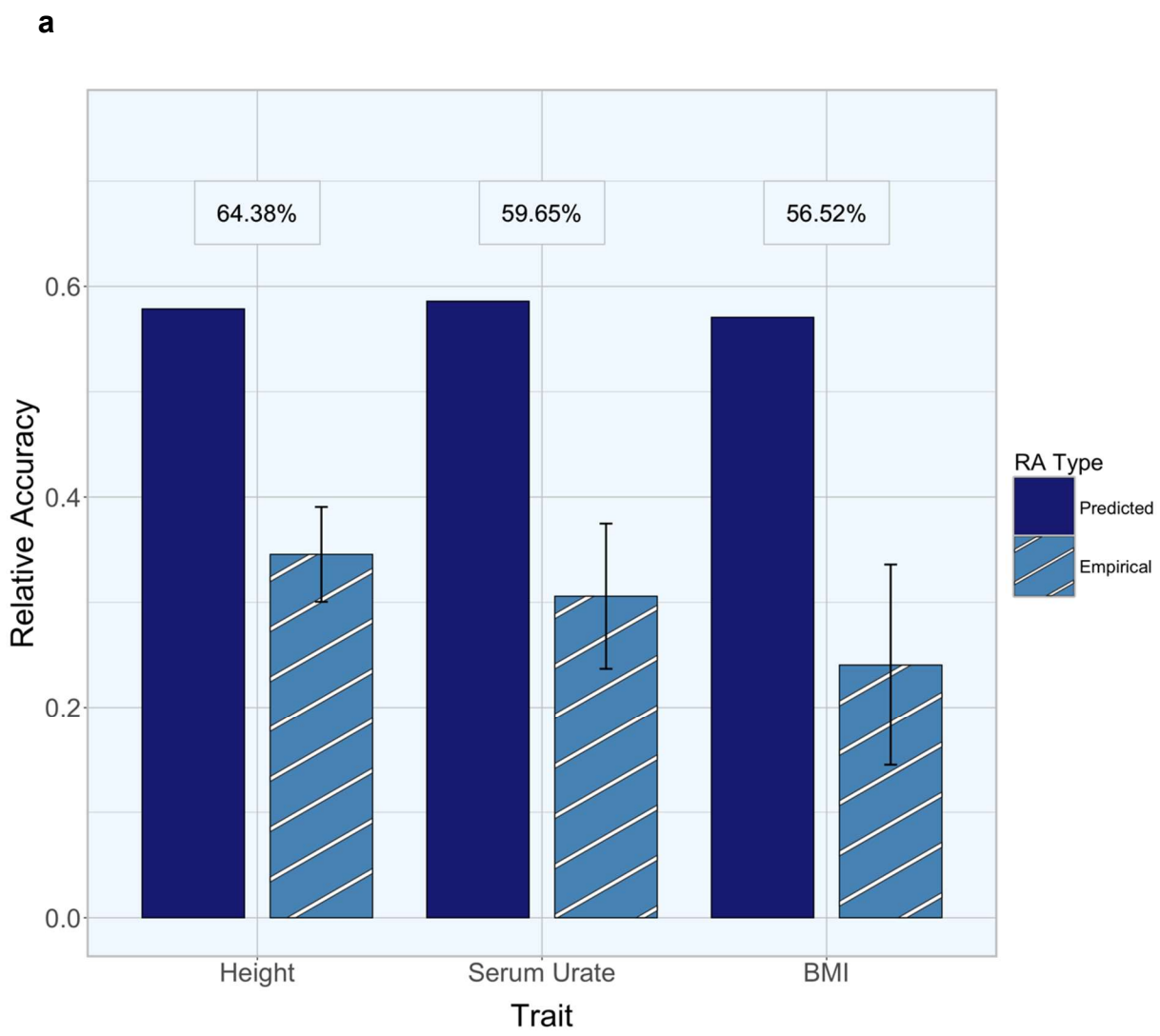
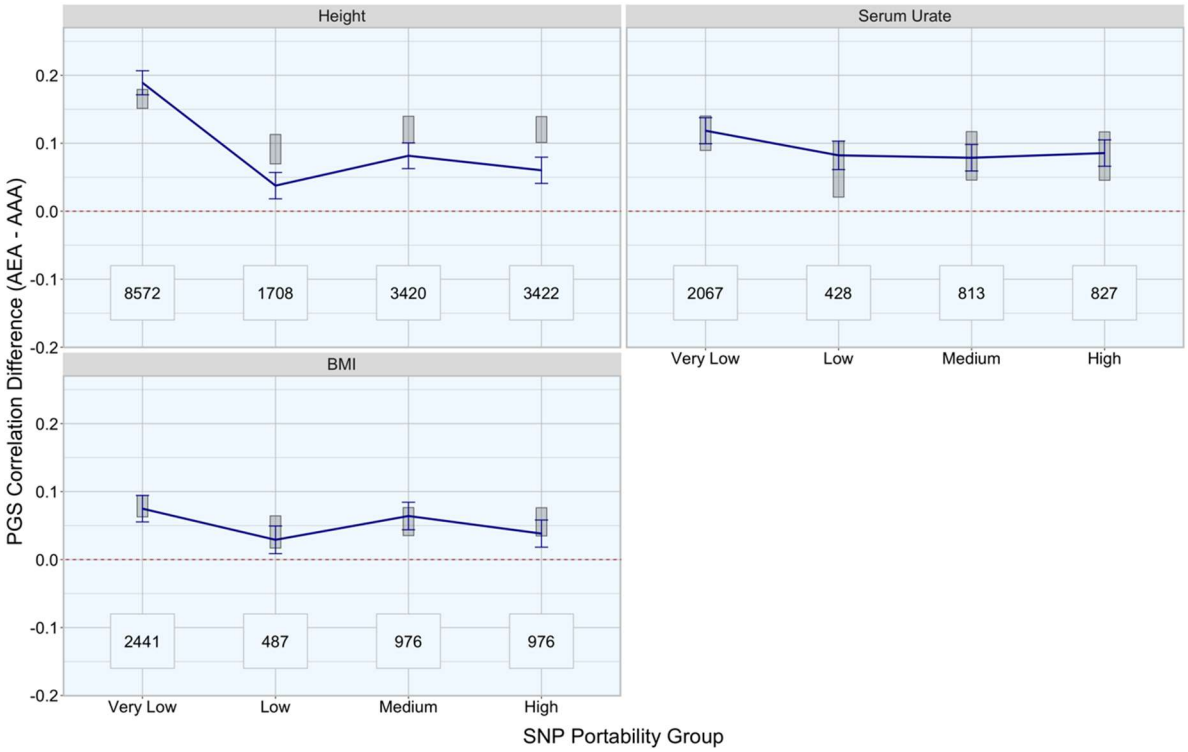


Figure B14: PGS relative accuracies (RA) and correlation differences in the ARIC study data set. (**a**) MC-ANOVA predicted relative accuracy (RA) versus empirical RA of UK Biobank European (EUR)-derived polygenic scores (PGS) when used to predict phenotypes of individuals from the ARIC study. The within-ancestry testing group is the ARIC study European Americans (AEA) and the cross-ancestry group is the ARIC study African Americans (AAA). A standard error (SE) bar is shown for each mean empirical RA estimate. The loss of accuracy (LOA, %) attributable to genome differentiation is shown on top of each bar set. (**b**) The vertical axis represents the difference between the within- and cross-ancestry polygenic score prediction correlation for SNP groups with Very Low, Low, Medium, and High MC-ANOVA predicted portability ($R^2_{1\rightarrow 2}$ groupings) by trait (height, serum urate, and body mass index [BMI]). A

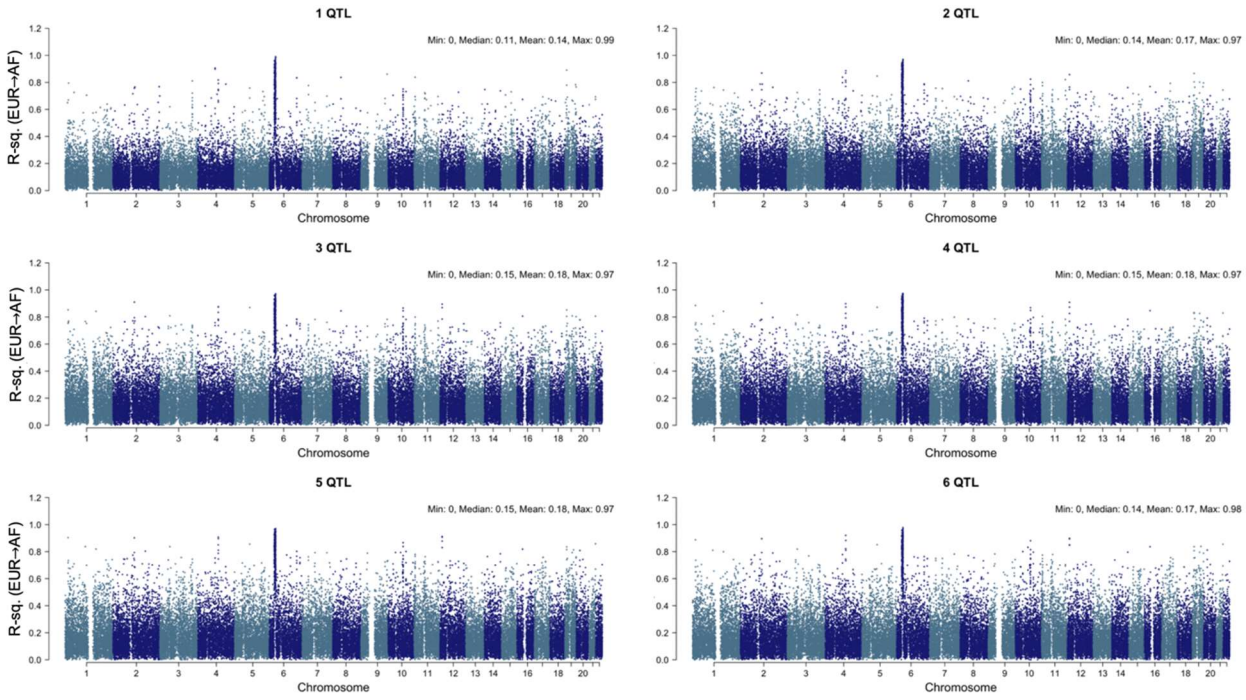
Figure B14 (cont'd)

positive difference in PGS prediction correlation indicates that the PGS of the SNP set had a higher prediction correlation in AEA (within-ancestry prediction) relative to AAA (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) for the correlation difference, where SNPs were assigned to portability groups completely at random, maintaining the number of SNPs in each subgroup. For both **a** and **b**, the sample sizes for the SE bars and the simulated null are n=9,628 AEA and n=3,130 AAA for height, n=9,627 AEA and n=3,046 AAA for serum urate, and n=9,625 AEA and n=3,127 AAA for BMI.

b



a: Varying the number of QTL per segment.



b: Varying the number of flanking SNPs to the sides of each segment.

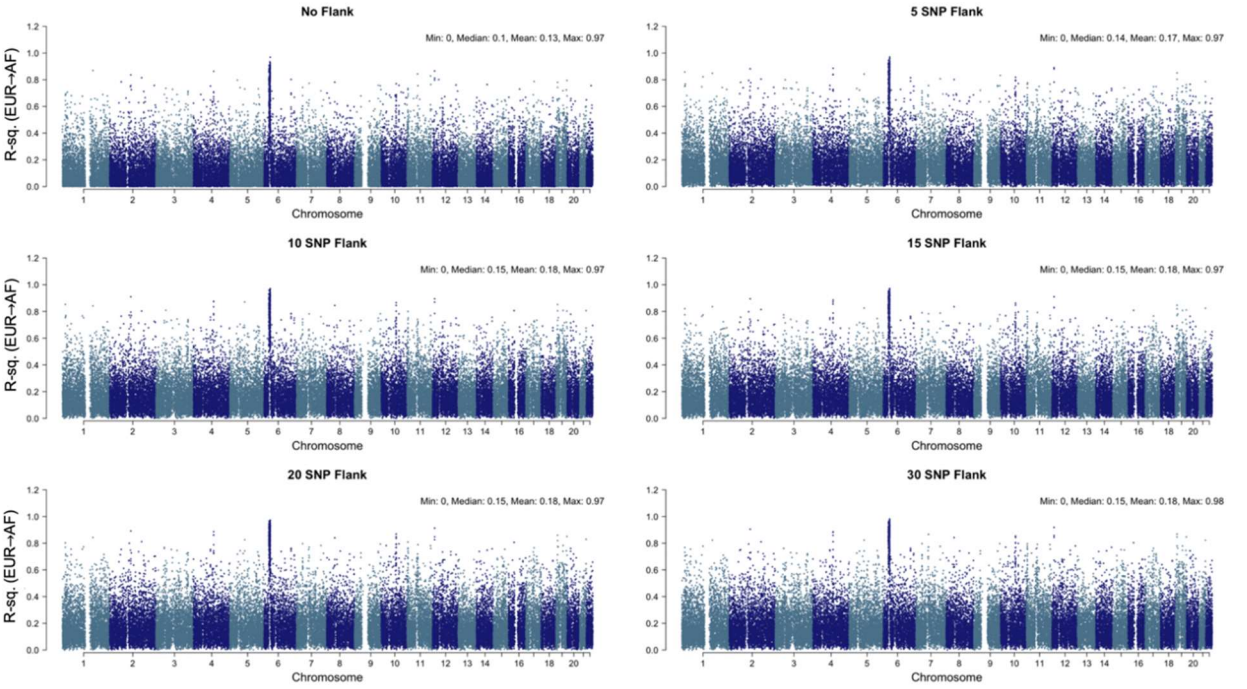


Figure B15: (a, b) Cross-ancestry R-squared by chromosome and position by number for

Figure B15 (cont'd)

varying numbers of causal variants in the segment and number of SNPs in the flanking regions (all results based on SNPs from the UK Biobank arrays). Each dot represents the estimated $R^2_{1\to2}$ [6] for a chromosome segment for the AF=African ancestry group by (**a**) the number of sampled QTL and (**b**) the number of SNPs included in the flanking regions to each side of the chromosome segment. (**c, d**) Cross-ancestry R-squared (R-sq.) from the baseline model (three QTL and ten flanking SNPs) subtracted from the model varying either the number of causal variants in the segment or the number of SNPs in the flank (based on SNPs from the UK Biobank arrays). Each histogram shows the distribution of the difference in $R^2_{1\to2}$ [6] between the sensitivity model minus the baseline model for the AF=African ancestry group by (**c**) the number of sampled QTL and (**d**) the number of SNPs included in the flanking regions to each side of the chromosome segment. There is a vertical red line at R-squared equals zero.

Figure B15 (cont'd)

c: Varying the number of QTL per segment compared to the baseline method (3 QTL).

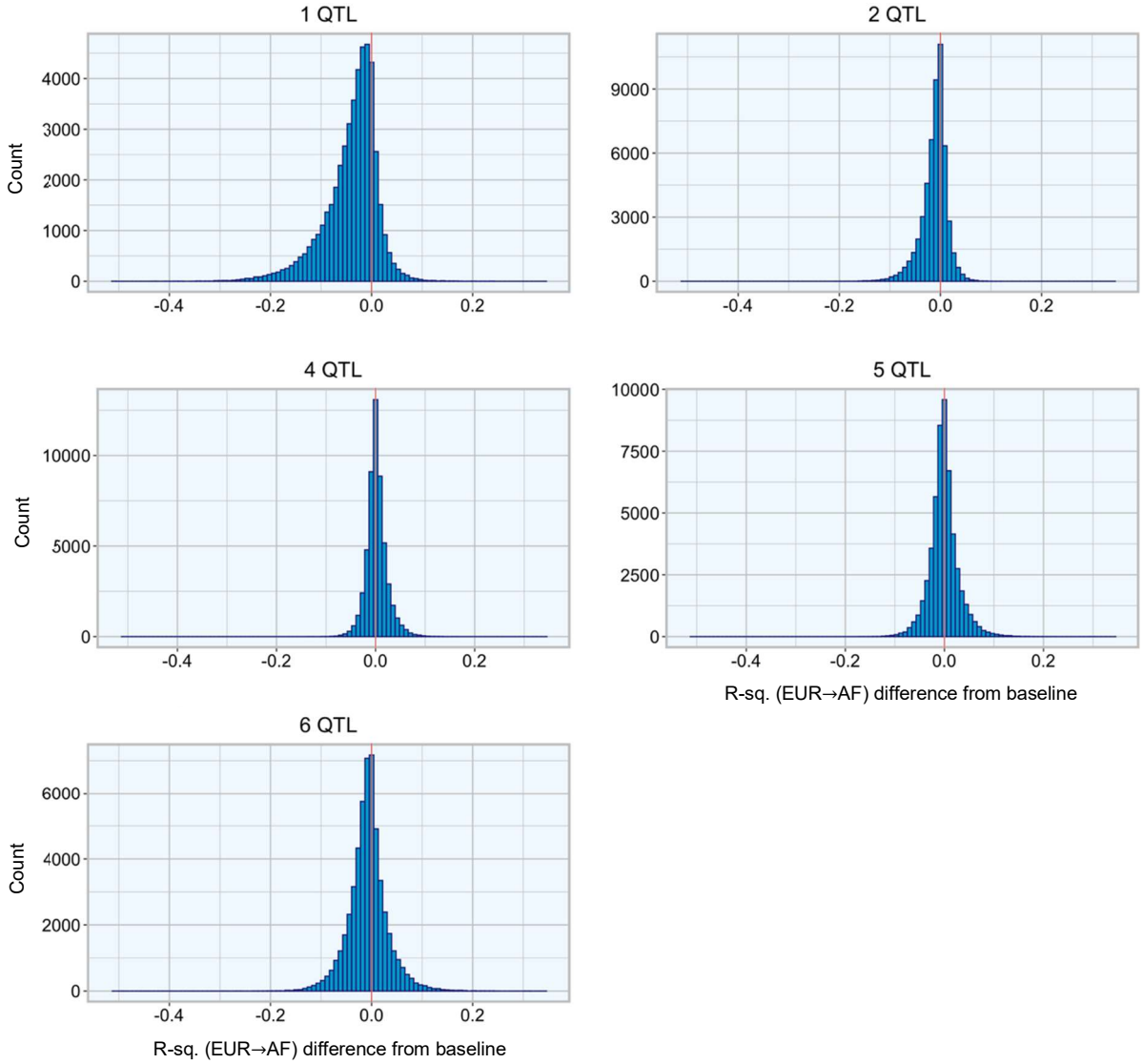
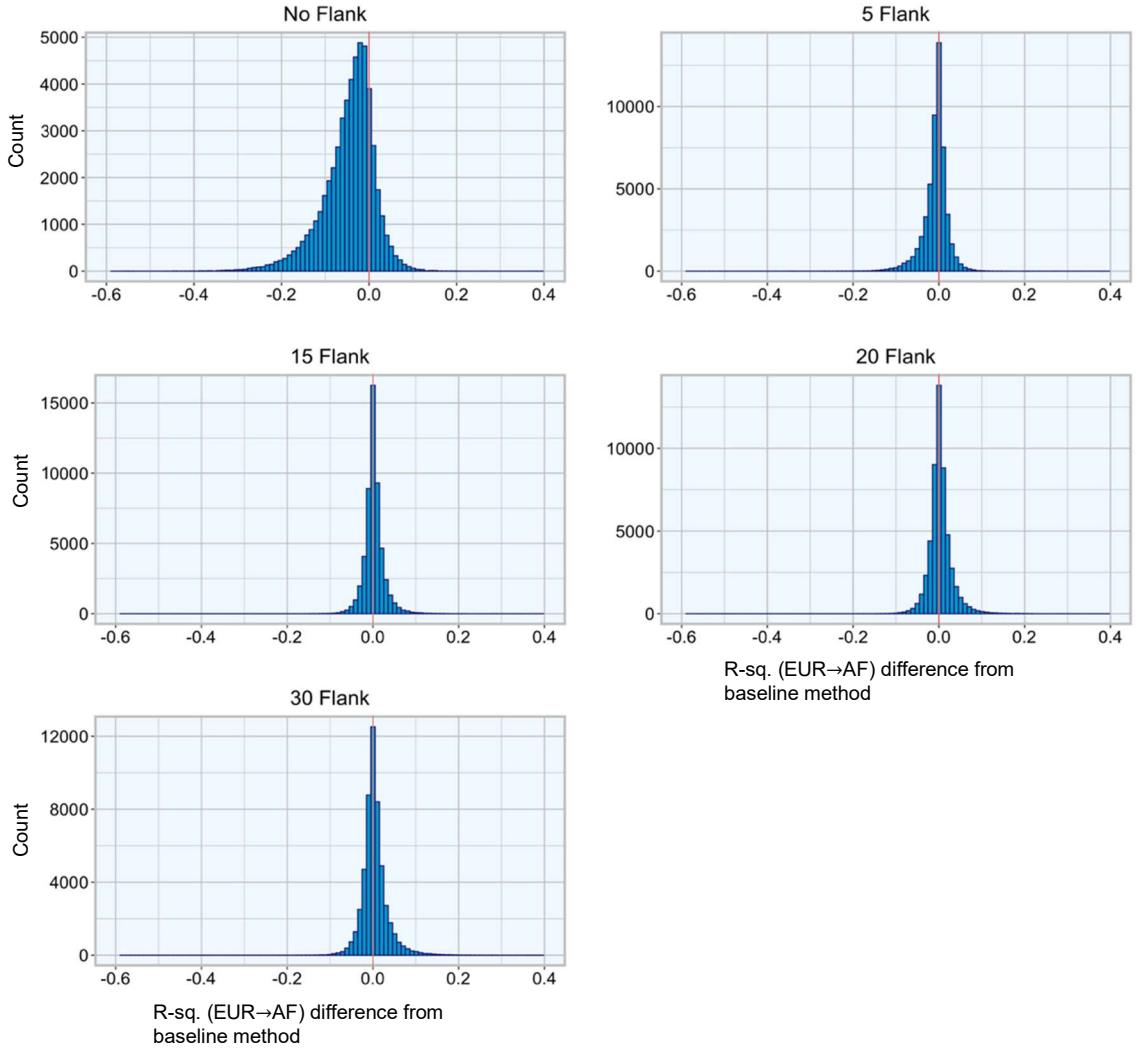


Figure B15 (cont'd)

d: Varying the flanking SNPs to the sides of each segment compared to the baseline method (3 QTL).



a: Cross-ancestry R-squared (EUR → AF).

b: Within-ancestry R-squared (EUR → EUR).

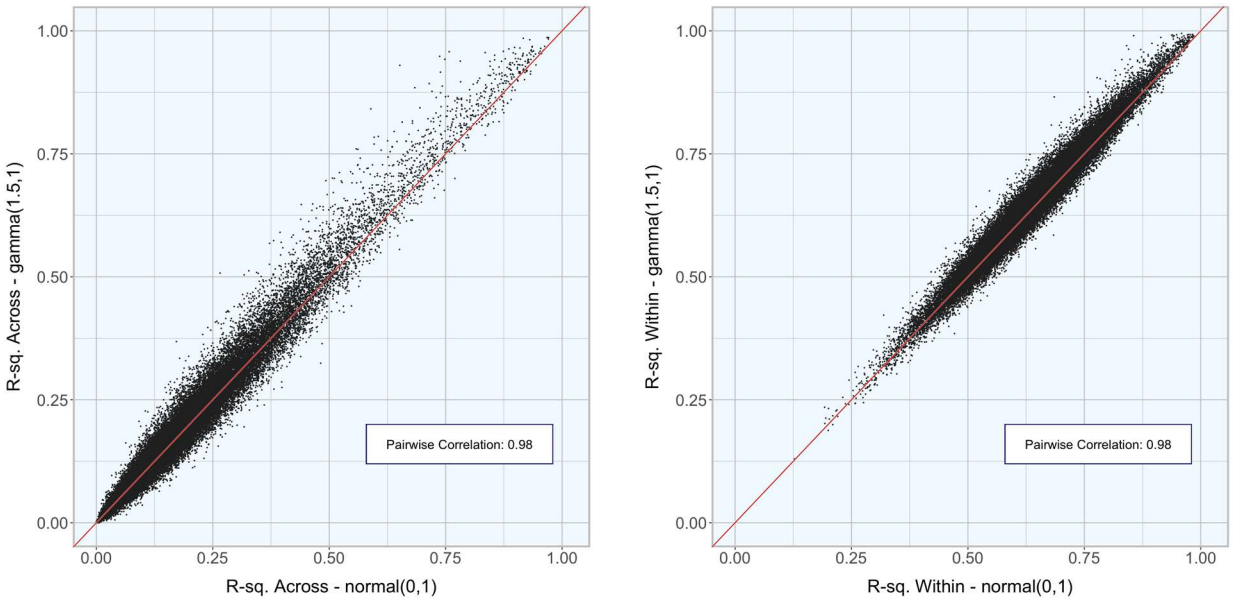


Figure B16: Cross- and within-ancestry R-squared for different causal variant effect distributions based on SNPs from the UK Biobank arrays. The MC-ANOVA cross-ancestry R-squared (R-sq.) estimates for the African (AF) ancestry group (**a**) and the within-ancestry (European [EUR]) R-squared estimates (**b**) when drawing causal variant effects from a normal distribution (shown in the main results) compared to a gamma distribution with a shape parameter of 1.5 and rate parameter of one. The pairwise Pearson correlation is noted for each subplot.

Supplementary Tables

Table B1: Average R-squared and relative accuracy (RA) by testing set using HapMap SNPs.

Ancestry Group	Sample Size	R-squared ($R^2_{1\to2}$)*	Relative Accuracy ($R^2_{1\to2}/R^2_{1\to1}$)	Standard Error of the RA***	Variance in RA Across Segments***
European (EUR)	230,000	0.926**	1.000		
African (AF)	3,083	0.596	0.638	0.021	0.030
Caribbean (CR)	3,343	0.629	0.674	0.020	0.026
East Asian (EA)	1,329	0.814	0.875	0.022	0.012
South Asian (SA)	7,919	0.868	0.935	0.010	0.003

* Subscript 1 always indicates an EUR training or testing set; 2 indicates non-EUR testing; ** $R^2_{1\to1}$; *** Median

Table B2: Descriptive statistics by ancestry group in the UK Biobank data set. Continuous variables are reported as the mean ± standard deviation and are followed by the number of samples missing in parentheses.

Variable	Units	European (EUR) Training	EUR Testing	South Asian (SA)	East Asian (EA)	Caribbean (CR)	African (AF)
Total Sample Size		230,000	6,698	7,919	1,329	3,343	3,083
Female	%	52.8	52.4	45.6	62.8	62.7	48.5
Age	years	56.8 ± 8.0	57.0 ± 7.9	53.2 ± 8.5	52.4 ± 7.6	52.8 ± 8.1	50.8 ± 7.9
Height	cm	169.1 ± 9.2	169.2 ± 9.3	164.4 ± 8.9 (305)	162.0 ± 7.7 (25)	167.3 ± 8.6 (51)	167.7 ± 8.6 (66)
HDL	mmol/L	1.5 ± 0.4	1.5 ± 0.4	1.3 ± 0.3 (1,053)	1.5 ± 0.4 (172)	1.5 ± 0.4 (428)	1.4 ± 0.4 (406)
Serum Urate	umol/L	309.8 ± 80.1	310.5 ± 79.8	318.7 ± 79.8 (410)	311 ± 76.9 (62)	305.5 ± 81.7 (183)	318.7 ± 80.5 (207)
LDL	mmol/L	3.6 ± 0.9	3.6 ± 0.9	3.3 ± 0.9 (419)	3.4 ± 0.8 (61)	3.3 ± 0.8 (187)	3.2 ± 0.8 (209)
BMI	kg/m^2	27.4 ± 4.7	27.4 ± 4.7	27.1 ± 4.4 (169)	24.1 ± 3.4 (8)	29.3 ± 5.5 (50)	29.6 ± 5.1 (51)
Serum Glucose	mmol/L	5.1 ± 1.2	5.1 ± 1.2	5.4 ± 1.9 (1,048)	5.1 ± 1.0 (172)	5.2 ± 1.6 (434)	5.1 ± 1.5 (407)

Table B3: Estimated relative accuracy (RA) of SNP windows grouped by the estimated cross-ancestry R-squared ($R^2_{1\to2}$ for 1=European [EUR] and 2=testing set) for the Caribbean (CR), East Asian (EA), and South Asian (SA) ancestry groups using SNPs from the UK Biobank array.

Testing Group	Portability Group	Quantile Group Cutoff	$R^2_{1\to2}$ Range	Number of SNPs	Average $R^2_{1\to1}$	Average $R^2_{1\to2}$	Average RA $(R^2_{1\to2}/R^2_{1\to1})$
Caribbean (CR)	High	(0.8,1]	(0.31,0.98]	122,158	0.752	0.447	0.592
	Medium	(0.6,0.8]	(0.23,0.31]	122,157	0.675	0.266	0.400
	Low	(0.5,0.6]	(0.20,0.23]	61,073	0.645	0.211	0.334
	Very Low	[0,0.5]	[0,0.20]	305,403	0.596	0.128	0.216
East Asian (EA)	High	(0.8,1]	(0.52,0.98]	112,673	0.771	0.642	0.835
	Medium	(0.6,0.8]	(0.41,0.52]	112,685	0.685	0.460	0.678
	Low	(0.5,0.6]	(0.36,0.41]	56,332	0.652	0.384	0.596
	Very Low	[0,0.5]	[0,0.36]	281,711	0.592	0.240	0.405
South Asian (SA)	High	(0.8,1]	(0.62,0.98]	122,156	0.784	0.712	0.908
	Medium	(0.6,0.8]	(0.53,0.62]	122,151	0.694	0.575	0.831
	Low	(0.5,0.6]	(0.50,0.53]	61,082	0.656	0.516	0.791
	Very Low	[0,0.5]	[0.04,0.50]	305,402	0.573	0.395	0.689

Table B4: The number of SNPs that were selected for each trait (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose) by the threshold used for the p-value in the GWAS (based on a two-sided test of a t-statistic, with the null hypothesis that the SNP effect is zero), and the proportion of variance of the (adjusted) phenotype explained by the European (EUR)-derived PGS in testing data, by ancestry group (African [AF], Caribbean [CR], East Asian [EA], South Asian [SA], ARIC European American [AEA], and ARIC African American [AAA]) using SNPs from the UK Biobank array.

Variable	# SNPs (p<1e-5)	# SNPs for GWAS (p<5e-8)	Proportion of Variance Explained in EUR (%)	Proportion of Variance Explained in AF (%)	Proportion of Variance Explained in CR (%)	Proportion of Variance Explained in EA (%)	Proportion of Variance Explained in SA (%)	Proportion of Variance Explained in ARIC AEA (%)	Proportion of Variance Explained in ARIC AAA (%)
Height	11,675	6,907	27.4	3.9	6.7	9.7	15.0	22.0	7.6
HDL	3,609	1,967	18.0	6.6	7.5	9.3	13.2	--	--
Serum Urate	3,151	1,751	11.4	5.1	4.6	6.2	8.8	9.0	2.7
LDL	2,272	1,210	10.1	6.0	7.9	5.4	4.1	--	--
BMI	2,371	830	3.8	0.3	0.9	0.9	2.8	3.6	0.9
Glucose	938	338	1.9	0.3	0.4	0.5	0.8	--	--

Table B5: The average cross-ancestry R-squared, $R^2_{1\to2}$ [6], by chromosome (Chr) and ancestry group, and the number of annotated genes for each using SNPs from the UK Biobank array.

African (AF)			Caribbean (CR)			East Asian (EA)			South Asian (SA)		
Chr	*Average* $R^2_{1\to2}$	*# Genes*	*Chr*	*Average* $R^2_{1\to2}$	*# Genes*	*Chr*	*Average* $R^2_{1\to2}$	*# Genes*	*Chr*	*Average* $R^2_{1\to2}$	*# Genes*
6	0.335	1060	6	0.379	1060	6	0.515	1041	6	0.605	1060
11	0.241	1303	11	0.292	1304	19	0.464	1351	17	0.571	1125
19	0.236	1406	16	0.286	802	17	0.455	1081	19	0.571	1406
16	0.233	802	17	0.281	1125	11	0.45	1254	22	0.569	463
17	0.232	1125	19	0.28	1406	21	0.448	237	11	0.565	1304
5	0.222	876	5	0.272	876	22	0.442	456	16	0.558	802
15	0.22	607	7	0.271	932	16	0.432	756	21	0.554	244
7	0.219	932	3	0.27	1122	1	0.43	1935	15	0.551	607
3	0.218	1122	22	0.268	463	5	0.428	847	1	0.547	2031
1	0.215	2031	15	0.268	607	15	0.421	591	2	0.546	1278
22	0.215	463	1	0.267	2031	7	0.418	910	7	0.544	932
2	0.212	1278	2	0.262	1278	2	0.417	1230	5	0.543	876
12	0.21	1053	12	0.26	1058	3	0.415	1068	20	0.54	564
10	0.21	776	20	0.259	564	10	0.415	762	12	0.537	1058
21	0.208	244	10	0.258	776	4	0.411	777	14	0.536	650
20	0.208	564	21	0.254	244	14	0.411	622	3	0.536	1122
14	0.207	650	9	0.254	768	20	0.41	545	10	0.536	776
9	0.206	768	4	0.252	802	12	0.405	1006	9	0.533	768
4	0.205	802	14	0.251	650	9	0.403	739	4	0.527	802
18	0.193	312	18	0.237	312	18	0.395	308	18	0.521	312
13	0.188	383	8	0.237	715	13	0.389	373	13	0.519	383
8	0.186	715	13	0.236	383	8	0.387	695	8	0.51	715

Table B6: The top fifteen most portable annotated genes (largest $R^2_{1\to2}$ [6]) for each ancestry group and the associated chromosome (Chr) and number of SNPs in each gene using SNPs from the UK Biobank array. The gene that was common between all ancestry groups is noted with an asterisk.

	African (AF)			Caribbean (CR)			East Asian (EA)			South Asian (SA)		
	Gene	*Chr*	# *SNPs*	*Gene*	*Chr*	# *SNPs*	*Gene*	*Chr*	# *SNPs*	*Gene*	*Chr*	# *SNPs*
1	*HIST1H2AD*	6	1	*TCF19*	6	13	*LOC100129195*	6	4	*ZBTB22*	6	4
2	*HLA-DQB1*	6	84	*HLA-DQB1*	6	84	*ZSCAN16*	6	5	*HIST1H1T*	6	7
3	*TCF19*	6	13	*HLA-F-AS1**	6	48	*HLA-F-AS1**	6	48	*OR51M1*	11	6
4	*CCHCR1*	6	58	*CCHCR1*	6	58	*ZFP57*	6	60	*LOC100129195*	6	4
5	*HLA-DRB1*	6	83	*HLA-DRB1*	6	83	*ZNFX1*	20	7	*ZSCAN16*	6	5
6	*HLA-F-AS1**	6	48	*HCG4*	6	6	*HIST1H1T*	6	7	*B3GALT4*	6	1
7	*LINC00116*	2	7	*LOC554223*	6	17	*HLA-F*	6	33	*PFDN6*	6	1
8	*HLA-DOB*	6	20	*HIST1H2AD*	6	1	*ZBTB22*	6	4	*WDR46*	6	12
9	*SFTA2*	6	7	*OR2B3*	6	7	*B3GALT4*	6	1	*ZFP57*	6	60
10	*BAG6*	6	21	*HLA-DOB*	6	20	*PFDN6*	6	1	*HLA-F*	6	33
11	*HCG4*	6	6	*SFTA2*	6	7	*WDR46*	6	12	*LOC553103*	5	2
12	*LOC554223*	6	17	*LINC00116*	2	7	*BTNL2*	6	36	*ADH1A*	4	4
13	*OR2B3*	6	7	*BTNL2*	6	36	*OR2B3*	6	7	*HLA-DPB2*	6	69
14	*BTNL2*	6	36	*BAG6*	6	21	*OR51M1*	11	6	*HLA-F-AS1**	6	48
15	*HLA-DQA2*	6	43	*BRD2*	6	19	*SFTA2*	6	7	*HIST1H2BG*	6	25

Table B7: From the top fifteen most portable genes from chromosome six from any ancestry group (African [AF], Caribbean [CR], East Asian [EA], and South Asian [SA]) using SNPs from the UK Biobank arrays, the 26 unique genes are grouped by base pair (BP) position (within 50 Kbp) only or base pair position as well as functional class. The three groups based on proximity as well as class were the *HIST* genes, the *HLA-F*/*V* genes, and the *HLA-D* genes.

Genes	BP position	Ancestry groups
H1-6 (*HIST1H1T*), *H2BC8* (*HIST1H2BG*), *H2AC7* (*HIST1H2AD*)	26106237-26216656	AF, CR, EA, SA
ZSCAN16-AS1 (*LOC100129195*), *ZSCAN16*	28092306-28103691	EA, SA
OR2B3	29045632-29054923	AF, CR, EA
ZFP57, HLA-F, HLA-F-AS1, HCG4, HLA-V (*LOC554223*)	29640785-29768123	AF, CR, EA, SA
SFTA2	30899163-30900150	AF, CR, EA
CCHCR1, *TCF19*	31108829-31130078	AF, CR
BAG6	31606813-31619576	AF, CR
BTNL2	32361762-32374640	AF, CR, EA
HLA-DRB1, *HLA-DQB1*, *HLA-DQA2*, *HLA-DOB*, *HLA-DPB2*	32542638-33101602	AF, CR, SA
BRD2	32938199-32948804	CR
B3GALT4, WDR46, PFDN6, ZBTB22	33245868-33283766	EA, SA

Table B8: Prediction correlation for each trait (height, high-density lipoprotein [HDL], serum urate, low-density lipoprotein [LDL], body mass index [BMI], and glucose) averaged over 50 replications in an external testing set (n-testing=300) from a cross-ancestry gradient descent algorithm for each ancestry group (AF=African, CR=Caribbean, and SA=South Asian) when using an adaptive learning rate based on relative accuracy compared to a fixed learning rate (LR).

Ancestry Group	Trait	Average Prediction Correlation (R-squared) with Fixed Learning Rate (LR)	Average Prediction Correlation (R-squared) with Adaptive LR	% Change in R-squared (Fixed to Adaptive LR)	% of Testing Sets in Which Using an Adaptive LR Improved Prediction R-squared*
African (AF)	Height	0.207 (0.043)	0.213 (0.045)	5.97	74.0
	HDL	0.248 (0.062)	0.255 (0.065)	5.64	88.0
	SU	0.203 (0.041)	0.206 (0.042)	2.34	56.0
	LDL	0.227 (0.052)	0.229 (0.052)	1.70	72.0
	BMI	0.059 (0.003)	0.059 (0.003)	-0.71	54.2*
	Glucose	0.021 (0.0004)	0.025 (0.001)	43.77	51.0*
Caribbean (CR)	Height	0.246 (0.061)	0.250 (0.062)	3.13	70.0
	HDL	0.245 (0.060)	0.249 (0.062)	3.13	78.0
	SU	0.170 (0.029)	0.178 (0.032)	9.16	79.5*
	LDL	0.262 (0.068)	0.264 (0.069)	1.48	73.8*
	BMI	0.083 (0.007)	0.084 (0.007)	0.95	50.0*
	Glucose	0.059 (0.003)	0.060 (0.004)	3.66	66.7*
South Asian (SA)	Height	0.345 (0.119)	0.346 (0.120)	0.40	64.0
	HDL	0.319 (0.102)	0.319 (0.102)	-0.15	46.9*
	SU	0.266 (0.071)	0.266 (0.071)	0	All replications were equal
	LDL	0.177 (0.031)	0.179 (0.032)	2.13	64.0
	BMI	0.160 (0.026)	0.163 (0.027)	3.02	70.0
	Glucose	0.086 (0.007)	0.088 (0.008)	6.54	62.0

* Percentage excludes training-testing partitions for which the adaptive and fixed R-squared were identical, which happened whenever the optimal number of iterations was zero or very large, in which cases varying learning rates do not affect estimates.

Supplementary Notes

Supplementary Note 1: Portability map availability and use.

The portability maps for SNPs from the UK Biobank arrays as well as SNPs from HapMap 3 variants are accessible in three ways: **1)** Supplementary Data, **2)** via an R package (downloadable as data objects as well as interactively through a Shiny app), and **3)** interactively through a Shiny app hosted on a webpage.

1. Supplementary Data

   The two portability maps can be downloaded directly as Supplementary Data 1 and Supplementary Data 2 (UK Biobank arrays and HapMap variants, respectively).

2. Webpage Shiny app

   A Shiny app graphical interface was created to provide portability map information based on user-provided base pair positions (or genes or RS IDs). It is available at: https://lupia.github.io/Cross-Ancestry-Portability/.

   This version of the Shiny app will run slower than the identical version accessible through the R package described next. However, this web-based version does not require R software or packages.

3. R package MCANOVA: data objects and Shiny app

   The maps are available in an R package, detailed on GitHub here: https://github.com/lupiA/MCANOVA/blob/main/README.md.

   The MCANOVA R package provides the portability maps in two ways. First, they are directly useable as data objects (see Examples, i) once the MCANOVA package is installed. Second, we have created an interactive Shiny app (see Examples, ii) in which users can input base pair positions (or genes or RS IDs) to obtain the relative accuracy estimates and other

portability information for those regions from the maps.

Additionally, the MCANOVA package provides a function implementing the MC-ANOVA method to estimate relative accuracy and functions to obtain the small chromosome segments (see Examples, iii) used in this paper.

Installation

To install the `MCANOVA` package in R, first install the `remotes` package:

```
install.packages("remotes")
library(remotes)
```

Then install the package:

```
install_github("lupiA/MCANOVA")
library(MCANOVA)
```

Examples

After installation is complete:

i) To load the portability maps into an R session as data objects:

```
data(MAP_UKB)
data(MAP_HAPMAP)
```

ii) Launching the Shiny app to interactively access the portability map information:

```
PGS_portability_app()
```

iii) Creating chromosome segments of a minimum base pair length and size (using a small example map):

```
data(geno_map_example)
minSNPs <- 10
minBP <- 10e3
MAP_example <- geno_map_example
MAP_example$segments <- getSegments(MAP_example$base_pair_position,
```

```
        chr = MAP_example$chromosome,
        minBPSize = minBP,
        minSize = minSNPs,
        verbose = TRUE)
```

iv) Running MC-ANOVA

```
  # install.packages("BGData")
  library(BGData)
  # Set seed
  set.seed(12345)
##
  # Generate genotypes (100 subjects and 500 SNPs)
  n <- 100
  p <- 500
  X <- matrix(sample(0:2, n * p, replace = TRUE), ncol = p)
  data(geno_map_example)
  colnames(X) <- geno_map_example$SNPs
  minSNPs <- 10
  minBP <- 10e3
  MAP_example <- geno_map_example
  MAP_example$segments <- getSegments(MAP_example$base_pair_position,
        chr = MAP_example$chromosome,
        minBPSize = minBP,
        minSize = minSNPs,
        verbose = TRUE)
##
  # Assign ancestry IDs (80% to ancestry 1, 20% to ancestry 2)
  n_1 <- round(0.8 * n)
```

```
    n_2 <- round(0.2 * n)
    ancestry <- rep(c("Group_1", "Group_2"), times = c(n_1, n_2))
    rownames(X) <- ancestry
##
    # Initialize portability estimates
    MAP_example$correlation_within <- NA
    MAP_example$correlation_across <- NA
    MAP_example$R_squared_within <- NA
    MAP_example$R_squared_across <- NA
##
    # Set parameters for MC-ANOVA
    lambda <- 1e-8
    nRep <- 300
    nQTL <- 3
##
    # Loop over segments and run MC-ANOVA
    # For whole genome applications, this can be run in parallel with one
job per
    segment in a High-Performance Computing Cluster
##
    for (i in min(MAP_example$segments):max(MAP_example$segments)) {
      core <- which(MAP_example$segments == i)
      flank_size <- 10
      chunk_start <- max(min(core)- flank_size, 1)
      chunk_end <- min(max(core) + flank_size, nrow(MAP_example))
      chunk <- chunk_start:chunk_end
      isCore <- chunk %in% core
```

```
 ##
      X_1 <- X[rownames(X) ==“Group_1”, chunk]
      X_2 <- X[rownames(X) ==“Group_2”, chunk]
##
      # Run MC-ANOVA
      out <- MC_ANOVA(X = X_1, X2 = X_2, core = which(isCore),
                lambda = lambda,
                nQTL = nQTL,
                nRep = nRep)
##
      # Extract portability estimates
      MAP_example$correlation_within[chunk[isCore]] <- out[1, 1]
      MAP_example$correlation_across[chunk[isCore]] <- out[2, 1]
      MAP_example$R_squared_within[chunk[isCore]] <- out[1, 1]^2
      MAP_example$R_squared_across[chunk[isCore]] <- out[2, 1]^2
    }
##
    RA <- MAP_example$R_squared_across/MAP_example$R_squared_within
```

REFERENCES

1. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).

2. Wright, S. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**, 395–420 (1965).

CHAPTER 3: The impact of sample size and the relative proportion of ancestry group on cross-ancestry prediction accuracy

Introduction

Over the past two decades, there has been a large increase in the publication of Genome-Wide Association Studies (GWAS), with initial studies relying on cohorts of a few thousand participants[1]. These early investigations identified numerous loci linked to various human traits and diseases. However, there was a lack of replication between studies which highlighted the need for larger sample sizes to better detect associations between single nucleotide polymorphisms (SNPs) and phenotypes, especially for SNPs with small effects and rare variants. Consequently, numerous GWAS were conducted by consortia, which meta-analyzed summary statistics from multiple cohorts, revealing many novel findings. Despite these advancements, consortia faced limitations such as reliance on summary statistics, inconsistent phenotype definitions, and a focus on single health issues. Thus, with the establishment of biobanks housing hundreds of thousands of individual phenotype-genotype records, sample sizes were increased drastically, overcoming some of the previous consortia limitations.

The advent of Big Data in genomics allowed for a more accurate identification of quantitative trait loci (QTL), and thus significantly enhanced our ability to predict complex traits and disease risk[2,3]. Polygenic scores (PGS) are a common method to estimate the disease or trait genetic predisposition for an individual. Now that genotyping platforms have become sufficiently dense, and with the availability of methods that can be used to impute several millions of variants, the overarching limiting factor of prediction accuracy in PGS is sample size[2,4–7].

Within-population PGS prediction accuracy is affected by three main factors. First, the trait heritability imposes an upper bound on PGS prediction accuracy. Theoretically, we could achieve a PGS prediction R-squared equal to the trait heritability if we knew all the causal variants (and were able to genotype them) and their effects without error[8,9]. However, for complex

traits, knowing all causal variants is nearly impossible. Therefore, PGS rely on using SNPs that are in linkage disequilibrium (LD) with causal variants. Thus, a second factor affecting PGS prediction accuracy is the strength of LD between causal variants and the SNPs used to build a PGS[9-11]. This depends on trait heritability, marker density, and sample size because these three factors affect the power to detect associations between SNPs and phenotypes. Third, PGS use SNP effect estimates; thus, a third factor affecting PGS prediction performance is the accuracy of SNP effect estimates[9-12]. Additionally, for cross-ancestry PGS prediction, the portability of SNP effects between ancestry groups also affects PGS prediction performance[13,14].

There is comparatively poor prediction performance and replication of PGS when applied across ancestries, particularly between more genetically distant ancestry groups, such as European (EU) and African (AF)[15–20]. Genomic differences between ancestry groups in allele frequencies, LD patterns, and LD strength are the primary factors contributing to poor PGS prediction accuracy[13,14,20,21]. Additional factors affecting prediction accuracy are genetic-by-genetic, genetic-by-environment interactions, and effect size differences, however, previous literature suggests that causal variants and their effect sizes are mostly shared between ancestry groups[13,23–26]. Nevertheless, cross-ancestry (EU to non-EU) prediction remains a necessity due to the lack of statistical power from small sample sizes available for within-non-EU prediction and the extreme overrepresentation of EU ancestry groups in genetic data. Recently, studies have found that including even a small number of non-EUs (the target ancestry group) in the training data, or incorporating non-EU summary statistics into the PGS construction, can improve prediction[20,27–32].

We hypothesize that in cross-ancestry prediction (e.g., EU to AF), as training sample size increases, there is more statistical power, typically resulting in more SNPs entering each PGS and

ultimately increasing prediction accuracy. However, we anticipate that the gains in prediction accuracy from increasing the training sample size of EU versus AF is not equivalent, and increasing the sample size of AF will have a bigger gain in prediction accuracy than the same increase in sample size of EU. Additionally, we expect that increasing the training sample sizes will improve SNP effect estimation precision and that the portability of SNP effects between ancestry groups is an additional factor affecting cross-ancestry prediction accuracy.

In this study, using EU ancestry data from the UK Biobank[33] and AF ancestry data from the All of Us platform[34] we evaluate how factors influence cross-ancestry prediction accuracy (EU to AF). We focus on three primary factors: SNP selection and the strength of LD between markers and QTL, SNP effect estimate precision, and SNP effect portability across ancestry groups. Additionally, we estimate the relative contribution to the prediction accuracy of additional cross-ancestry samples compared to within-ancestry samples, hypothesizing that they are not a one-to-one equivalent. Our analysis provides insight into the need for prioritizing non-EU data collection and explores the main bottlenecks in cross-ancestry prediction accuracy.

Materials

UK Biobank cohort

This study selected distantly related individuals of European (EU) ancestry from the UK Biobank who had complete data on height, sex, and age. Participants were between 18 and 75 years old and were not excluded from kinship inference, were included in phasing, and were not identified as an outlier in heterozygosity and missing rates. Following Lupi et al., 2024[14], samples were excluded if they withdrew from the study, "if they had a mismatch of reported and genetic sex, 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."

All of Us cohorts

We selected distantly related African (AF) ancestry individuals from the All of Us cohort[34] with complete data on height, sex, and age (18-75 years old). Relatedness and ancestry were both defined by Controlled Tier data provided by the platform (relatedness-based kinship scores and predicted ancestry[35]). The principal components (PCs) used for this cohort were also supplied by the platform as Controlled Tier data.

Methods

Study overview

One of the main factors limiting PGS prediction accuracy is sample size. To evaluate how different factors, some affected by training sample size, impact cross-ancestry PGS prediction accuracy, we used European (EU) data from the UK Biobank (UKB) and the unprecedentedly large African (AF) ancestry data from the All of Us (AoU) platform to evaluate different scenarios to produce PGS. For each scenario, we constructed PGS for height at varying AF and EU training set sample sizes (used for effect estimation) and evaluated the prediction in the same two testing sets every time: AF (TST_{AF}, $n_{TS\ \ AF} = 9{,}078$) and EU (TST_{EU}, $n_{TST_{EU}} = 10{,}000$). The scenarios differed by: (1) varying both the training sample sizes and the number of SNPs selected (a typical PGS), (2) fixing the number of SNPs but varying the training sample sizes (isolating how sample size affects effect estimation), (3) incorporating SNP effect portability, by comparing PGS consisting of SNPs estimated to be more portable across ancestry group to SNPs estimated to be less portable. The AF training (TRN_{AF}) sample size for effect estimation ranged from $n_{TRN_{AF}} = 0$ to $40{,}000$ and EU training (TRN_{EU}) sample size ranged from $n_{TRN_{EU}} = 0$ to $250{,}000$, with a grid of eight additional sample sizes in between for each ancestry group (more

details can be found in the cohort descriptions in the Methods section).

Scenario 1: A typical PGS (SNP filtering and estimation depend on sample size). To examine the impact of both QTL signal detection and SNP effect estimation, which are dependent on sample size, we varied the training sample sizes used for both SNP filtering and SNP effect estimation. We evaluated each PGS with a standard approach in which SNPs entering into each PGS were selected based on meta-analysis p-values (more details on this can be found in the section 'Genome-wide association study (GWAS)' in Methods) from combining the single marker GWAS (p-value < 1e-4) using the given training sample size of each training ancestry (AF and EU).

Scenario 2: Isolating effect estimation (fixing the SNP set). Next, to distinguish the impact of effect estimation on prediction accuracy from the impact of the number of SNPs selected, we evaluated each PGS with a predetermined SNP set. Thus, in this scenario, the training sample size used for effect estimation varied but the SNP set was fixed (p=5,234 SNPs, filtered from a meta-GWAS from $n_{TRN_{AF}} = 25{,}000$ and $n_{TRN_{EU}} = 100{,}000$). An additional fixed SNP set was filtered from the meta-GWAS of AF ($n_{TRN_{AF}} = 25{,}000$) and EU ($n_{TRN_{EU}} = 25{,}000$).

Scenario 3: SNP portability. Previous literature has suggested that some regions of the genome will be portable across ancestry groups in PGS and others will not[14]. Therefore, to examine SNP portability (more details on this follow in the 'SNP selection' section of Methods), we evaluated the PGS in two different genomic scores. The first PGS included the most portable SNPs across ancestry group, i.e., the SNPs that were in the top 20th percentile of MC-ANOVA predicted cross-ancestry R-squared (most portable SNPs)[14] and the second consisted of the SNPs in the bottom 20th percentile (least portable SNPs).

Design

Out of the 270,859 selected EU individuals, a random sample of $n_{TST_{EU}} = 10{,}000$ individuals was designated as the EU testing set (TST_{EU}), and then nine distinct training sets (TRN_{EU}) of varying sample sizes were randomly drawn: $n_{TRN_{EU}} = \{0,\ 5{,}000,\ 10{,}000,\ 25{,}000,\ 50{,}000,\ 75{,}000,\ 100{,}000,\ 150{,}000,\ 200{,}000,\ 250{,}000\}$. The training sets, TRN_{EU}, did not include any individuals from the testing set, TST_{EU}. Additionally, smaller training sets were subsets of the larger training sets.

A random sample of $n_{TST_{AF}} = 9{,}078$ individuals was selected to be the AF testing set (TST_{AF}), and nine distinct training sets (TRN_{AF}) of varying sample sizes were randomly drawn: $n_{TRN_{AF}} = \{0,\ 5{,}000,\ 7{,}500,\ 10{,}000,\ 15{,}000,\ 20{,}000,\ 25{,}000,\ 30{,}000,\ 35{,}000,\ 40{,}000\}$. The training sets, TRN_{AF}, did not include any individuals from the testing set, TST_{AF}. Additionally, smaller training sets were subsets of the larger training sets.

Genotypes

Since this analysis involved combining data from two cohorts, UKB and AoU, we used the intersection between the AoU genotyped SNPs and the UKB imputed SNPs. SNPs were excluded from this set if they had a minor allele frequency of less than 0.01 or missingness of over 0.1 in either dataset (among the full sample set), resulting in p = 522,170 SNPs retained for analysis. If a sample subset contained a missing SNP, it was imputed with the mean.

Phenotypes

For the AF cohort, the height measurement selected was the one closest to 60 years old for each individual, and outliers for height were removed, defined as larger or smaller than the median ± three times that of the middle 50th percentile for height. For the EU cohort, the height measurement selected was from the first instance or the second if the first was missing.

Some steps in this study involved preadjusting the height phenotype (e.g., the effect estimation for the PGS). For this, the residuals from an ordinary least squares (OLS) regression of height on sex, age, and the first five genotyped principal components were used as the adjusted phenotype. The EU OLS regression also included batch and center.

Genome-wide association study (GWAS)

A GWAS for height, including sex, age, and the first five SNP-derived PCs as additional covariates, was evaluated for each sample size set and for each data cohort (EU and AF) using PLINK[36]. That is, a single marker regression for TRN_*, where '*' = EU or AF, was evaluated as:

$$\mathbf{y}_* = \mathbf{SNP}_{*,j}\beta_{*,j} + \mathbf{Z}_*\alpha_* + \mathbf{e}_* \quad [10]$$

for $j = 1...p$ SNPs, where $\mathbf{y}_*$ is the height vector for the '*' = EU or AF ancestry group and $\mathbf{SNP}_{*,j}$ is the vector of the number of allele copies for the j^{th} SNP and the '*' = EU or AF ancestry group. $\mathbf{Z}_* \in \mathbb{R}^{n_{TRN_*} \text{ x } 7}$ is a predictor matrix, consisting of sex, age, and PC1 to PC5 (the first five genotype-derived principal components).

To obtain GWAS p-values that considered both data cohorts (EU and AF), we combined the ancestry group GWAS SNP effects ($\hat{\beta}_{EU,j}$ and $\hat{\beta}_{AF,j}$) estimated in [10] to obtain a meta-analysis-based estimate, $\hat{\beta}_{META,j}$, for each SNP ($j = 1...p$)[37]:

$$\hat{\beta}_{META,j} = \frac{w_{EU,j}\hat{\beta}_{EU,j} + w_{AF,j}\hat{\beta}_{AF,j}}{w_{EU,j} + w_{AF,j}} \quad [11]$$

where $w_{EU,j} = \frac{1}{SE(\hat{\beta}_{EU,j})^2}$ and $w_{AF,j} = \frac{1}{SE(\hat{\beta}_{AF,j})^2}$. The variance of the meta-estimator is $SE(\hat{\beta}_{META,j}) = \sqrt{\frac{1}{w_{EU,j} + w_{AF,j}}}$ and the meta-test statistic for each SNP ($j = 1...p$) was defined as:

$$Z_{META,j} = \frac{\hat{\beta}_{META,j}}{SE(\hat{\beta}_{META,j})}. \quad [12]$$

The final meta-GWAS p-value was then defined to be twice the area under a standard normal

distribution of negative infinity to the negative absolute value of $Z_{META,j}$ [12].

PGS and prediction accuracy calculation

SNP selection. The primary difference between the PGS evaluated in each of the scenarios described above was in the SNPs entering into each PGS. In all three scenarios, in the cases where $n_{TRN_{EU}} = 0$ or $n_{TRN_{AF}} = 0$, the single-cohort GWAS p-value was used instead of the meta-GWAS.

In Scenario 1 ('A typical PGS'), the number of SNPs varied based on the training sample sizes $n_{TRN_{EU}}$ and $n_{TRN_{AF}}$. For each sample size combination, the SNPs were selected for the PGS if the meta-GWAS p-value [12], if applicable, was less than 1e-4.

In Scenario 2 ('Isolating effect estimation'), SNPs entering into each PGS were selected if they had a p-value < 1e-4 based on the meta-GWAS [12], if applicable, using the sample sets 1) $n_{TRN_{EU}} = 100{,}000$ and $n_{TRN_{AF}} = 25{,}000$ to select *p*=5,234 SNPs and 2) $n_{TRN_{AF}} = 25{,}000$ and $n_{TRN_{EU}} = 25{,}000$ to select *p*=817 SNPs.

In Scenario 3 ('SNP portability'), the SNPs entering into each PGS were selected if they had a p-value < 1e-2 from the meta-GWAS, if applicable. Then, the PGS for each sample size combination was split into two sub-PGS. Since the SNPs were subset into two PGS for each sample size combination, the threshold for selection was relaxed to allow for an adequate number of SNPs in each PGS. One PGS sub-score was based on the SNPs with the top 20% of predicted portability, and the second was based on the SNPs with the lowest 20% of predicted portability. Portability was defined by the MC-ANOVA[14] predicted cross-ancestry R-squared for a small chromosome segment based on UKB (EU to AF) array data: $R^2_{EU\rightarrow AF} = \mathrm{Corr}\left(\mathbf{x}'_{i_{AF}}\boldsymbol{\beta}_{EU}, \mathbf{z}'_{i_{AF}}\boldsymbol{\alpha}\right)^2$, where $\mathbf{z}_{i_{AF}}$ is the (centered) vector of SNP genotypes at causal loci (QTL) for the AF ancestry group, $\mathbf{x}_{i_{AF}}$ is the (centered) vector of SNP genotypes at markers for the AF ancestry group, $\boldsymbol{\alpha}$ is

the vector of effects, and $\boldsymbol{\beta}_{EU} = \boldsymbol{\Sigma}_{X_{EU}}^{-1}\boldsymbol{\Sigma}_{X_{EU}Z_{EU}}\boldsymbol{\alpha}$ are the EU ancestry group (population) marker effects. The portability estimates used were obtained from the portability maps provided by Lupi et al., 2024[14].

SNP effects. Once the SNP set was determined, every PGS was based on summary statistics from the AF and EU training cohorts. To jointly estimate effects for *J* PGS SNPs, $\hat{\mathbf{b}}$, we fit a Bayesian ridge regression model using a Markov Chain Monte Carlo (MCMC) Gibbs sampler algorithm:

$$\mathbf{y}_* = 1\mu_* + \sum_{j=1}^{J} \mathbf{SNP}_{*,j}\, \mathbf{b}_{*,j} + \boldsymbol{e}_* \qquad [13]$$

where, for the '*' = EU or AF ancestry group, $\mathbf{y}_*$ is the preadjusted vector of height, μ_* is an intercept, $\mathbf{SNP}_{*,j}$ is the vector of the number of allele copies for the j^{th} SNP, and $\boldsymbol{e}_* = \{e_{1*}, \ldots e_{n*}\}$ are independent normal residuals. The residual variance has a scaled inverse-chi squared prior and the shrinkage parameter lambda (λ_*) in the Bayesian ridge regression prior was kept fixed for each ancestry group. The ridge estimator is $\hat{\mathbf{b}}_{*,j} = (\mathbf{X}_*'\mathbf{X}_* + \lambda_*\mathbf{I})^{-1}\mathbf{X}_*'\mathbf{y}_*$, where $\mathbf{X}_*$ is the design matrix of the intercept and SNPs. The prior mean, μ_b, was defined as an unweighted average across the two training cohorts:

$$\mu_b = \frac{\left(\sum_{i=1}^{n_{AF}} \boldsymbol{y}_{TR\ \ AF} + \sum_{i=1}^{n_{EU}} \boldsymbol{y}_{TRN_{EU}}\right)}{n_{TRN_{AF}} + n_{TRN_{EU}}}, \qquad [14]$$

and the prior variance, since the two cohorts are independent, was defined simply as the unweighted combined variance:

$$\sigma_b^2 = \frac{\left((\boldsymbol{y}'\boldsymbol{y})_{TRN_{AF}} + (\boldsymbol{y}'\boldsymbol{y})_{TRN_{EU}}\right)}{n_{TRN_{AF}} + n_{TR\ \ EU}} - {\mu_b}^2, \qquad [15]$$

where μ_b is the expression defined in [14], $\mathbf{y}_{TRN_{AF}}$ and $\mathbf{y}_{TRN_{EU}}$ are the AF and EU preadjusted height vectors, respectively, and $n_{TRN_{AF}}$ and $n_{TR\ \ EU}$ are the ancestry group-specific training set sample sizes.

We ran the BLR algorithm with 50,000 MC iterations and used a burn-in of 2,000 using the 'BLRCross' function from the R package BGLR[38]. This function takes summary statistics as inputs rather than the traditional inputs of an incidence matrix **X** and phenotype vector. That is, for p SNPs entering into a PGS, the model involved the summary statistics $\mathbf{X}'\mathbf{X}_{TRN_*}$, $\mathbf{X}'\mathbf{y}_{TRN_*}$, $\mathbf{y}'\mathbf{y}_{TRN_*}$, and n_{TRN_*}, where, for the ancestry group subscript (* = AF or EU), n_{TRN_*} is the sample size, $\mathbf{X}_{TRN_*} \in \mathbb{R}^{n\,x\,p}$ is the centered and imputed matrix of genotypes coded as 0, 1, or 2 (the count of reference allele at each SNP), and $\mathbf{y}_{TRN_*} \in \mathbb{R}^{n\,x\,1}$ is the preadjusted vector of the height phenotype. The training sets were then combined additively (without weights) into one set of summary statistics, and the combined summary statistics are were entered into the BLR algorithm described above.

Prediction accuracy estimation. To evaluate each PGS in a testing set, we used summary statistics from each test set (for AF, $\mathbf{X}'\mathbf{X}_{TST_{AF}}$, $\mathbf{X}'\mathbf{y}_{TST_{AF}}$, and $\mathbf{y}'\mathbf{y}_{TST_{AF}}$, and for EU, $\mathbf{X}'\mathbf{X}_{TST_{EU}}$, $\mathbf{X}'\mathbf{y}_{TST_{EU}}$, and $\mathbf{y}'\mathbf{y}_{TST_{EU}}$) and the estimated SNP effects, $\hat{\mathbf{b}}$, to estimate the prediction correlation for each ancestry group, $R_{TST_{AF}}$ and $R_{TST_{EU}}$ (for ease of notation we will drop the ancestry group subscript here, with the understanding that TST either equals TST_{AF} or TST_{EU}):

$$R_{TST} = Corr(\hat{\mathbf{y}}_{TST}, \mathbf{y}_{TST})$$

$$= \frac{Cov(\hat{\mathbf{y}}_{TST}, \mathbf{y}_{TST})}{\sqrt{Var(\mathbf{y}_{TST})Var(\hat{\mathbf{y}}_{TST})}}$$

$$= \frac{E(\hat{y}'_{TST}\mathbf{y}_{TST}) - \mu_{\hat{y}_{TST}}\mu_{y_{TST}}}{\sqrt{\frac{1}{n_{TST}}\mathbf{y}'\mathbf{y}_{TST}\hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}_{TST}\hat{\mathbf{b}}}}.$$

Since $E(\hat{\mathbf{y}}_{TST}) = \mathbf{X}\hat{\mathbf{b}}$, $E(\mathbf{y}_{TST}) = \mathbf{y}_{TST}$, and since $\mathbf{y}_{TST}$ is centered around zero, $\mu_{TST} = 0$:

$$= \frac{\hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}_{TST}}{\sqrt{\frac{1}{n_{TST}}\mathbf{y}'\mathbf{y}_{TST}\,\hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}_{TST}\hat{\mathbf{b}}}}. \qquad [16]$$

The prediction R-squared, R^2_{TST}, is the prediction correlation squared.

Results

In this study, we explored how factors, particularly sample size and the ancestry group of the training set, affected cross-ancestry prediction accuracy. We evaluated different scenarios of PGS varying the size of the training set consisting of two ancestry groups: EU ancestry data from the UKB and AF from AoU. Age, sex, and height were well-balanced across the sample size sets and datasets (Table C1). The average height across the UKB sets was 169.0 ± 9.2 cm and was 169.4 ± 9.8 cm across the AoU sets. The AoU samples were younger on average (48.8 ± 13.4 years old) and more female (56.3%) than the UKB sets (56.8 ± 8.0 years old and 53.3% female).

Scenario 1: A typical PGS (SNP filtering and estimation depend on sample size)

In this scenario, the training set was composed of varying sample sizes of AF and EU ($n_{TRN_{AF}}$ and $n_{TRN_{EU}}$), and was used for both SNP selection as well as SNP effect estimation. Figure 9a shows the prediction accuracy in the AF testing set, TST_{AF}, while Figure 9b shows the prediction accuracy in the EU testing set, TST_{EU}. In the AF testing set (Figure 9a), the pure within-ancestry (EU $n_{TRN_{EU}} = 0$) is the first column of results and the pure cross-ancestry (AF $n_{TRN_{AF}} = 0$) is the last row of results. The prediction correlation increased as the training sample size increased for both the pure cross-ancestry and pure within-ancestry prediction (Figure 9a). The maximum prediction correlation achieved with pure within-ancestry prediction was 0.19, and it was at the maximum sample size explored ($n_{TRN_{AF}} = 40{,}000$). This was, as expected, larger than the maximum prediction correlation achieved with pure cross-ancestry prediction ($R_{TS\ _{AF}} = 0.15$). The maximum pure cross-ancestry prediction correlation was approximately equivalent to pure within-ancestry at $n_{TRN_{AF}} = 25{,}000$ (Figure 9a), implying that for comparing strictly cross-ancestry and within-ancestry PGS, ten EU individuals were required for every one AF individual.

The 10:1 (EU:AF) relationship was not linear though (Figure 9a), as the ancestry ratio was 1.3:1 to achieve a prediction correlation of about half as much ($R_{TST_{AF}} = 0.08$), and the ratio was 5.6:1 to achieve a prediction correlation of about twice as much ($R_{TST_{AF}} = 0.29$).

The number of SNPs selected for each PGS varied depending on the training sample sizes used for the meta-GWAS, and, interestingly, the SNP sets filtered from AF only ($n_{TRN_{EU}} = 0$) tended to be compared to the SNP sets filtered from EU only ($n_{TRN_{AF}} = 0$). For example, when $n_{TRN_{EU}} = 0$ but $n_{TRN_{AF}} = 25{,}000$, 510 SNPs were selected. However, when this was reversed such that $n_{TRN_{AF}} = 0$ but $n_{TRN_{EU}} = 25{,}000$, 770 SNPs were selected. This could be because the EU ancestry group tends to have more LD compared to the AF ancestry group.

a. Scenario 1: African testing set, TST_{AF}

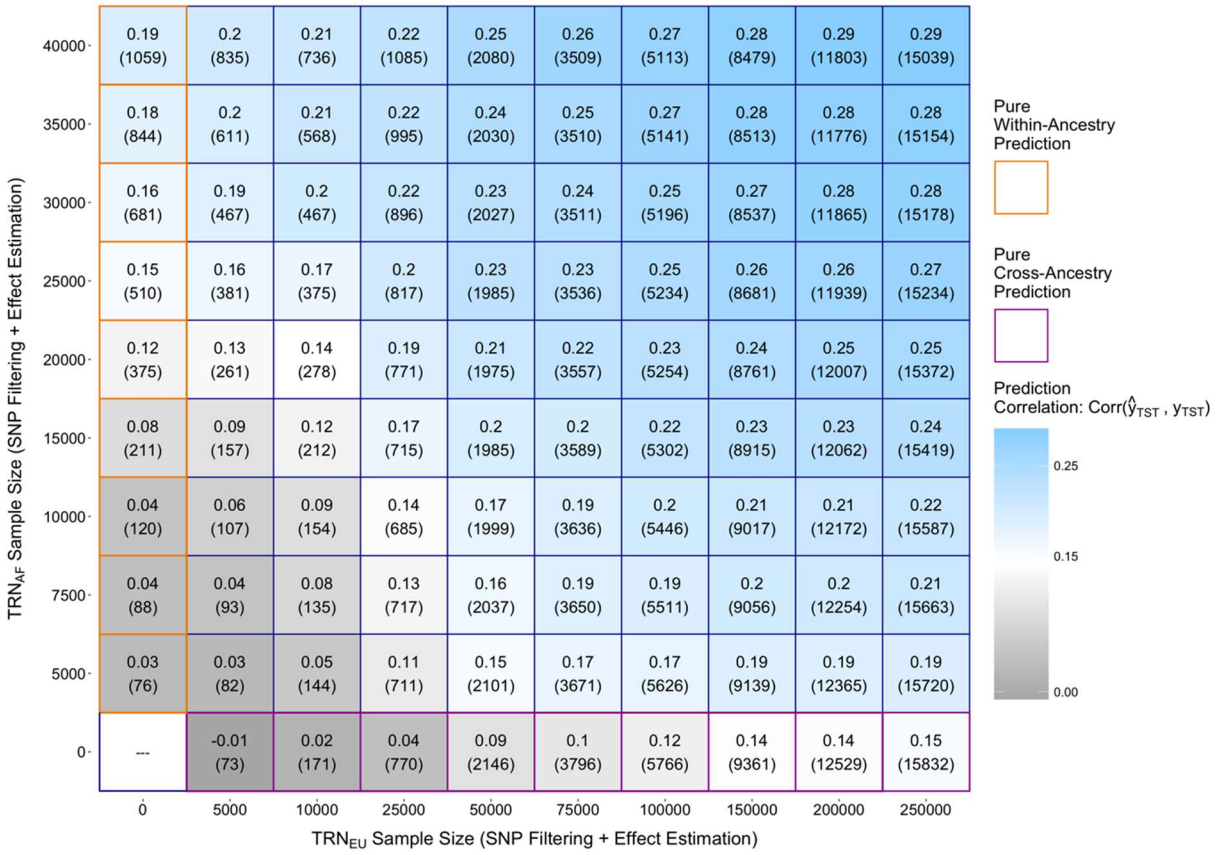
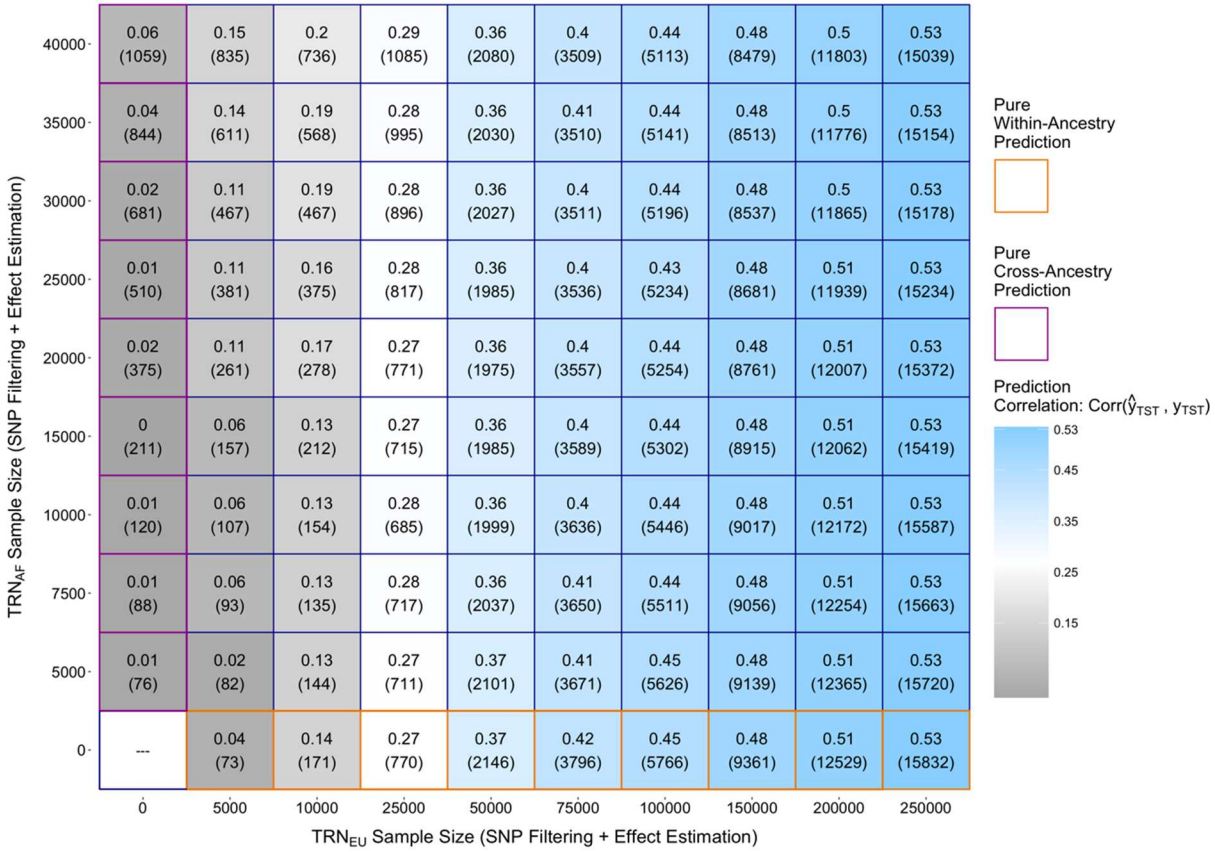


Figure 9: PGS varying the training sample sizes used to filter SNPs and estimate effects. The prediction correlation, R_{TST}, for height using different PGS for: **(a)** the African testing set, TST_{AF}, and **(b)** the European testing set, TST_{EU}, for different combinations of TRN_{AF} and TRN_{EU} sample sizes used for both SNP filtering and SNP effect estimation. SNPs entering into each PGS are based on the training sample size combinations ($n_{TRN_{AF}}$ and $n_{TRN_{EU}}$) and the number of SNPs, p, for each PGS is shown in parenthesis. For TST_{AF}, when there are no AF in the training ($n_{TRN_{AF}} = 0$), this is pure cross-ancestry prediction and when there are no EU in the training ($n_{TRN_{EU}} = 0$), this is pure within-ancestry prediction (and vice versa for TST_{EU}).

Figure 9 (cont'd)

b. Scenario 1: European testing set, TST_{EU}



When training the PGS using a combination of AF and EU data, prediction correlation testing in the AF group, TST_{AF}, tended to increase as both training sample sizes increased (Figure 9a), such that the maximum prediction correlation achieved was for the maximum sample sizes of both (AF $n_{TRN_{AF}} = 40{,}000$ and EU $n_{TRN_{EU}} = 250{,}000$). As seen comparing the pure within-ancestry and cross-ancestry PGS, the number of EU individuals included in the training data required to obtain equivalent prediction accuracy was more than that required of the AF number of individuals. For example, to achieve a prediction correlation of 0.27, four scenarios in Figure 9a achieved this: AF $n_{TRN_{AF}} = 35{,}000$ and 40,000 with EU $n_{TRN_{EU}} = 100{,}000$ (2.7:1), AF $n_{TRN_{AF}} = 30{,}000$ and EU $n_{TRN_{EU}} = 150{,}000$ (5:1), and AF $n_{TRN_{AF}} = 25{,}000$ and EU

$n_{TRN_{EU}} = 250{,}000$ (10:1). This means at this prediction accuracy level, the additional EU individuals added was worth increasingly less than adding more within-ancestry (AF) individuals. Conversely, we observed from the results in Figure 9b when testing in EU, TST_{EU}, that adding AF samples did not have much impact on prediction accuracy, rather, the prediction accuracy increased as the EU sample size increased.

Figure 10 has contour lines over the prediction correlations testing in AF (TST_{AF}), identifying the sample sizes required of each training ancestry to achieve equivalent prediction correlation. As the EU sample size ($n_{TRN_{EU}}$) increased in size, the contour lines started to level off, showing the relative decrease of information added after about $n_{TRN_{EU}} = 100{,}000$ EU individuals. This leveling off suggests that beyond some threshold for sample size, increasing the size of the non-target ancestry group in the training data (e.g., more EU individuals) contributes less to enhancing prediction accuracy and there are diminishing returns from additional data of this type. Additionally, the larger negative slopes at smaller EU sample sizes means that, while additional AF individuals increased prediction correlation more than the equivalent number of EU individuals, they were closer to a 1:1 equivalency (slope becomes closer to negative one) than at the higher EU sample sizes (where the slope becomes closer to zero).

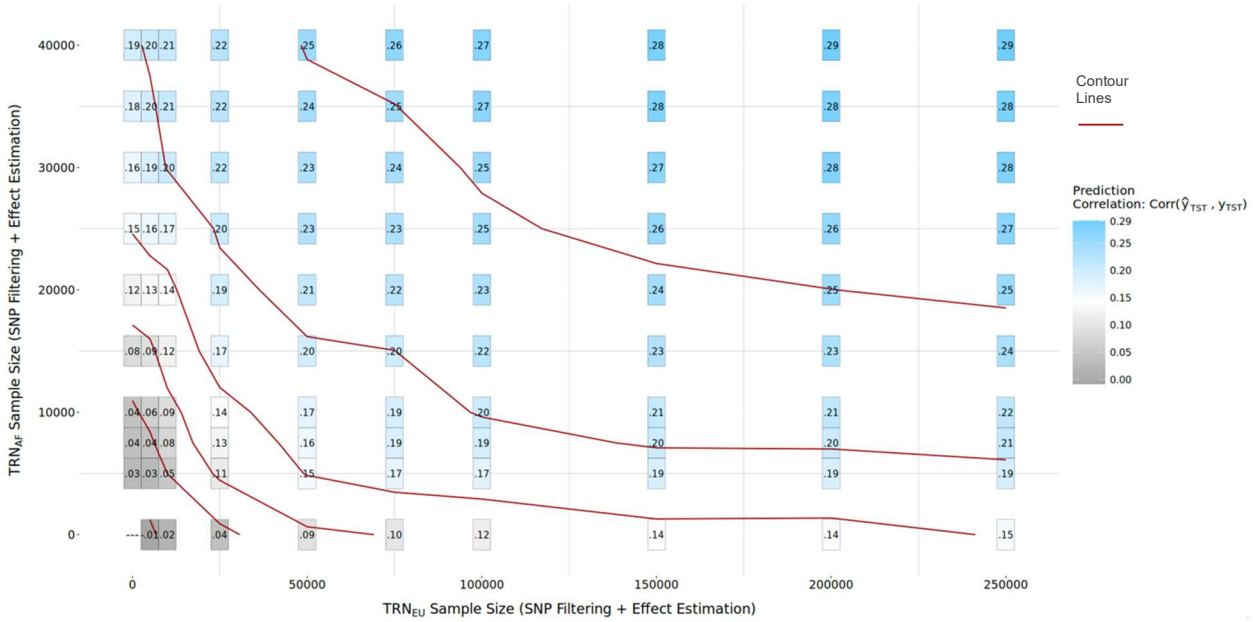


Figure 10: Prediction correlation contour lines. Contour lines of predictive correlations (shown in red) over the estimated prediction correlation, R_{TST}, for height using different PGS for the AF testing set, TST_{AF}. The axes show different combinations of AF and EU training set sample size combinations ($n_{TRN_{AF}}$ and $n_{TRN_{EU}}$). The training sample sizes were used for both SNP filtering and SNP effect estimation.

Scenario 2: Isolating effect estimation (fixing the SNP set)

To explore how SNP effect estimation accuracy affects PGS accuracy, we evaluated each PGS with a fixed SNP set of p = 5,234 SNPs (Figure 11) but estimated the SNP effects in different training sample sizes of EU and AF. In Figure 11a (testing in AF, TST_{AF}), as the EU sample size increased relative to the AF sample size, the prediction accuracy decreased since the SNP effect estimates converged toward the EU-specific estimates. The highest prediction accuracy was with the largest sample size of AF and a small number of EU ($n_{TRN_{AF}} = 40{,}000$ and $n_{TRN_{EU}} = 10{,}000$; Figure 11a). This is due to the EU sample having a substantially higher total size than the AF sample. For both testing in AF and EU, Figures 11a and 11b, respectively, as expected, pure cross-ancestry prediction (training in one ancestry group and testing in the other ancestry group) generally had the lowest prediction accuracy. When testing in the EU group

(Figure 11b), the highest accuracy was among the PGS with the largest sample size of EU. The range of prediction correlation estimates (Figure 11a) within the AF testing set, TST_{AF}, was narrow compared to the range in Scenario 1 (Figure 9a). Excluding pure within- and cross-ancestry cases, for the AF testing set, TST_{AF}, the range of prediction correlation was 0.19 – 0.28. Similarly, for the EU testing set, TST_{EU}, the range of prediction correlation was 0.33 – 0.46. In Scenario 1, both testing sets (TST_{AF} and TST_{EU}) had a range starting from nearly zero and a larger maximum.

When comparing the fixed SNP set combinations in Figure 11a to another fixed SNP set of p = 817 SNPs (Figure C1a), the prediction correlations were typically higher (although six cases when $n_{TRN_{AF}} = 0$ were smaller), which is what we hypothesized for PGS selecting a larger number of SNPs. The pattern of the correlation estimates was the same, in that the highest correlation estimates were among a smaller number of $n_{TRN_{EU}}$ rather than the maximum.

a. Scenario 2: African testing set, TST_{AF}

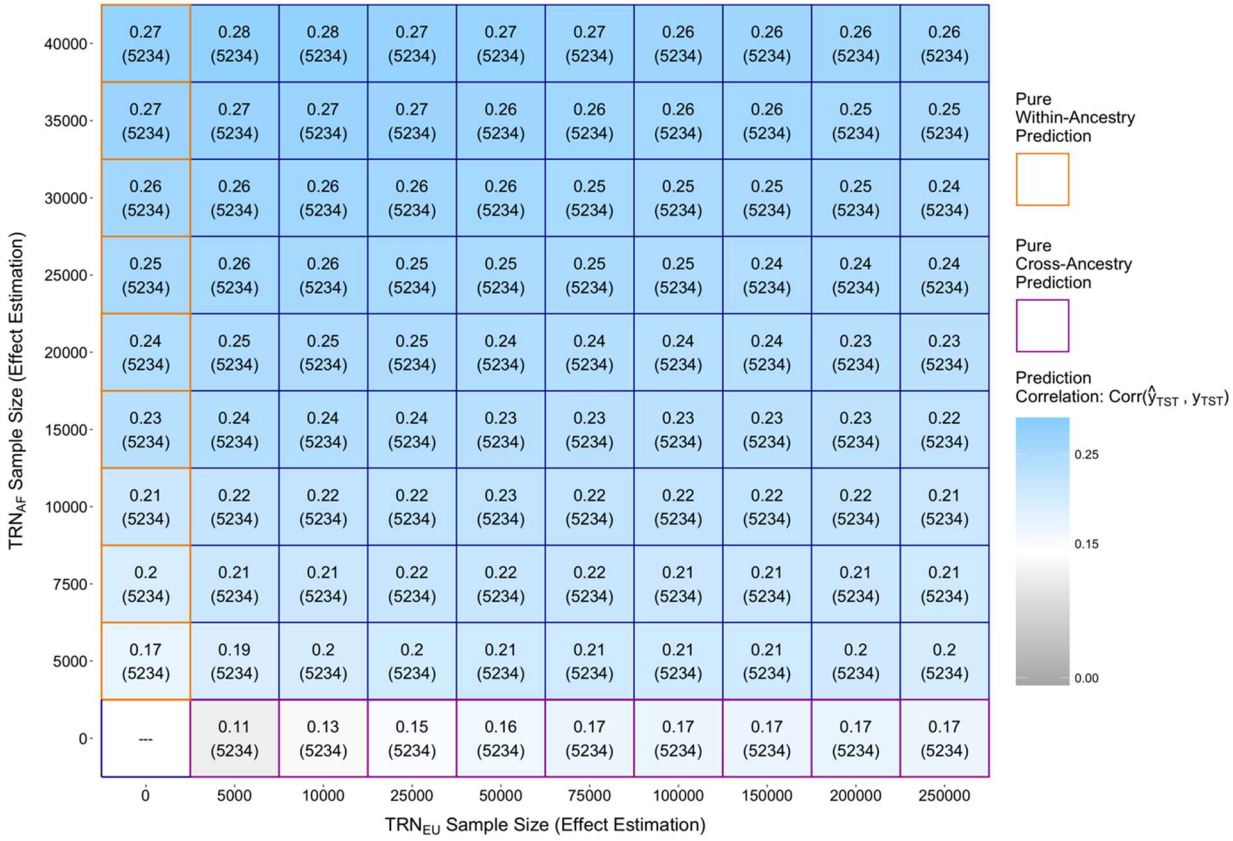
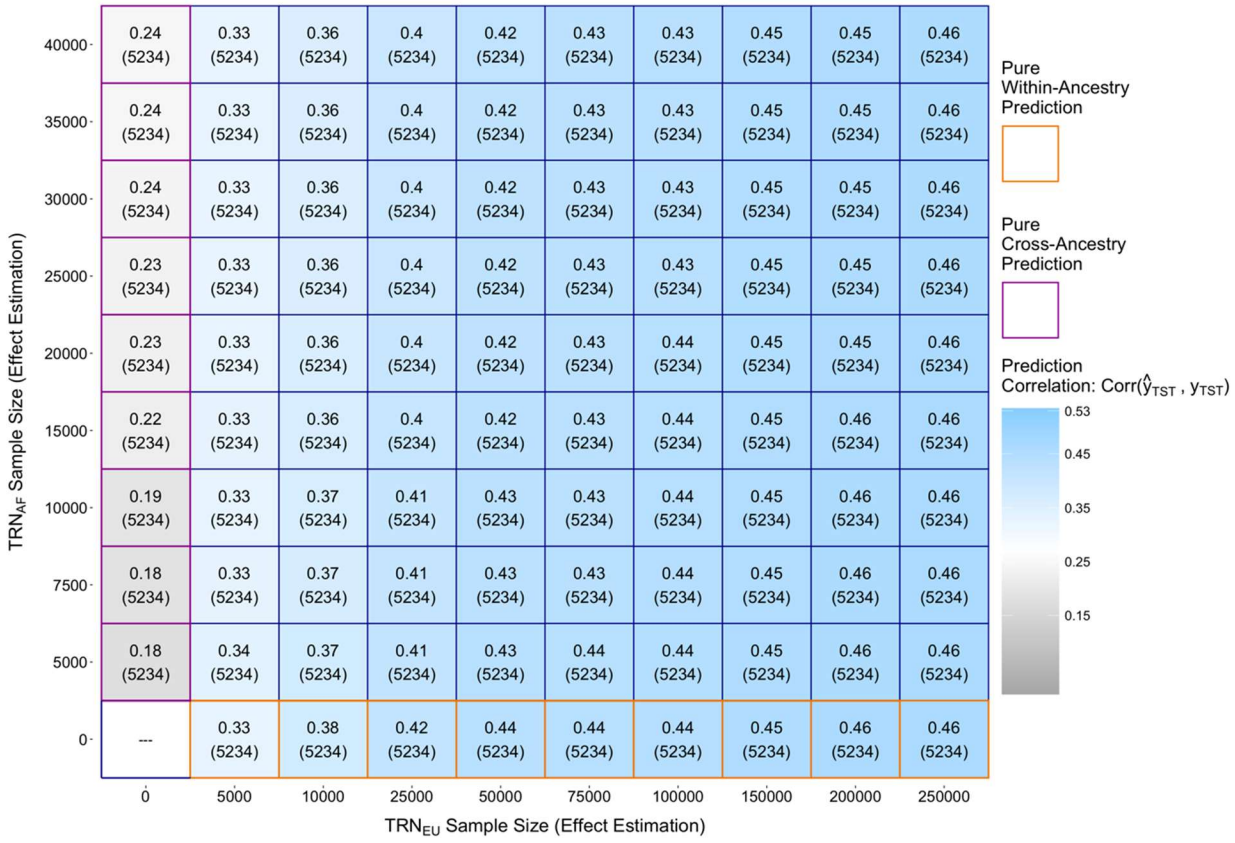


Figure 11: PGS fixing the SNP set but varying the training sample sizes used to estimate SNP effects. The prediction correlation, R_{TST}, for height using different PGS for: **(a)** the African testing set, TST_{AF}, and **(b)** the European testing set, TST_{EU}, for different combinations of TRN_{AF} and TRN_{EU} sample sizes used for effect estimation. The SNPs entering into each PGS are the same (p=5,234 SNPs) and is noted in parentheses. For TST_{AF}, when there are no AF in the training ($n_{TRN_{AF}} = 0$), this is pure cross-ancestry prediction and when there are no EU in the training ($n_{TRN_{EU}} = 0$), this is pure within-ancestry prediction (and vis versa for TST_{EU}). The scale of the prediction correlation is based on the values from Figure 9a for Figure 11a (and from Figure 9b for Figure 11b) to allow for straightforward comparison between the plots.

Figure 11 (cont'd)

b. Scenario 2: European testing set, TST_{EU}

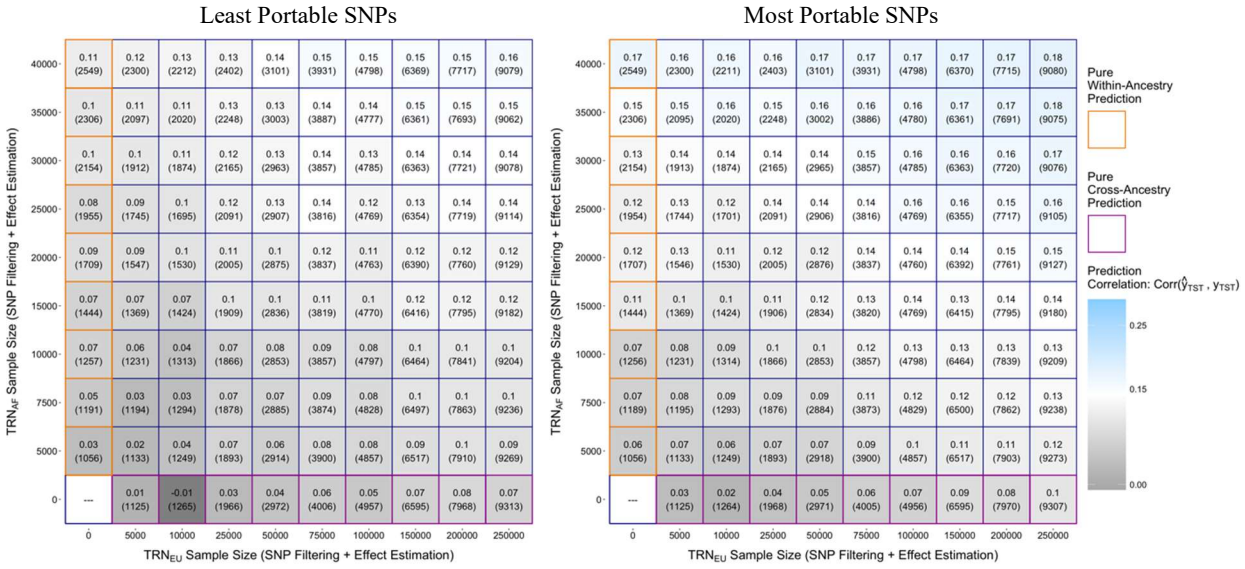


Scenario 3: SNP portability

In Scenario 3 we used cross-ancestry PGS portability estimates from the Relative Accuracy maps derived from the UK Biobank arrays, available at Lupi et al., 2024[14] (see 'SNP selection' in Methods for details on portability estimates). To build each PGS for height we partitioned the SNPs into two groups using a p < 1e-2 inclusion level. The two groups were the top 20th percentile of portable SNPs and the bottom 20th percentile of portable SNPs. Similar to Scenario 1, the training sample sizes ($n_{TRN_{AF}}$ and $n_{TRN_{EU}}$) for this scenario determined both SNP selection as well as SNP effect estimation. Figure 12a shows the prediction accuracy when testing in AF (TST_{AF}) for 99 combinations of training sample sizes (different sizes and ancestries) by SNP portability. Figure 12b shows the prediction accuracies for the same settings but testing in

EU (TST_{EU}). Figure 12a shows that even among the pure within-ancestry PGS, that did not involve any EU data, the PGS involving the SNPs predicted to be most portable across ancestry group had higher prediction correlation compared to the PGS involving the SNPs predicted to be the least portable. Interestingly, for the most portable SNPs testing in AF (TST_{AF}), the maximum prediction correlation achieved was only 0.18, and was with the maximum training sample sizes used for both SNP filtering and estimation ($n_{TRN_{AF}} = 40{,}000$ and $n_{TRN_{EU}} = 250{,}000$). For the least portable SNPs testing in AF (TST_{AF}), the maximum prediction correlation achieved was only 0.16, and was also with the maximum training sample sizes used for both SNP filtering and estimation ($n_{TRN_{AF}} = 40{,}000$ and $n_{TRN_{EU}} = 250{,}000$). While this sample size combination had fewer SNPs compared to that in Figure 9a (about 9,080 SNPs versus 15,039 SNPs, respectively), which used the same sample sizes for effect estimation but differed in the SNP set, other combinations had a larger number of SNPs entering into the PGS but still had poorer prediction accuracy. One example of this is for $n_{TRN_{AF}} = 25{,}000$ and $n_{TRN_{EU}} = 25{,}000$, in which the typical PGS (Figure 9a) had 817 SNPs and a prediction correlation of 0.20, while the most portable SNP-based PGS (Figure 12a) had 2,091 SNPs and a prediction correlation of 0.14 (0.12 for the least portable SNP-based PGS).

a. Scenario 3: African testing set, TST_{AF}



b. Scenario 3: European testing set, TST_{EU}

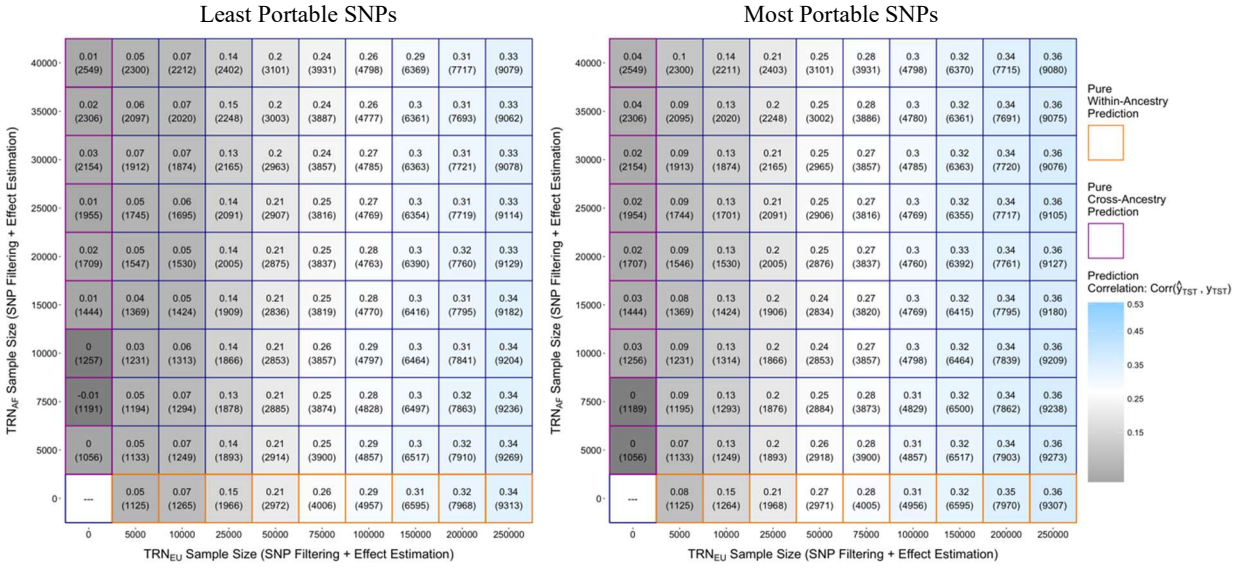


Figure 12: PGS subset by SNP portability. The prediction correlation, R_{TST}, for height using different PGS for: (**a**) AF (TST_{AF}) and (**b**) EU (TST_{EU}) when using the top 20% most portable SNPs (based on MC-ANOVA's cross-ancestry R-squared[14]) compared to the bottom 20% most portable SNPs by training set sample size (AF [TRN_{AF}] and EU [TRN_{EU}]). The scale of the prediction correlation is based on the values from Figure 9a for Figure 12a (and from Figure 9b for Figure 12b) to allow for straightforward comparison between the plots.

Comparing the most portable PGS to the least in the AF testing set, TST_{AF} (Figure C2a), the pure within-ancestry PGS had an average (median) increase in prediction R-squared (prediction correlation squared) of 103.8% (103.1%), the pure cross-ancestry PGS had an average (median) increase of 119.1% (83.7%), and the PGS involving both EU and AF in training had an average (median) increase of 78.8% (49.1%). In the EU testing set, TST_{EU} (Figure C2b), the most portable SNPs set tended to have higher prediction R-squared compared to the least portable SNPs. There was an average (median) increase of 158.5% (27.1%) across all PGS combinations (including pure within- and cross-ancestry PGS).

Our prediction correlations for Scenario 3 (Figure 12) are generally lower than in Scenario 1 (Figure 9). Of the cases when Scenario 1 had fewer SNPs than Scenario 3 (59 out of 99), only 13 of those had a smaller prediction correlation. The thirteen combinations with lower prediction correlation and a lower number of SNPs selected were all when $n_{TRN_{AF}} \leq 15{,}000$ and $n_{TRN_{EU}} \leq 25{,}000$. Since Scenario 3 represents the genomic regions with high portability instead of being genome-wide, if the SNPs with higher LD (portability) are located near one another, they could be picking up the same QTL signal, leading to less diversity in the signal being picked up by markers. Indeed, in Table C2 when defining AF peaks (unique local chromosome regions consisting of GWAS-significant hits in LD picking up the same QTL signal) from the two portability-based SNP hit sets (GWAS p-value < 1e-2), the SNP set consisting of the highly portable SNPs condensed to 511 peaks on average across sample size cases (on average 9.3% of the total SNPs) and the SNPs with low portability condensed to 500 peaks (8.6% of the total SNPs). Conversely, the Scenario 1 SNPs condensed 2,015 peaks on average (on average 50.2% of the total SNPs).

Discussion

In this study, we postulate that training sample size and ancestry group are the major limitations in cross-ancestry prediction accuracy. Previous studies have shown that the sample size used to train models affects prediction accuracy by influencing both the identification of phenotype-associated SNPs selected for inclusion in the PGS and the precision of SNP effect estimates[8–12]. To shed light on this problem, we evaluated polygenic scores (PGS) under various scenarios using both within and cross-ancestry training data from the UK Biobank and All of Us.

Using individuals from multiple ancestry groups is important in identifying genomic regions with QTL signal. However, we found that cross-ancestry prediction did not produce accurate marker effect estimates across ancestries. When the training sample size of the target ancestry group (AF) was very small (or zero), the estimated effects were poor and prediction accuracy among AF was low. However, when the AF ancestry group had a certain training sample size (e.g., at 15,000 AF samples in Scenario 2) it was more precise to use a European (EU) training size smaller than available since we observed that over-increasing the number of EU ancestry samples compared to the number of AF samples dominated the effect estimation, resulting in poorer prediction accuracy when testing among AF ancestry. These results might suggest that the QTL, or QTL effects, are not the same across ancestry groups, but Hou et al., 2023[39] suggest that causal variants tend to be similar across ancestry groups. Our results may align with this if different LD patterns exist between SNP markers and QTL across ancestry groups, as these variations could alter the QTL effects captured by markers in each group.

Indeed, previous studies have shown that there are LD and allele frequency differences between ancestry groups[13,14,19]. Thus, the QTL effect that we may capture in a marker in one ancestry group may not exist in the other. SNP portability describes the transferability of a PGS

across ancestry groups based on how similar the (local) genetic regions in the PGS are across ancestry groups. As shown by Lupi et al.[14], the higher the portability, the better conserved the LD and allele frequency in the region between groups and the more portable a cross-ancestry PGS will be. Classifying SNPs based on their portability across ancestries showed that SNPs in higher portability regions improved predictive accuracy in PGS compared to SNPs with lower portability (Scenario 3). This was true even in within-ancestry predictions. This suggests that cross-ancestry SNP portability is a valuable tool for identifying regions that are more suitable for prediction across ancestries and those that are less suitable.

Classifying SNPs by portability also demonstrated that the number of SNPs included in the PGS doesn't necessarily translate to better selection or identification of QTL or QTL markers. The SNP filtering threshold was less conservative (larger) in Scenario 3 compared to Scenario 1, yet under Scenario 3, some of the PGS yielded lower prediction correlation estimates (among TST_{AF}) even when more SNPs were selected. Since the portability-based SNP sets included substantially fewer GWAS peaks than the typical PGS, there was less variation in the QTL signal picked up by the portability-based SNP sets. This is similar to that observed by Kim et al., 2017[2], who compared SNP selection based on LD blocks to selecting the top SNPs independent of LD blocks and found that for a small number of SNPs, the top SNP method underperformed due to the top SNPs clustering in regions, thus, having poor genome coverage.

Our findings demonstrate the inefficiency of using cross-ancestry data (EU) compared to within-ancestry data (AF) for PGS predictions in AF individuals, similar to that found by Lehmann et al., 2023[6]. The study suggests that a significantly larger number of EU individuals is required to achieve the same prediction accuracy as a smaller number of AF individuals. This inefficiency is particularly pronounced at higher levels of prediction accuracy, where the required

EU sample size disproportionately increases compared to the AF sample size required. This non-linear relationship indicates diminishing returns from adding more cross-ancestry data beyond a certain point. Yet, our results highlight the need for further increasing non-EU data collection, since PGS involving both within and cross-ancestry data still greatly improved upon the pure within-ancestry prediction at the limited sample sizes available of AF individuals.

Our study has some limitations. First, this study used data from different cohorts, and both cohorts had different genotyping platforms. Therefore, we obtained a common set of SNPs with sufficient genome coverage (calls for All of Us and the UK Biobank imputed SNP set). This is a limitation since SNP imputation can induce artifacts related to the reference panels used for imputation, which are often EU-dominant. Additionally, imputed SNPs have a higher marker density and higher LD compared to genotyped SNPs. Another limitation is that this study only evaluated height. Thus, our results are not necessarily representative of other traits with different heritability, polygenicity, and genetic architecture. Lehmann et al., 2023[6], found that for cross-ancestry prediction, the optimal training strategy, e.g., the sample sizes of each ancestry (EU and non-EU), varied substantially depending on the trait. Nevertheless, their conclusion that additional non-EU genomic data collection is critical is consistent with our findings.

These findings have important implications for genomic research and the development of PGS. First, they highlight the necessity of increasing sample sizes for non-European ancestry groups to achieve more accurate prediction. Second, our findings show the value of cross-ancestry information borrowing to identify genomic regions with QTL signals. Finally, they highlight the limitations in estimating effects across ancestry groups. Overall, by highlighting the limitations and inefficiencies in using cross-ancestry data, our findings advocate for prioritizing the continuation of ongoing efforts in collecting data for underrepresented ancestry

groups.

REFERENCES

1. GWAS Catalog [Internet]. [cited 2023 Apr 20]. Available from: https://www.ebi.ac.uk/gwas/

2. Kim H, Grueneberg A, Vazquez AI, Hsu S, de los Campos G. Will Big Data Close the Missing Heritability Gap? Genetics. 2017;207(3):1135–45.

3. Lello L, Avery SG, Tellier L, Vazquez AI, de los Campos G, Hsu SDH. Accurate Genomic Prediction of Human Height. Genetics. 2018;210(2):477–97.

4. de los Campos G, Vazquez AI, Hsu S, Lello L. Complex-Trait Prediction in the Era of Big Data. Trends Genet. 2018;34(10):746–54.

5. Albiñana C, Zhu Z, Schork AJ, Ingason A, Aschard H, Brikell I, et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. Nat Commun. 2023 Aug 5;14(1):4702.

6. Lehmann B, Mackintosh M, McVean G, Holmes C. Optimal strategies for learning multi-ancestry polygenic scores vary across traits. Nat Commun. 2023 Jul 7;14(1):4023.

7. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat Commun. 2018 Mar 7;9(1):989.

8. de los Campos G, Sorensen D, Gianola D. Genomic Heritability: What Is It? Barsh GS, editor. PLoS Genet. 2015 May 5;11(5):e1005048.

9. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 2009 Jun;136(2):245–57.

10. Goddard ME, Wray NR, Verbyla K, Visscher PM. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statist Sci. 2009 Nov 1;24(4):517-529.

11. Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet. 2011 Dec;128(6):409–21.

12. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. Weedon MN, editor. PLoS ONE. 2008 Oct 14;3(10):e3395.

13. Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020 Jul 31;11(1):3865.

14. Lupi AS, Vazquez AI, de los Campos G. Mapping the relative accuracy of cross-ancestry prediction. Nat Commun. 2024; *accepted 2024*.

15. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019 Mar;177(1):26–31.

16. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.

17. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019 Jul 25;10(1):3328.

18. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. The American Journal of Human Genetics. 2015 Oct;97(4):576–92.

19. Privé F, Aschard H, Carmi S, Folkersen L, Hoggart C, O'Reilly PF, 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. 2022 Jan 6;109(1):12–23.

20. Zhao Z, Fritsche LG, Smith JA, Mukherjee B, Lee S. The construction of cross-population polygenic risk scores using transfer learning. Am J Hum Genet. 2022 Nov 3;109(11):1998–2008.

21. Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. Human Genetics and Genomics Advances. 2021 Jan;2(1):100017.

23. Guo J, Bakshi A, Wang Y, Jiang L, Yengo L, Goddard ME, et al. Quantifying genetic heterogeneity between continental populations for human height and body mass index. Sci Rep. 2021 Mar 4;11(1):5240.

24. Ding Y, Hou K, Xu Z, Pimplaskar A, Petter E, Boulier K, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature. 2023 Jun 22;618(7966):774–81.

25. Shi H, Burch KS, Johnson R, Freund MK, Kichaev G, Mancuso N, et al. Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. The American Journal of Human Genetics. 2020 Jun;106(6):805–17.

26. Hu S, Ferreira LAF, Shi S, Hellenthal G, Marchini J, Lawson DJ, et al. Leveraging fine-scale population structure reveals conservation in genetic effect sizes between human populations across a range of human phenotypes [Internet]. 2023 [cited 2024 Aug 26]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.08.08.552281

27. Majara L, Kalungi A, Koen N, Tsuo K, Wang Y, Gupta R, et al. Low and differential polygenic score generalizability among African populations due largely to genetic diversity. Human Genetics and Genomics Advances. 2023 Apr;4(2):100184.

28. Wang Y, Namba S, Lopera E, Kerminen S, Tsuo K, Läll K, et al. Global Biobank analyses

provide lessons for developing polygenic risk scores across diverse cohorts. Cell Genomics. 2023 Jan;3(1):100241.

29. Ruan Y, Lin YF, Feng YCA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nat Genet. 2022 May;54(5):573–80.

30. Márquez-Luna C, Loh PR, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol. 2017 Dec;41(8):811–23.

31. Hoggart CJ, Choi SW, García-González J, Souaiaia T, Preuss M, O'Reilly PF. BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. Nat Genet. 2024 Jan;56(1):180–6.

32. Mester R, Hou K, Ding Y, Meeks G, Burch KS, Bhattacharya A, et al. Impact of cross-ancestry genetic architecture on GWASs in admixed populations. The American Journal of Human Genetics. 2023 Jun;110(6):927–39.

33. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

34. The All of Us Research Program Investigators. The "All of Us" Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.

35. Controlled CDR Directory [Internet]. User Support. 2024 [cited 2024 Oct 18]. Available from: https://support.researchallofus.org/hc/en-us/articles/4616869437204-Controlled-CDR-Directory

36. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaSci. 2015 Dec;4(1):7.

37. Lee CH, Cook S, Lee JS, Han B. Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. Genomics Inform. 2016;14(4):173.

38. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014 Oct;198(2):483–95.

39. Hou K, Ding Y, Xu Z, Wu Y, Bhattacharya A, Mester R, et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat Genet. 2023 Mar 20;55:549-558.

APPENDIX C: Chapter 3

Supplementary Figures

a. Scenario 2: African testing set, TST_{AF}, with a p = 817 SNP set

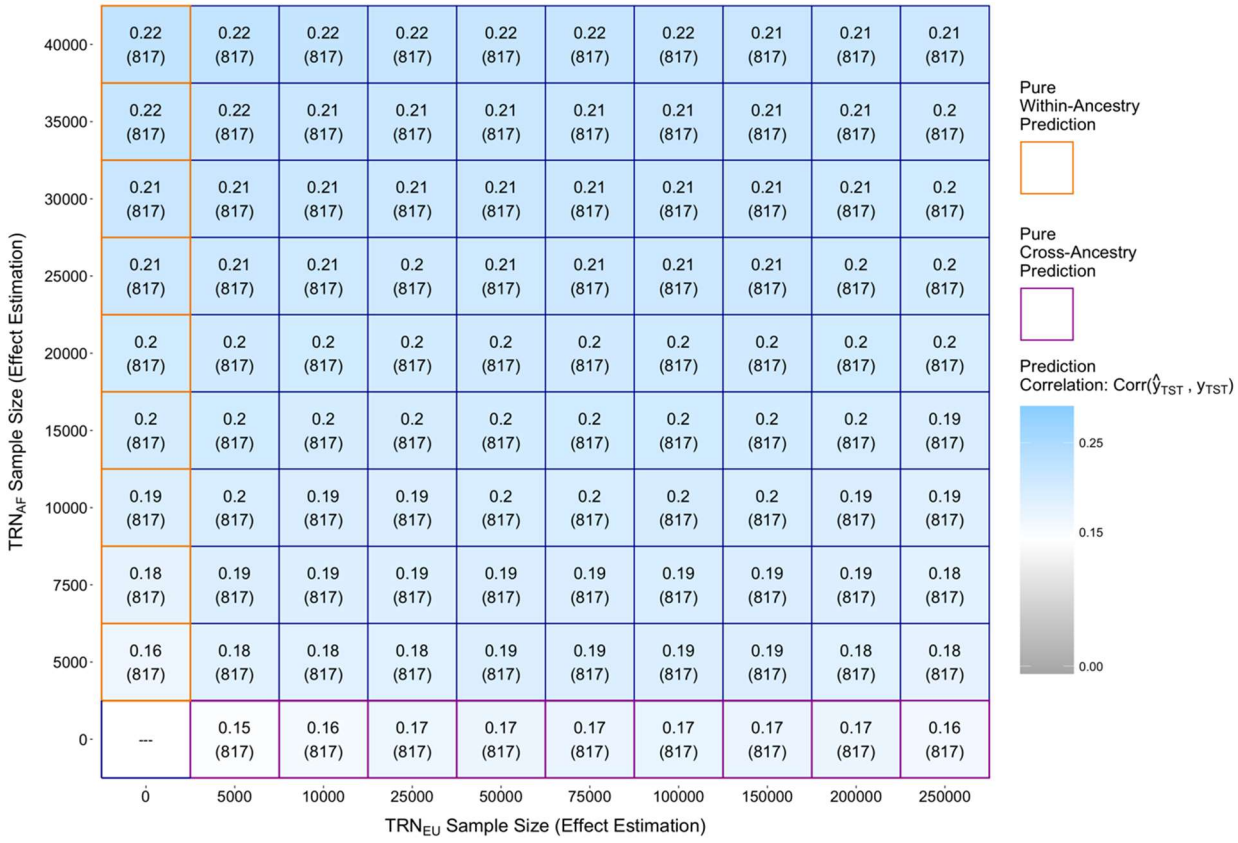
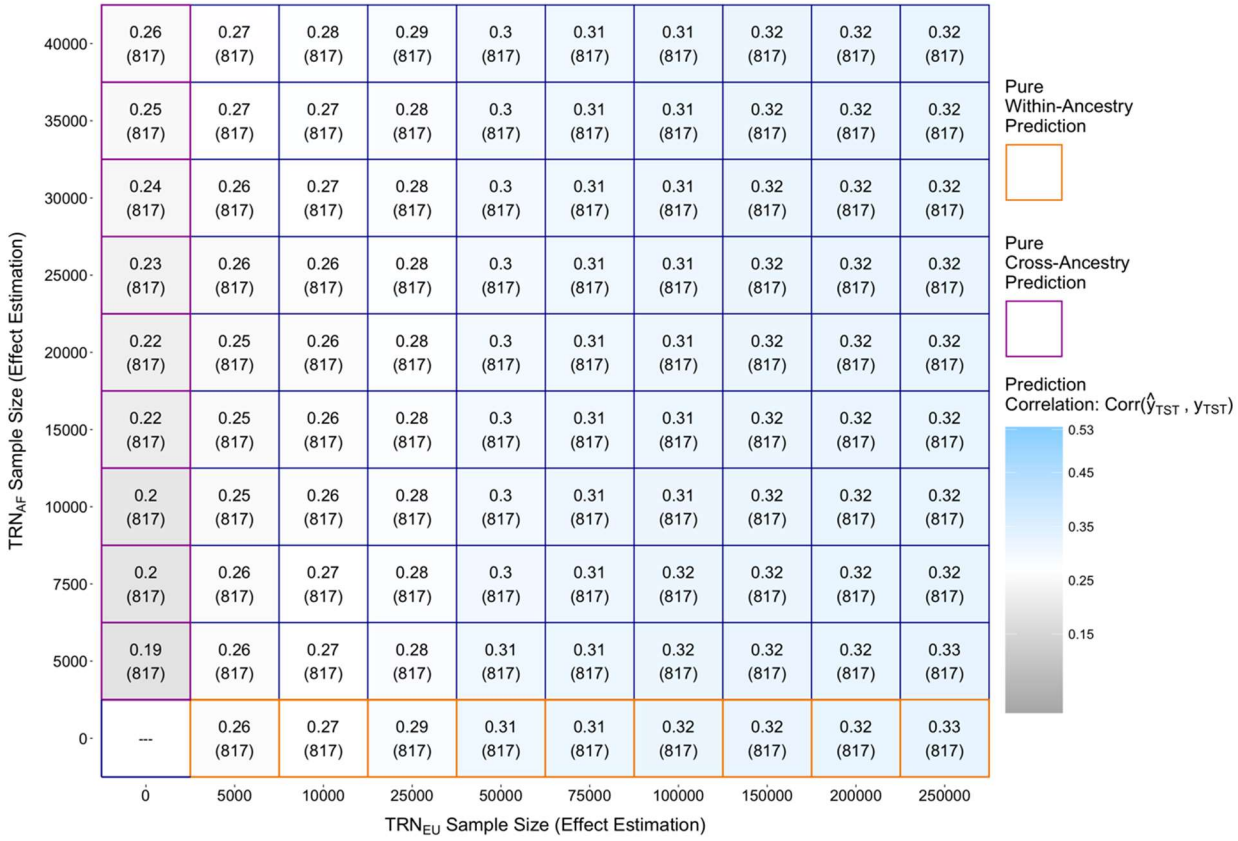


Figure C1: PGS fixing the SNP set but varying the training sample sizes used to estimate SNP effects. The prediction correlation, R_{TST}, for height using different PGS for: **(a)** the African testing set, TST_{AF}, and **(b)** the European testing set, TST_{EU}, for different combinations of TRN_{AF} and TRN_{EU} sample sizes used for effect estimation. The SNPs entering into each PGS are the same (p = 817 SNPs) and is noted in parentheses. For TST_{AF}, when there are no AF in the training ($n_{TRN_{AF}} = 0$), this is pure cross-ancestry prediction and when there are no EU in the training ($n_{TRN_{EU}} = 0$), this is pure within-ancestry prediction (and vis versa for TST_{EU}). The scale of the prediction correlation is based on the values from Figure 9a for Figure 11a (and from Figure 9b for Figure 11b) to allow for straightforward comparison between the plots.

Figure C1 (cont'd)

b. Scenario 2: European testing set, TST_{EU}, with a p = 817 SNP set

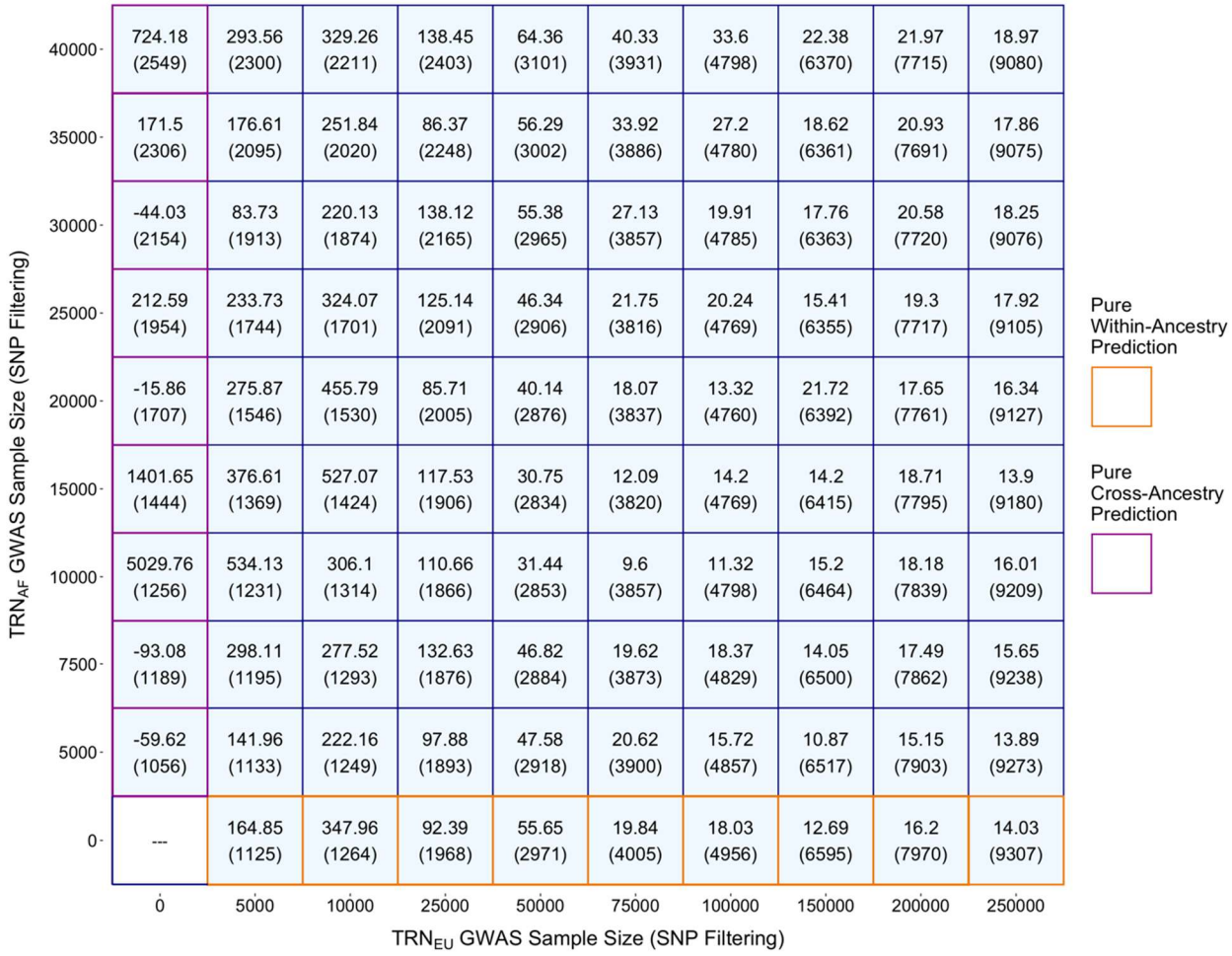


a. Scenario 3: African testing set, TST_{AF}

Figure C2: The percent increase in prediction R-squared, $\frac{\left(R^2_{TST}(most\ portable) - R^2_{TST}(least\ portable)\right)}{R^2_{TST}(least\ portable)} \times 100\%$[14], for height using different PGS for: (**a**) AF (TST_{AF}) and (**b**) EU (TST_{EU}) when using the top 20% most portable SNPs (based on MC-ANOVA's cross-ancestry R-squared[14]) compared to the bottom 20% most portable SNPs by training set sample size (AF [TRN_{AF}] and EU [TRN_{EU}]) used for SNP filtering and estimation.

Figure C2 (cont'd)

b. Scenario 3: European testing set, TST_{EU}



Supplementary Tables

Table C1: Descriptive statistics of the European (EU) and African (AF) training and testing sets. Continuous traits are described by the mean plus or minus one standard deviation.

Ancestry Group	Sample Size	Female (%)	Age (years)	Height (cm)
European (EU)	n=5,000	53.7	56.9 ± 8.0	168.9 ± 9.2
	n=10,000	53.2	56.8 ± 8.0	169.0 ± 9.1
	n=25,000	53.1	56.7 ± 8.0	169.1 ± 9.2
	n=50,000	53.3	56.8 ± 8.0	169.0 ± 9.2
	n=75,000	53.2	56.8 ± 8.0	169.1 ± 9.2
	n=100,000	53.1	56.8 ± 8.0	169.1 ± 9.2
	n=150,000	53.2	56.8 ± 8.0	169.1 ± 9.2
	n=200,000	53.3	56.8 ± 8.0	169.1 ± 9.2
	n=250,000	53.2	56.8 ± 8.0	169.1 ± 9.2
	Testing Set (n=10,000)	54.1	56.7 ± 8.0	168.9 ± 9.2
African (AF)	n=5,000	55.7	48.7 ± 13.4	169.5 ± 9.8
	n=7,500	55.7	48.7 ± 13.4	169.5 ± 9.8
	n=10,000	55.9	48.6 ± 13.5	169.4 ± 9.9
	n=15,000	56.5	48.7 ± 13.5	169.4 ± 9.8
	n=20,000	56.6	48.8 ± 13.5	169.4 ± 9.8
	n=25,000	56.6	48.8 ± 13.4	169.3 ± 9.7
	n=30,000	56.6	48.9 ± 13.4	169.4 ± 9.7
	n=35,000	56.7	48.9 ± 13.4	169.3 ± 9.8
	n=40,000	56.7	48.9 ± 13.4	169.3 ± 9.8
	Testing Set (n=9,078)	56.0	49.1 ± 13.4	169.5 ± 9.7

Table C2: The number of AF peaks at an R-squared threshold of 0.1 for each sample size combination SNPs (EU and AF). SNP set refers to the scenario under which the SNPs were selected. '1e-4' is the p-value cutoff used in Scenario 1, and '1e-2 Low' and '1e-2 High' are the p-value cutoffs and portability sets used in Scenario 3.

AF Sample Size	EU Sample Size	SNP Set	# of Peaks	Total # of SNPs	# of Peaks out of # of SNPs (%)
5000	0	1e-4	72	75	96.0
7500	0	1e-4	76	87	87.4
10000	0	1e-4	102	119	85.7
15000	0	1e-4	147	210	70.0
20000	0	1e-4	205	374	54.8
25000	0	1e-4	273	509	53.6
30000	0	1e-4	351	680	51.6
35000	0	1e-4	413	843	49.0
40000	0	1e-4	513	1058	48.5
5000	5000	1e-4	70	81	86.4
7500	5000	1e-4	74	92	80.4
10000	5000	1e-4	83	106	78.3
15000	5000	1e-4	123	156	78.8
20000	5000	1e-4	170	260	65.4
25000	5000	1e-4	238	380	62.6
30000	5000	1e-4	282	466	60.5
35000	5000	1e-4	346	610	56.7
40000	5000	1e-4	455	834	54.6
5000	10000	1e-4	97	143	67.8
7500	10000	1e-4	100	134	74.6
10000	10000	1e-4	109	153	71.2
15000	10000	1e-4	155	211	73.5
20000	10000	1e-4	180	277	65.0
25000	10000	1e-4	233	374	62.3
30000	10000	1e-4	289	466	62.0
35000	10000	1e-4	332	567	58.6
40000	10000	1e-4	424	735	57.7
5000	25000	1e-4	345	710	48.6
7500	25000	1e-4	366	716	51.1
10000	25000	1e-4	353	684	51.6
15000	25000	1e-4	369	714	51.7
20000	25000	1e-4	405	770	52.6
25000	25000	1e-4	433	816	53.1
30000	25000	1e-4	475	895	53.1
35000	25000	1e-4	507	994	51.0
40000	25000	1e-4	553	1084	51.0
5000	50000	1e-4	886	2100	42.2
7500	50000	1e-4	878	2036	43.1
10000	50000	1e-4	879	1998	44.0
15000	50000	1e-4	896	1984	45.2
20000	50000	1e-4	908	1974	46.0
25000	50000	1e-4	921	1984	46.4
30000	50000	1e-4	956	2026	47.2
35000	50000	1e-4	977	2029	48.2
40000	50000	1e-4	1022	2079	49.2
5000	75000	1e-4	1519	3670	41.4
7500	75000	1e-4	1524	3649	41.8
10000	75000	1e-4	1517	3635	41.7

Table C2 (cont'd)

15000	75000	1e-4	1513	3588	42.2
20000	75000	1e-4	1506	3556	42.4
25000	75000	1e-4	1512	3535	42.8
30000	75000	1e-4	1537	3510	43.8
35000	75000	1e-4	1547	3509	44.1
40000	75000	1e-4	1570	3508	44.8
5000	100000	1e-4	2228	5625	39.6
7500	100000	1e-4	2211	5510	40.1
10000	100000	1e-4	2188	5445	40.2
15000	100000	1e-4	2144	5301	40.4
20000	100000	1e-4	2136	5253	40.7
25000	100000	1e-4	2134	5233	40.8
30000	100000	1e-4	2150	5195	41.4
35000	100000	1e-4	2154	5140	41.9
40000	100000	1e-4	2165	5112	42.4
5000	150000	1e-4	3569	9138	39.1
7500	150000	1e-4	3533	9055	39.0
10000	150000	1e-4	3530	9016	39.2
15000	150000	1e-4	3478	8914	39.0
20000	150000	1e-4	3510	8760	40.1
25000	150000	1e-4	3463	8680	39.9
30000	150000	1e-4	3414	8536	40.0
35000	150000	1e-4	3422	8512	40.2
40000	150000	1e-4	3433	8478	40.5
5000	200000	1e-4	4880	12364	39.5
7500	200000	1e-4	4849	12253	39.6
10000	200000	1e-4	4816	12171	39.6
15000	200000	1e-4	4794	12061	39.7
20000	200000	1e-4	4795	12006	39.9
25000	200000	1e-4	4805	11938	40.2
30000	200000	1e-4	4784	11864	40.3
35000	200000	1e-4	4768	11775	40.5
40000	200000	1e-4	4794	11802	40.6
5000	250000	1e-4	6236	15719	39.7
7500	250000	1e-4	6209	15662	39.6
10000	250000	1e-4	6206	15586	39.8
15000	250000	1e-4	6172	15418	40.0
20000	250000	1e-4	6167	15371	40.1
25000	250000	1e-4	6104	15233	40.1
30000	250000	1e-4	6109	15177	40.3
35000	250000	1e-4	6116	15153	40.4
40000	250000	1e-4	6085	15038	40.5
5000	0	1e-2 Low	18	1055	1.7
7500	0	1e-2 Low	22	1190	1.8
10000	0	1e-2 Low	25	1256	2.0
15000	0	1e-2 Low	27	1443	1.9
20000	0	1e-2 Low	58	1708	3.4
25000	0	1e-2 Low	69	1954	3.5
30000	0	1e-2 Low	81	2153	3.8
35000	0	1e-2 Low	105	2305	4.6
40000	0	1e-2 Low	115	2548	4.5
5000	5000	1e-2 Low	16	1132	1.4
7500	5000	1e-2 Low	15	1193	1.3
10000	5000	1e-2 Low	14	1230	1.1
15000	5000	1e-2 Low	16	1368	1.2
20000	5000	1e-2 Low	38	1546	2.5
25000	5000	1e-2 Low	57	1744	3.3

Table C2 (cont'd)

30000	5000	1e-2 Low	71	1911	3.7
35000	5000	1e-2 Low	81	2096	3.9
40000	5000	1e-2 Low	115	2299	5.0
5000	10000	1e-2 Low	15	1248	1.2
7500	10000	1e-2 Low	12	1293	0.9
10000	10000	1e-2 Low	15	1312	1.1
15000	10000	1e-2 Low	24	1423	1.7
20000	10000	1e-2 Low	33	1529	2.2
25000	10000	1e-2 Low	44	1694	2.6
30000	10000	1e-2 Low	61	1873	3.3
35000	10000	1e-2 Low	73	2019	3.6
40000	10000	1e-2 Low	96	2211	4.3
5000	25000	1e-2 Low	69	1892	3.6
7500	25000	1e-2 Low	73	1877	3.9
10000	25000	1e-2 Low	71	1865	3.8
15000	25000	1e-2 Low	77	1908	4.0
20000	25000	1e-2 Low	88	2004	4.4
25000	25000	1e-2 Low	89	2090	4.3
30000	25000	1e-2 Low	88	2164	4.1
35000	25000	1e-2 Low	104	2247	4.6
40000	25000	1e-2 Low	104	2401	4.3
5000	50000	1e-2 Low	191	2913	6.6
7500	50000	1e-2 Low	195	2884	6.8
10000	50000	1e-2 Low	195	2852	6.8
15000	50000	1e-2 Low	202	2835	7.1
20000	50000	1e-2 Low	200	2874	7.0
25000	50000	1e-2 Low	196	2906	6.7
30000	50000	1e-2 Low	196	2962	6.6
35000	50000	1e-2 Low	213	3002	7.1
40000	50000	1e-2 Low	218	3100	7.0
5000	75000	1e-2 Low	372	3899	9.5
7500	75000	1e-2 Low	375	3873	9.7
10000	75000	1e-2 Low	375	3856	9.7
15000	75000	1e-2 Low	367	3818	9.6
20000	75000	1e-2 Low	354	3836	9.2
25000	75000	1e-2 Low	359	3815	9.4
30000	75000	1e-2 Low	368	3856	9.5
35000	75000	1e-2 Low	368	3886	9.5
40000	75000	1e-2 Low	370	3930	9.4
5000	100000	1e-2 Low	543	4856	11.2
7500	100000	1e-2 Low	535	4827	11.1
10000	100000	1e-2 Low	536	4796	11.2
15000	100000	1e-2 Low	550	4769	11.5
20000	100000	1e-2 Low	536	4762	11.3
25000	100000	1e-2 Low	546	4768	11.5
30000	100000	1e-2 Low	541	4784	11.3
35000	100000	1e-2 Low	537	4776	11.2
40000	100000	1e-2 Low	529	4797	11.0
5000	150000	1e-2 Low	884	6516	13.6
7500	150000	1e-2 Low	874	6496	13.5
10000	150000	1e-2 Low	884	6463	13.7
15000	150000	1e-2 Low	867	6415	13.5
20000	150000	1e-2 Low	844	6389	13.2
25000	150000	1e-2 Low	859	6353	13.5
30000	150000	1e-2 Low	857	6362	13.5
35000	150000	1e-2 Low	849	6360	13.3
40000	150000	1e-2 Low	853	6368	13.4

Table C2 (cont'd)

5000	200000	1e-2 Low	1236	7909	15.6
7500	200000	1e-2 Low	1237	7862	15.7
10000	200000	1e-2 Low	1220	7840	15.6
15000	200000	1e-2 Low	1219	7794	15.6
20000	200000	1e-2 Low	1214	7759	15.6
25000	200000	1e-2 Low	1210	7718	15.7
30000	200000	1e-2 Low	1191	7720	15.4
35000	200000	1e-2 Low	1175	7692	15.3
40000	200000	1e-2 Low	1220	7716	15.8
5000	250000	1e-2 Low	1606	9268	17.3
7500	250000	1e-2 Low	1609	9235	17.4
10000	250000	1e-2 Low	1594	9203	17.3
15000	250000	1e-2 Low	1574	9181	17.1
20000	250000	1e-2 Low	1594	9128	17.5
25000	250000	1e-2 Low	1587	9113	17.4
30000	250000	1e-2 Low	1577	9077	17.4
35000	250000	1e-2 Low	1571	9061	17.3
40000	250000	1e-2 Low	1574	9078	17.3
5000	0	1e-2 High	18	1055	1.7
7500	0	1e-2 High	15	1188	1.3
10000	0	1e-2 High	18	1255	1.4
15000	0	1e-2 High	37	1443	2.6
20000	0	1e-2 High	50	1706	2.9
25000	0	1e-2 High	62	1953	3.2
30000	0	1e-2 High	91	2153	4.2
35000	0	1e-2 High	97	2305	4.2
40000	0	1e-2 High	131	2548	5.1
5000	5000	1e-2 High	9	1132	0.8
7500	5000	1e-2 High	14	1194	1.2
10000	5000	1e-2 High	22	1230	1.8
15000	5000	1e-2 High	36	1368	2.6
20000	5000	1e-2 High	46	1545	3.0
25000	5000	1e-2 High	59	1743	3.4
30000	5000	1e-2 High	66	1912	3.5
35000	5000	1e-2 High	91	2094	4.3
40000	5000	1e-2 High	118	2299	5.1
5000	10000	1e-2 High	27	1248	2.2
7500	10000	1e-2 High	27	1292	2.1
10000	10000	1e-2 High	34	1313	2.6
15000	10000	1e-2 High	45	1423	3.2
20000	10000	1e-2 High	57	1529	3.7
25000	10000	1e-2 High	73	1700	4.3
30000	10000	1e-2 High	87	1873	4.6
35000	10000	1e-2 High	91	2019	4.5
40000	10000	1e-2 High	118	2210	5.3
5000	25000	1e-2 High	116	1892	6.1
7500	25000	1e-2 High	123	1875	6.6
10000	25000	1e-2 High	117	1865	6.3
15000	25000	1e-2 High	123	1905	6.5
20000	25000	1e-2 High	131	2004	6.5
25000	25000	1e-2 High	143	2090	6.8
30000	25000	1e-2 High	151	2164	7.0
35000	25000	1e-2 High	157	2247	7.0
40000	25000	1e-2 High	163	2402	6.8
5000	50000	1e-2 High	252	2917	8.6
7500	50000	1e-2 High	254	2883	8.8
10000	50000	1e-2 High	268	2852	9.4

Table C2 (cont'd)

15000	50000	1e-2 High	262	2833	9.2
20000	50000	1e-2 High	266	2875	9.3
25000	50000	1e-2 High	265	2905	9.1
30000	50000	1e-2 High	262	2964	8.8
35000	50000	1e-2 High	269	3001	9.0
40000	50000	1e-2 High	281	3100	9.1
5000	75000	1e-2 High	405	3899	10.4
7500	75000	1e-2 High	415	3872	10.7
10000	75000	1e-2 High	414	3856	10.7
15000	75000	1e-2 High	403	3819	10.6
20000	75000	1e-2 High	410	3836	10.7
25000	75000	1e-2 High	409	3815	10.7
30000	75000	1e-2 High	396	3856	10.3
35000	75000	1e-2 High	404	3885	10.4
40000	75000	1e-2 High	420	3930	10.7
5000	100000	1e-2 High	593	4856	12.2
7500	100000	1e-2 High	593	4828	12.3
10000	100000	1e-2 High	583	4797	12.2
15000	100000	1e-2 High	566	4768	11.9
20000	100000	1e-2 High	587	4759	12.3
25000	100000	1e-2 High	576	4768	12.1
30000	100000	1e-2 High	575	4784	12.0
35000	100000	1e-2 High	569	4779	11.9
40000	100000	1e-2 High	566	4797	11.8
5000	150000	1e-2 High	913	6516	14.0
7500	150000	1e-2 High	917	6499	14.1
10000	150000	1e-2 High	901	6463	13.9
15000	150000	1e-2 High	895	6414	14.0
20000	150000	1e-2 High	897	6391	14.0
25000	150000	1e-2 High	870	6354	13.7
30000	150000	1e-2 High	861	6362	13.5
35000	150000	1e-2 High	861	6360	13.5
40000	150000	1e-2 High	868	6369	13.6
5000	200000	1e-2 High	1204	7902	15.2
7500	200000	1e-2 High	1187	7861	15.1
10000	200000	1e-2 High	1206	7838	15.4
15000	200000	1e-2 High	1177	7794	15.1
20000	200000	1e-2 High	1186	7760	15.3
25000	200000	1e-2 High	1187	7716	15.4
30000	200000	1e-2 High	1181	7719	15.3
35000	200000	1e-2 High	1190	7690	15.5
40000	200000	1e-2 High	1185	7714	15.4
5000	250000	1e-2 High	1485	9272	16.0
7500	250000	1e-2 High	1504	9237	16.3
10000	250000	1e-2 High	1494	9208	16.2
15000	250000	1e-2 High	1474	9179	16.1
20000	250000	1e-2 High	1476	9126	16.2
25000	250000	1e-2 High	1454	9104	16.0
30000	250000	1e-2 High	1462	9075	16.1
35000	250000	1e-2 High	1470	9074	16.2
40000	250000	1e-2 High	1451	9079	16.0

CONCLUSION

In this dissertation, I discuss three projects that address challenges in the analysis and prediction of high-dimensional genetic data, focusing on utilizing local genetic information and improving the accuracy of models across underrepresented ancestry groups. One approach in genomics is to do a genome-wide analysis. However, genome-wide analyses can have challenges, such as heavy computational burdens or interpretation difficulties. If instead the analysis is evaluated locally, some of these challenges can be overcome.

Both Chapters 1 and 2 present approaches that leverage local information, such as linkage disequilibrium. Chapter 1 estimates local genetic covariances within local segments in linkage disequilibrium, identifying segments with opposing directionality to the overall genetic correlation that would typically be masked in genome-wide correlation analyses. In the context of cross-ancestry prediction, the second study develops the MC-ANOVA method, which estimated the loss of prediction accuracy due to (local) differences in linkage disequilibrium and allele frequencies between ancestry groups. The study highlights the significant variability in prediction accuracy across local SNP segments, identifying some segments that are portable in PGS across ancestry groups and other segments that are not portable.

Chapter 2 highlighted limitations in the non-European data available, and the importance of continuing ongoing efforts to collect non-European genomic data. Thus, Chapter 3 explores the impact of sample size on cross-ancestry prediction accuracy by meta-analyzing data from the UK Biobank and All of Us to investigate how varying European and African ancestry training sample sizes affect prediction in African ancestry. The findings further demonstrate the importance of cross-ancestry sample sizes in improving prediction accuracy among underrepresented ancestry groups and emphasize the need for increased sample sizes of non-Europeans.

Collectively, these projects contribute important methodological advancements and computational tools in the field of statistical genetics, particularly in the context of leveraging local genetic and ancestry information.