Three Open Questions in Polygenic Score Portability

Joyce Y. Wang[1], Neeka Lin[1], Michael Zietz[2], Jason Mares[3], Paul J. Rathouz[4,5] and Arbel Harpak[1,5,+]

[1] Department of Integrative Biology, The University of Texas at Austin, Austin, TX

[2] Department of Biomedical Informatics, Columbia University, New York, NY

[3] Department of Neurology, Columbia University, New York, NY

[4] Department of Statistics and Data Science, The University of Texas at Austin, Austin, TX

[5] Department of Population Health, The University of Texas at Austin, Austin, TX

[+] Correspondence should be addressed to A.H. (arbelharpak@utexas.edu)

Abstract

A major obstacle hindering the broad adoption of polygenic scores (PGS) is their lack of "portability" to people that differ—in genetic ancestry or other characteristics—from the GWAS samples in which genetic effects were estimated. Here, we use the UK Biobank to measure the change in PGS prediction accuracy as a continuous function of individuals' genome-wide genetic dissimilarity to the GWAS sample ("genetic distance"). Our results highlight three gaps in our understanding of PGS portability. First, prediction accuracy is extremely noisy at the individual level and not well predicted by genetic distance. In fact, variance in prediction accuracy is explained comparably well by socioeconomic measures. Second, trends of portability vary across traits. For several immunity-related traits, prediction accuracy drops near zero quickly even at intermediate levels of genetic distance. This quick drop may reflect GWAS associations being more ancestry-specific in immunity-related traits than in other traits. Third, we show that even qualitative trends of portability can depend on the measure of prediction accuracy used. For instance, for type 2 diabetes, precision stays roughly constant, while recall surprisingly increases, with genetic distance. Together, our results show that portability cannot be understood through global ancestry groupings alone. There are other, understudied factors influencing portability, such as the specifics of the evolution of the trait and its genetic architecture, social context, and the construction of the polygenic score. Addressing these gaps can aid in the development and application of PGS and inform more equitable genomic research.

Introduction

Polygenic scores (PGS), genetic predictors of complex traits based on genome-wide association studies (GWAS), are gaining traction among researchers and practitioners[15,17,30]. Yet a major problem hindering their broad application is their highly variable performance across prediction samples[23,7,12,16]. Often, prediction accuracy appears to decline in groups unlike the GWAS sample—in genetic ancestry, social context or environmental exposures[19,23,12,38,49,36], restricting the contexts in which PGS can be used reliably.

This so-called "portability" problem is a subject of intense study. Typically, portability is evaluated through variation in the within-group phenotypic variance explained by a PGS (i.e., the coefficient of determination, R^2) among genetic ancestry groups. Indeed, population genetics theory gives clear predictions for the relationship between genetic dissimilarity to the GWAS sample and PGS prediction accuracy under some models (neutral evolution[33,52,4], directional[28], or stabilizing selection[52,28]), all else being equal (including, e.g., assumptions about environmental effects).

However, inference based on empirical variation in R^2 can be misleading for various reasons. For one, R^2 can be arbitrarily low even when the model fitted to the data is correct. It also cannot be compared across transformations of the data. R^2 is not comparable across datasets, because, for instance, it depends on the extent of variation in the independent variable[46,18,40]. In the context of inference about the causes of PGS portability, these issues can manifest in different ways[22]. For example, heterogeneity in within-group genetic variance and environmental variance can each greatly affect group differences in R^2.

A related issue is that the impacts of environmental and social factors on portability are not well understood, despite evidence illustrating these impacts can be substantial[23,12,50,25,22]. To complicate matters, such factors may be confounded with genetic ancestry, limiting our ability to make inferences based on the typical decay of R^2 between PGS and trait value in ancestries less represented in GWAS samples[23,30,12,50,42].

With these limitations of R^2, and the possible confounding with environmental and so-

cial factors, it remains unclear how well genetic ancestry would predict the applicability of PGS for individuals. Recent work implied that individual-level prediction accuracy should be largely explained by genome-wide genetic dissimilarity to the GWAS sample (see figure 3 in [8] and figure 5 in [45]). However, we note that this work focused on the relationship between genetic distance and the length of the prediction interval, i.e. expected uncertainty in prediction under an assumed model, rather than the relationship with the realized prediction accuracy. Understanding the drivers of variation in prediction accuracy is especially pertinent for personalized clinical risk predictions and decisions regarding their reporting to patients[17,14].

This motivated us to empirically study PGS prediction accuracy at the individual level. In what follows, we highlight three puzzling observations that also point to three gaps in our understanding of the portability problem: (1) Genetic dissimilarity to the GWAS sample poorly predicts portability at the individual level, (2) portability trends (with respect to genetic distance) can be trait-specific; and (3) portability trends depend on the measure of prediction accuracy. Informed by our results, we suggest avenues of future research that can help bridge these gaps.

Results

Portability and individual-level genetic distance from the GWAS sample. We examined PGS portability as a function of genetic distance from the GWAS sample in the UK Biobank (UKB). For each of 15 continuous physiological traits, we performed a GWAS in a sample of 336,923 individuals. For 69,500 individuals not included in the GWAS sample (henceforth referred to as "prediction sample"), we predicted the trait value using the PGS and covariates. Using a Principal Component Analysis (PCA) of the genotype matrix of the entire sample, we quantify each individual's genetic distance from the GWAS sample as distance from the centroid of GWAS individuals' coordinates in PCA space (**Fig. 1A**). This measure is quicker to compute, yet highly correlated with F_{st} between the GWAS sample and single individuals in the prediction sample ($r > 0.98$), albeit noticeably less reflective of F_{st}

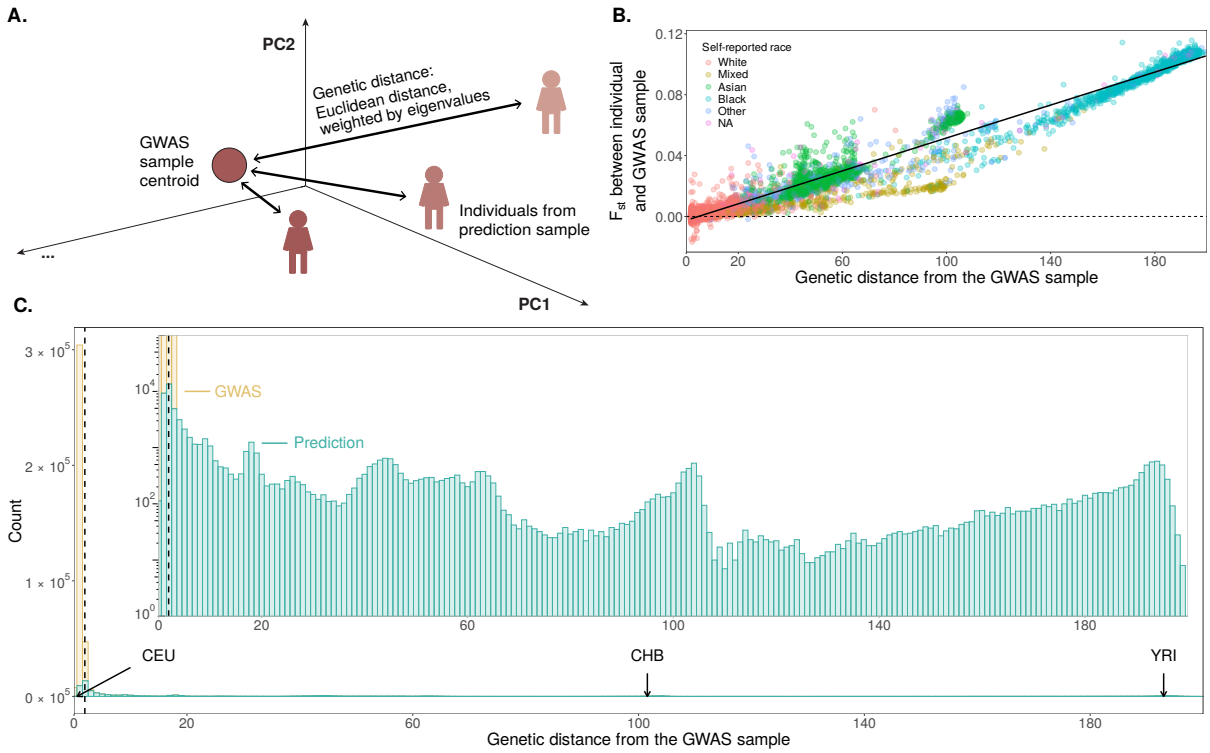

Figure 1: Measuring "genetic distance" from the GWAS sample. **A.** Across 336,923 individuals in the GWAS sample and 69,500 individuals in the prediction set, we measure "genetic distance" from the GWAS sample as the weighted Euclidean distance from the centroid of GWAS individuals in PCA space, with each PC weight being proportional to its respective eigenvalue. **B.** Across 10,000 individuals from the prediction set, genetic distance to the GWAS sample (calculated with 40 PCs) is highly correlated with F_{st} between the GWAS sample and the individual (**Fig. S1**). Under a theoretical model where portability is driven by genetic ancestry alone and the trait evolves neutrally, F_{st} should perfectly predict variation in prediction accuracy. We note that genetic distance is less reflective of F_{st} for intermediate genetic distances. **C.** The distribution of genetic distance. For reference, we show the mean genetic distances for subsets of the 1000 Genomes dataset[3]: CEU, Utah residents of primarily Northern and Western European descent; CHB, Han Chinese in Beijing, China; YRI, Yoruba in Ibadan, Nigeria. The dashed line represents the 97.5th percentile of genetic distance from among GWAS sample individuals. In what follows, our reports are based on individuals with genetic distances larger than this value. The inset is a zoomed-in view of a smaller range and on a log-scale, to better visualize the distribution within the prediction sample.

at intermediate genetic distances (**Fig. 1B**). The imperfect correlation may be a result of our use of only the top 40 PCs[29,33]. Under some theoretical conditions (such as neutral evolution, additive contribution of genotype and environment, fixed environmental variance)—F_{st} should perfectly predict variation in prediction accuracy due to genetic ancestry[31,33,4]. We standardized genetic distance such that its mean is 1 across GWAS sample individuals.

In the prediction sample, we observed a continuum of genetic distance from the GWAS

sample with several clear modes, the main one at short distances: 38,992 individuals have a genetic distance of up to 10 and the remaining 30,508 individuals at distances between 10-199.6 (**Fig. 1B, C**). To ground our expectations, we estimated the mean genetic distance for three 1000 Genomes[3] subsamples: Utah residents of primarily Northern and Western European descent (CEU) average at 0.5, Han Chinese in Beijing, China (CHB) average at 101.6, and Yoruba in Ibadan, Nigeria (YRI) average at 193.0 (**Fig. 1C**).

For each of the 15 continuous physiological traits, we measure the prediction accuracy at the group and individual level with slightly different prediction models (**Methods**). In both cases, we fit a prediction model regressing the trait to the polygenic score and other covariates. To evaluate group-level accuracy, we split individuals into 250 bins of genetic distance comprising of 278 individuals each. Within each bin we measure the partial R^2 of the polygenic score and the trait value. To evaluate individual-level accuracy, we measure the squared difference between the PGS-predicted value and the trait, after residualizing the trait for covariates.

Prediction accuracy is weakly predicted by genetic distance. For some traits, such as height, group-level prediction accuracy decayed monotonically with genetic distance from the GWAS sample, as expected and reported previously (**Fig. 2A**)[47,33,8]. A major factor driving this decay appears to be an associated decay in heterozygosity in the PGS marker SNPs (**Figs. 4B,S22**; see [28, 47]). Lower heterozygosity in PGS markers impacts the genetic variance a polygenic score can capture because it makes for a less variable predictor. The impact of genetic distance on LD with causal variation is less straightforward[47,52,2].

Previous work implied that variation in individual-level prediction accuracy should be largely explained by genetic distance[8,45]. However, that was not the case in our analysis. While individual-level accuracy generally decayed with distance for most traits, this correlation was weak (**Figs. 2B,S2**). Even a flexible cubic spline fit of genetic distance explains little of the variance in prediction accuracy ($R^2 = 0.51\%$).

In fact, individual-level prediction accuracy is explained comparably well by socioeconomic measures (**Figs. S18-S21**). For example, we observed a steady mean increase in

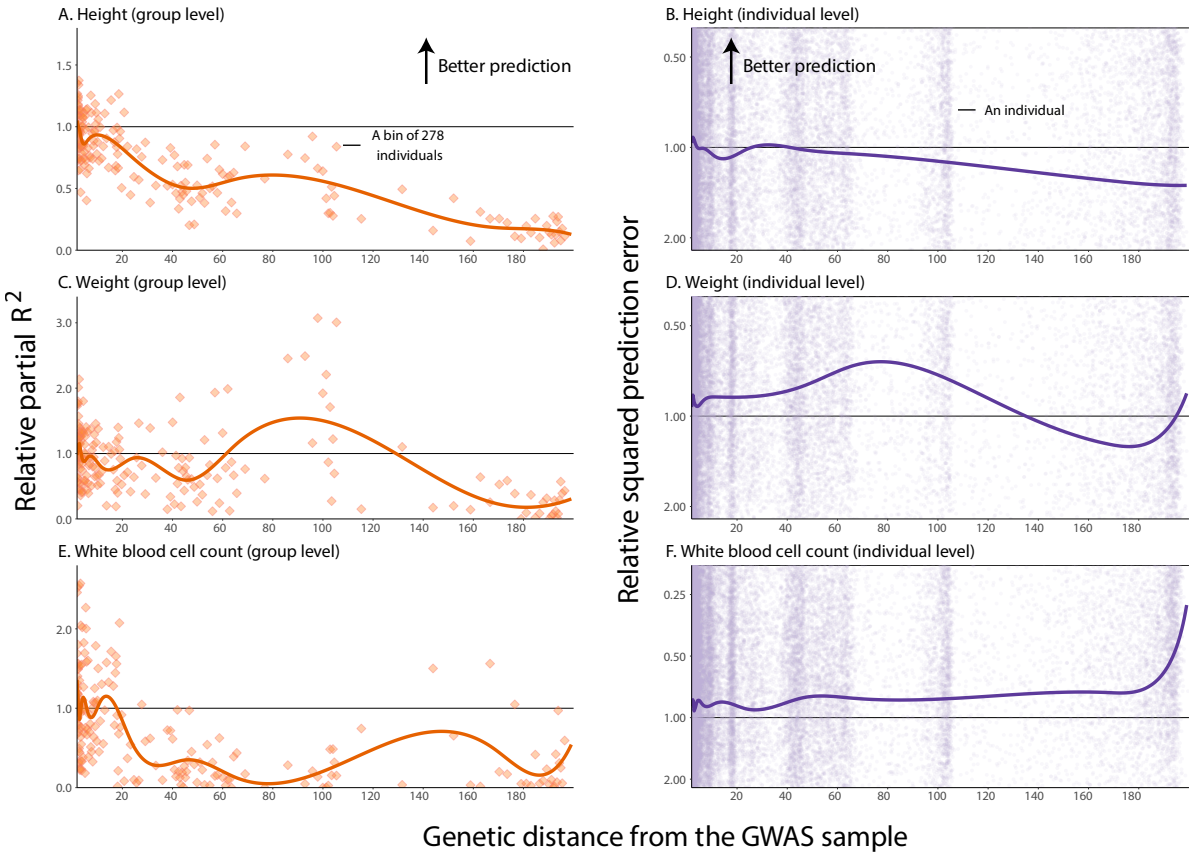

Figure 2: Trends of portability vary across traits and measures. At the group level (left panels), we measured prediction accuracy with the squared partial correlation between the PGS and the trait value in 250 bins of 278 individuals each. At the individual level (right panels), we measured prediction accuracy as the squared difference between the predicted phenotype and the true phenotype value. At both the group and individual levels, y-axis values show relative prediction accuracy, i.e. prediction accuracy divided by a baseline value. The baseline value is the mean prediction accuracy in 25 bins with median genetic distances that are closest to the mean genetic distance for GWAS individuals. **Table S1** details these baseline values for each trait. Curves show cubic spline fits, with 8 knots placed based on the density of data points. **A, B.** For height, prediction accuracy decays nearly monotonically with genetic distance at both the group (A) and individual (B) levels. **C, D.** For weight, prediction accuracy does not monotonically decay with genetic distance. **E, F.** For white blood cell count, at the group level, prediction accuracy drops near zero at a short genetic distance from the GWAS sample (E); yet at the individual level, it increases (F). See **Figs. S2-S5** for other traits and **Figs. S6-S9** for plots showing the full ranges of individual-level prediction accuracy.

118 squared prediction error across quantiles of Townsend Deprivation Index[44] for 11/15 of the
119 traits examined, suggesting poorer prediction in individuals of lower socioeconomic status
120 (**Figs. 3A,S14,S17,S16**; the four exceptions being white blood cell-related traits, **Fig. S15**;
121 see also similar reports in [23, 12, 25]). Like genetic distance, the Townsend Deprivation In-
122 dex only explains between 0.02% and 0.53% of the variance in squared prediction error

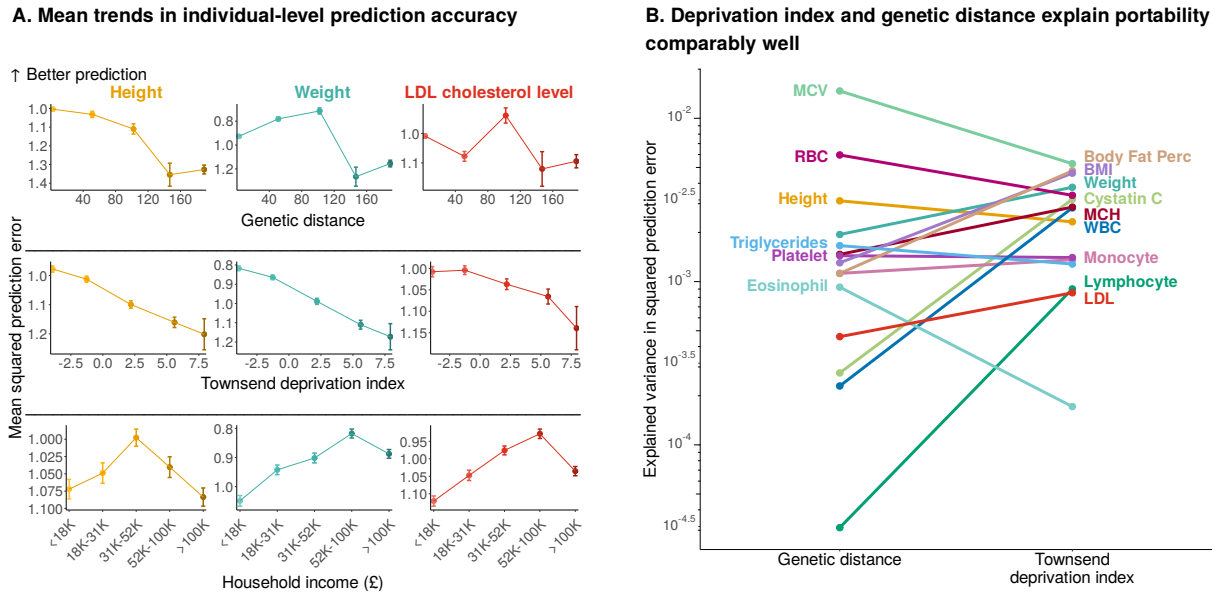

Figure 3: Genetic distance and socioeconomic factors explain individual-level prediction accuracy comparably well. **A.** Data points confer to mean (±SE) squared prediction errors of individuals in the prediction sample (divided by a constant, the mean squared prediction error in a reference group), binned into 5 equidistant strata. The x-axis shows the median measure value for each stratum. "Household income" refers to average yearly total household income before tax. See **Figs. S14-S17** for other traits. **B.** We compared the variance in squared prediction error explained by a cubic spline fit to genetic distance to the variance explained by a cubic spline fit to the Townsend deprivation index. MCV: mean corpuscular volume. MCH: mean corpuscular hemoglobin. RBC: red blood cell count. Body fat perc: body fat percentage. WBC: white blood cell count. LDL: LDL cholesterol level. See **Figs. S18-S21** for the variance explained by other genetic and socioeconomic measures.

across traits with a cubic spline. Notably, however, for the majority of traits, more variance is explained by this measure of socioeconomic status than by genetic distance (**Fig. 3B**).

Trends of portability vary across traits. One might expect a qualitatively similar, monotonic relationship between genetic distance and prediction accuracy across traits. Previous analyses (that have not examined individual level prediction accuracy) observed similarly monotonic[19], and even linear[33,8], relationship regardless of the trait examined. However, we observed variation in this relationship among traits. Unlike the case of height, the prediction accuracy for many other traits did not decay monotonically with genetic distance. Weight and body fat percentage peaked in accuracy at intermediate genetic distances (**Figs. 2D,S2**).

In other traits we examined, in particular white blood cell-related traits, group-level

prediction accuracy dropped near zero even at a short genetic distance (**Figs. 2E,S3**). There are multiple possible drivers of trait-specific portability trends. We considered, in particular, variable selective pressures on the immune system across time and geography. We hypothesized that these would lead to less portable genetic associations (across ancestry) compared to other traits. To test this prediction, we re-estimated the effects of index SNPs (SNPs included in the PGS, ascertained in the original GWAS sample) in two subsets of the prediction sample, one closer and another farther (in terms of genetic distance) from the GWAS sample. The prediction sample based allelic effect estimates were least consistent with the original GWAS for lymphocyte count, compared, e.g., to triglyceride levels, a trait of similar SNP heritability (**Fig. 4A**). To further illustrate this point, 31.7% of index SNPs for lymphocyte count had a different sign when estimated in the original GWAS and in the "closer" GWAS, compared to 9.6% for triglyceride levels.

The rapid turnover of allelic effects may also interact with statistical biases. Consider, for example, "winners curse", whereby effect estimates are inflated due to the ascertainment of index SNPs and the estimation of their effects in the same sample[20]. Winners curse would be most severe in large effect PGS index SNPs: These SNPs are typically at lower frequencies in the GWAS sample than small effect index SNPs, because GWAS power scales with the product of squared allelic effect and heterozygosity[41,27,28]. If causal effects on lymphocyte count change rapidly, then large effect index SNPs may be under weaker selective constraint in the prediction sample than in the GWAS sample, and segregate at high allele frequencies. Indeed, for lymphocyte count, the heterozygosity of large effect variants increases with genetic distance from the GWAS sample (**Fig. 4B**; see **Fig. S22** for other traits). As a result of the trends of heterozygosity, the variance in the polygenic score (a sum over index SNP heterozygosity multiplied by their squared effect estimates) quickly increases with genetic distance for white blood cell count, lymphocyte count, and monocyte count, despite decreasing for all other traits we have examined (**Figs. 4C, S66**). And so, taken together, the PGS variance increases quickly and allelic effect estimates become non-predictive even close to the GWAS sample (**Fig. S24**). Together, this may drive the immediate drop in prediction

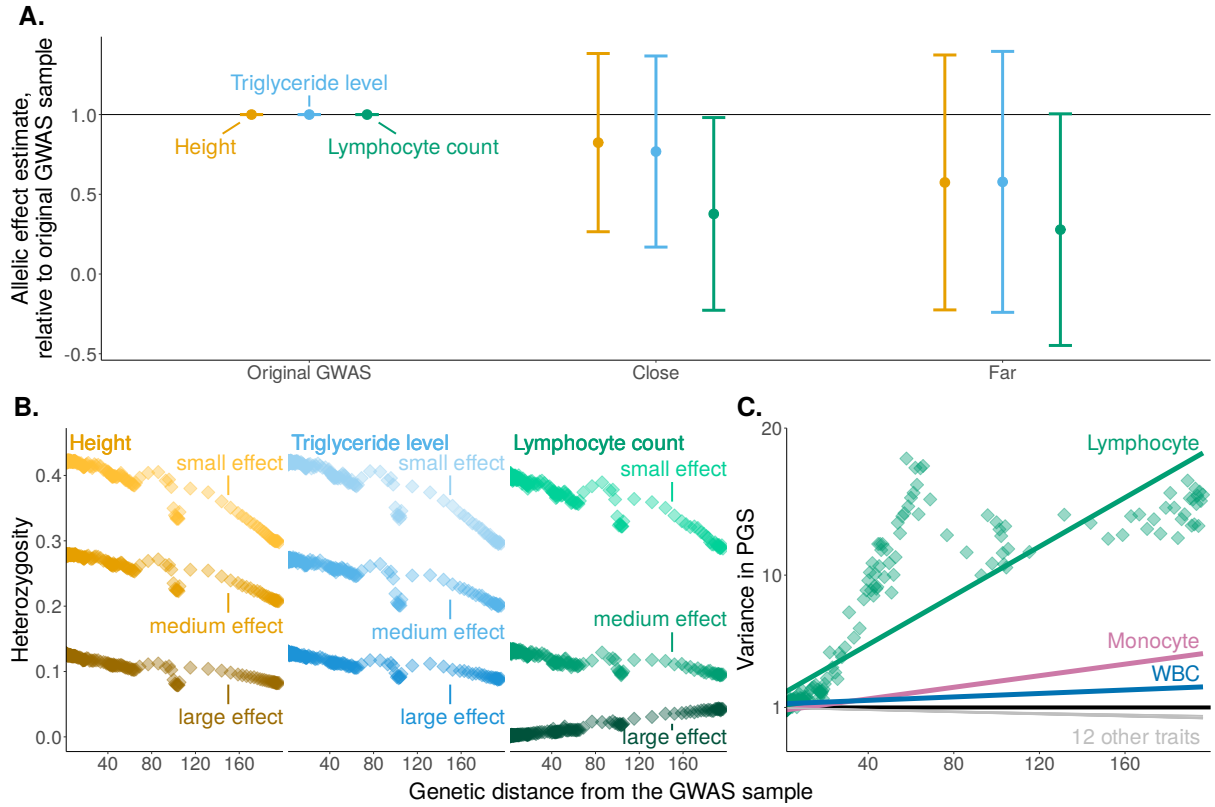

Figure 4: Lymphocyte count as an example of trait-dependent factors influencing portability. **A.** We re-estimated the allelic effects of PGS index SNPs in subsamples of the prediction set: "close" (genetic distance ≤ 10, with 38,992 individuals), and "far" (genetic distance > 10, with 30,508 individuals). For each index SNP of each PGS, we computed the allelic effect estimate relative to the effect estimate in the original GWAS sample. Shown are means $\pm$ standard deviations across PGS index SNPs for three traits, highlighting the poorer agreement between allelic effect estimates for lymphocyte count. **B.** We compared the mean heterozygosity of index SNPs for height, triglycerides, and lymphocyte count. For each trait, SNPs are stratified into three equally-sized bins of squared allelic effect estimate (**Fig. S23**). Each data point is the mean heterozygosity of a stratum in a bin of genetic distance. Unlike other traits, the heterozygosity of large effect variants for lymphocyte count increases with genetic distance from the GWAS sample. See **Fig. S22** for other traits. **C.** We compared the variance of PGS, in each bin, relative to the variance of PGS in the reference group, across traits. Among the 15 traits we have examined, only for lymphocyte count, monocyte count, and white blood cell count (WBC) the PGS variance increased with genetic distance. Green points show the PGS variance for lymphocyte count in genetic ancestry bins. Lines show the ordinary least squares linear fit to the respective bin-level data for each trait.

accuracy of white blood-cell related traits.

The measure of predictive performance can alter our view of portability. Finally, the trends of portability, even qualitative trends, can depend on the measure of prediction accuracy. For triglyceride levels, lymphocyte count, and white blood cell count, group-level prediction accuracy is near zero far from the GWAS sample (**Figs. 2E,S3**) whereas at

the individual level, prediction accuracy increases (**Figs. 2F,S3,S11,S13**).

Trait-specific, measure-specific portability trends are of specific importance for diseases, since the choice of performance metrics often ties in with an intervention or policy that is being considered. For example, when a broadly accessible, safe intervention exists, we might prioritize the identification of all cases over all else and therefore focus on recall (number of true positives divided by number of true cases). In other cases, for example when the intervention includes a risky treatment, we might prioritize avoiding false positives and hence focus on precision (true positives over all positive classifications). Therefore, we sought to know whether trends of portability depend on the measures of predictive performance in disease.

To address this question, we analyzed precision and recall in two diseases with over 5,000 cases in the prediction sample. For asthma, precision and accuracy depended on genetic distance in a qualitatively similar way (**Fig. 5A**). In contrast, for type 2 diabetes, precision appeared roughly constant for medium and large genetic distances, while recall generally increased with distance far from the GWAS sample (**Fig. 5B**). In conclusion, predictive performance showed trait-specific, measure-specific trends in disease risk prediction as well.

Discussion

Through an examination of empirical trends of portability at the individual level, we highlighted three gaps in our current understanding of the portability problem. Below, we discuss possible avenues towards filling these gaps.

The driver of portability that has been extensively discussed in the literature is ancestral similarity to the GWAS sample[19,33,47,4,17,8]. Yet our results show that, at the individual level, prediction accuracy is poorly predicted by genome-wide genetic ancestry. We note that our measure of genetic distance (also similar to that used in other studies[33,8,12,11]) is plausibly sub-optimal, as suggested, for example, by the noisiness of its relationship with F_{st} at intermediate genetic distances (**Fig. 2B**). Therefore, one path forward is to ask whether refined measures of genetic distance from the GWAS sample, in particular ones that capture

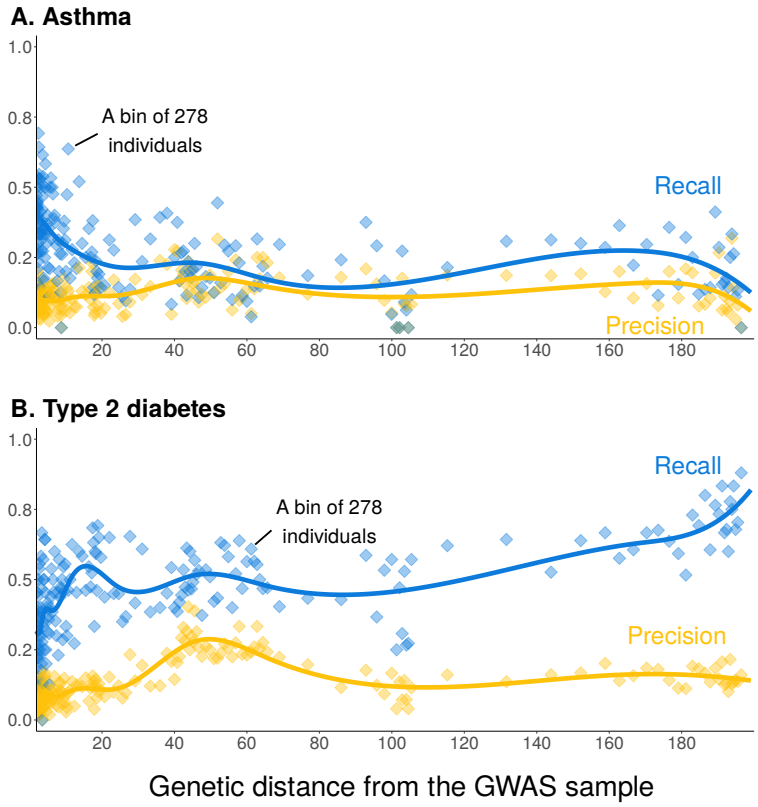

Figure 5: Trends of portability vary across medical conditions and measures of predictive performance. Individuals in the prediction sample were stratified into 250 bins by genetic distance. Each data point confers to a bin of 278 individuals. We calculated the precision (number of correctly-predicted cases divided by the number of predicted cases) and recall (number of correctly-predicted cases divided by the number of cases). (**A**) Type 2 diabetes, with 5,416 cases in the prediction sample. (**B**) Asthma, with 5,111 cases in the prediction sample. Curves show cubic spline fits, with 8 knots placed based on the density of data points. Note that in this figure, unlike Fig. 2, raw measures of prediction accuracy (precision and recall) are shown on the y-axis, rather than relative ones.

local ancestry[13,37] (e.g., in the genomic regions containing the PGS index SNPs), better explain portability. Another direction is in quantifying how environmental and social context, such as access to healthcare, affect portability (See [12] for a recent method in this vein). The relative importance of these factors will also inform the efforts to diversify participation in GWAS.

Second, we observed some trait-specific trends in portability, and hypothesize that they reflect the specifics of natural selection and evolutionary history of genetic variants affecting the trait. While previous work considered the impact of directional[4,9,6,28] and stabilizing

selection[52,47,28,2] on portability, the trait-specific (and PGS-specific) impact—notably for disease prediction—is yet to be studied empirically. Here, we showed empirical support for the hypothesis that heterogeneous selective pressures on the immune system across time and geography may drive unique portability trends in immune-related polygenic scores. Further research is needed to test this hypothesis and its generalizability. More generally, evolutionary perspectives on genetic architectures and other facets of GWAS data have been transformative[39,24]. This may also prove to be the case for understanding PGS portability.

Third, we show that individual-level measures, which are arguably the most relevant to eventual applications of PGS, can yield different results to group level measures that are widely used. PGS research has been focused on coefficient of determination (R^2) analyzed at the group level[19,23,47,8,33,52,4]. More generally, different applications and questions call for different measures of prediction accuracy, for instance when considering the utility of a public health intervention applied to communities, as opposed to asking about the cost-effectiveness of an expensive drug for an individual patient (see [1] for related discussion). Therefore, future research on predictive performance could benefit from more focus on the metrics most relevant to the intended application.

Addressing these gaps in our understanding of PGS portability will be key for evaluating the utility of a PGS, and for its equitable application in the clinic and beyond.

Acknowledgements

We thank Ipsita Agarwal, Maryn Carlson, Yi Ding, Doc Edge, Kangcheng Hou, Hakhamanesh Mostafavi, Bogdan Pasaniuc, Molly Przeworski, Sam Smith, Jeff Spence and members of the Harpak Lab for helpful feedback. This work was funded by NIH grant R35GM151108, a fellowship from the Simons Foundation's Society of Fellows (#633313) and a Pew Scholarship to A.H. This study was conducted using the UK Biobank resource under application 61666, as approved by the University of Texas at Austin institutional review board (protocol 2019-02-0125). We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed

to the research results reported within this paper.

Methods

Data

Data overview. All analyses were conducted with data from the UK Biobank, a large-scale biomedical database with a sample size of 502,490 individuals[43]. In this study, we considered 406,423 individuals who passed quality control (QC) checks, which included the removal of 359 individuals who withdrew from the study as of December 17, 2024, 651 samples identified by the UK Biobank as having sex chromosome aneuploidy (data field 22019), and an additional 14,418 individuals whose self-reported biological sex (data field 31) differed from sex determined from that implied by their sex chromosome karyotype (data field 22001). We removed 963 individuals who are outliers in heterozygosity or genotype missingness (data field 22027) and 6,851 individuals with genotype missingness greater than 2% (data field 22005). We the, removed 72,825 individuals with 3rd-degree relatives or closer (data field 22020). In total, we removed 96,067 individuals. In the selection of the GWAS sample, we first classified all the individuals who passed the filters into "White British" (WB) and "Non-White British" (NWB). WB individuals are those who self-identified as "White" and "British" and closely clustered together in PC space (336,923 individuals, data field 22006). All other individuals are NWB (69,500 individuals) We used all the WB individuals as the GWAS set and the NWB as the prediction set.

Genotype data. We started with 765,067 biallelic variants out of a total of 784,256 genotyped variants on the autosomes. We first removed 10,543 SNPs within the major histocompatibility complex (MHC) and extended region in strong LD with it (chromosome 6, positions 26,477,797-35,448,354 in the GRCh37 genome build). We excluded variants with a Hardy-Weinberg equilibrium p-value (`--hwe`) lower than 1×10^{-10} among White British (WB) individuals, removing another 45,930 variants. We also removed an additional 39,996 variants by setting the minor allele frequency threshold (`--maf`) among WB to $> 0.01\%$.

After filtering, we had 668,598 variants which we analyzed going forward.

Phenotype data. We analyzed 15 highly heritable traits, as determined based on Neale Lab SNP heritability estimates[26] (**Table S1**). These included both physiological measurements to biomarkers: standing height (data field 50), cystatin C level (data field 30720), platelet count (data field 30080), mean corpuscular volume (MCV, data field 30040), weight (data field 21002), mean corpuscular hemoglobin (MCH, data field 30050), body mass index (BMI, data field 21001), red blood cell count (RBC, data field 30010), body fat percentage (data field 23099), monocyte count (UKB data field 30130), triglyceride level (data field 30870), lymphocyte count (data field 30120), white blood cell count (WBC, data field 30000), eosinophil count (data field 30150), and LDL cholesterol level (data field 30780) (**Table S1**). For all analyses, we removed individuals with missing trait data.

Genetic distance calculations

The fixation index (F_{st}) is a natural metric, a single number, to measure the divergence between two sets of chromosomes and we considered using it to measure the distance between the pair of chromosomes of an individual and chromosomes in the GWAS sample. However, calculating F_{st} was computationally costly. Since previous work[33] showed it is tightly correlated with Euclidean distance in the PC space in the UKB, we used Euclidean distance as a single number proxying genetic distance from the GWAS sample. We used the pre-computed PCA provided by the UK Biobank (data field 22009). To calculate individual-level scores on each PC, we used the genotype matrix of the full post-filtering sample of individuals (data field 22009).

The genetic distance is the weighted PC distance between an individual coordinates vector in PCA subsapce of the first K PCs, x, and the centroid of M individuals $\{x^m\}_{m=1}^{M}$ in the GWAS sample, $C = \frac{\sum_{m=1}^{M} x^m}{M}$, is

$$\sqrt{\sum_{k=1}^{K} w_k \left(x_k - c_k\right)^2}$$

with weights

$$w_k = \frac{\lambda_k}{\sum_{n=1}^{40} \lambda_n},$$

where λ_k is the k'th eigenvalue.

To identify K, the number of PCs we used and to confirm the approximation is reasonable for our data, we examined the correlation of genetic distance with F_{st} as a function of K on a small subset of the prediction sample.

We randomly selected 10,000 prediction sample individuals with a weighted PC distance greater than the weighted PC distance of 95% of the GWAS set (based on weighted PC distance calculated from the $K = 10$). For those individuals, we estimated their F_{st} and weighted PC distance to the GWAS centroid for $K \in 1, ..., 40$ (UKB provides pre-computed individual-level coordinates for each of 40 PCs). We estimated F_{st} in this subsample with the Weir and Cockerham method[48] using the `--fst` flag in *PLINK 1.9*[34,5]

Since the PC distance calculated from using $K = 40$ correlated most strongly with F_{st} ($r > 0.98$) (**Fig. S1**), we used this number of PC to estimate the genetic distance for all test individuals (**Fig. 1B-C**). We note that genetic distance is less reflective of F_{st} for intermediate genetic distances (**Fig. 1B**).

We divided the raw genetic distances by the (raw) mean genetic distance among GWAS sample individuals. To gain intuition about these standardized units of genetic distance, we wished to estimate where on this scale we would find individuals from three subsamples from the 1000 Genomes Phase 3 dataset[3]: CEU, Utah residents (CEPH) from primarily Northern and Western European descent; CHB, Han Chinese in Beijing, China; and YRI, Yoruba in Ibadan, Nigeria. To this end, we ran a PCA with a dataset that includes both the UKB individuals and the CEU, CHB, and YRI individuals. We identified the UKB individuals with the shortest weighted Euclidean distance to the centroid of each of the three 1000 Genomes populations, and used the genetic distance of those three UKB individuals in our PCA of only UKB individuals as a proxy of where the three 1000 Genomes subsamples fall on the scale of our genetic distance measurement (**Fig. 1C**).

The distribution of genetic distance is heavily right-skewed, with most individuals falling

close to the GWAS centroid. Since we wanted to focus on the individuals far away from the GWAS set, we only analyzed data for individuals with a genetic distance greater than the 97.5^{th} percentile of genetic distance from among GWAS sample individuals (**Fig. 1C**), with the exception of the analyses behind **Fig. 3** and **Figs. S14-S21**.

For group level analyses, we binned the prediction samples by genetic distance using 250 equally-sized bins, with 278 individuals per bin.

PGS and evaluating PGS prediction accuracy

GWAS. For each trait, we used the `--glm` flag from *PLINK 2.0*[35,5] to run GWAS on the GWAS set. We used the following covariates: the first 20 PCs from UKB (data field 22009, age (data field 21022), age^2, sex (data field 31), age*sex, and age^2*sex, where the asterisk (*) denotes the product of two variables, referring to an interaction term.

Using *PLINK 1.9* with the flag `--glm` along with `--1`, we also ran a GWAS for 2 binary disease traits: type 2 diabetes (ICD-10 code E11) and asthma, using the same covariates as the ones used for continuous traits. We marked an individual as positive for a disease if they are either positive for main ICD-10 (UKB field 41202) or secondary ICD-10 (UKB field 41204).

For both continuous and disease traits, we clumped SNPs using the `--clump` flag in *PLINK 1.9*[34,5], setting the association p-value threshold for clumping (`--clump-p1`) to 0.01, LD r^2 threshold (`--clump-r2`) to 0.2, and window size (`--clump-kb`) to 250 kb.

PGS construction. After clumping and thresholding the SNPs with marginal association $p < 1 \times 10^{-5}$, we calculated PGS for each individual for every phenotype. The calculations were carried out with the `--score` flag in *PLINK 2.0*[35,5].

PGS prediction accuracy at the group level. To evaluate prediction accuracy for continuous trait PGS at the group level, we linearly regressed the phenotype on the covariates (array type (data field 22000), age, age^2, sex, age*sex, and age^2*sex) and PGS within each genetic distance bin (phenotype $\sim$ covariates + PGS), which is the full model. We then performed another linear regression of the phenotype on the covariates, excluding the PGS,

within each bin (phenotype $\sim$ covariates), which is the reduced model. Using these two squared correlations, we calculated partial R^2 for the PGS with the sum of squared errors (SSE) of these two models as

$$R^2_{\text{partial}} = \frac{SSE\,(\text{reduced_model}) - SSE\,(\text{full_model})}{SSE\,(\text{reduced_model})},$$

which represents the prediction accuracy of PGS for each bin.

As a baseline prediction accuracy, we identified the 25 bins (of 278 individuals each) with the median genetic distance most similar to the mean genetic distance for GWAS individuals; This reference group represents individuals from the prediction set that are closest to "typical" GWAS individuals in terms of genetic distance. The mean PGS prediction accuracy across these 25 bins served as the baseline value. Throughout the paper, we report the prediction accuracy at the group level as a bin's squared partial correlation between the PGS and the trait divided by this baseline value.

For binary traits, we calculated the precision (fraction of true cases out of predicted cases) and the recall (fraction of predicted cases out of all cases) in each genetic distance bin. To get the predicted phenotype, we used the PGS and predicted the phenotype to be positive if their PGS is in the 75th percentile of the PGS of the GWAS set for type 2 diabetes disease and 70th percentile for asthma.. These These thresholds were determined based on the percentile cutoff which maximized the F1 score in the GWAS set for each polygenic score.

Prediction error at the individual level. For the individual-level prediction error, we first derived phenotypic values adjusted for covariates Z in two steps, involving residualizing some covariates in each genetic ancestry bin independently and some covariates globally. First, we regress raw phenotype values Y independently in each bin on covariates,

$$Y \sim \text{array type} + \text{age}^2 + \text{sex} + \text{age*sex} + \text{age}^2\text{*sex}.$$

In bins in which only a single individual was genotyped with a particular array type, we did not include array type as a covariate. We then regress the residual X (where $X = Y - \hat{Y}$

and $\hat{Y}$ is the fitted value from the first step) globally on covariates,

$$X \sim \text{genetic_distance_polynomial} + \text{sex} + \text{sex} * \text{genetic_distance},$$

where genetic_distance_polynomial is a 20-degree polynomial in genetic distance. Finally, we regress the residual of this second regression, $Z = X - \hat{X}$ onto the PGS in a simple (univariate) linear regression. We refer to the squared residual of this regression,

$$\left(Z - \hat{Z}\right)^2,$$

as the unstandardized squared prediction error. Similar to the group-level analysis, we computed the mean unstandardized squared prediction error in the 25 reference bins as a baseline value (**Table S1** details the baseline values across traits). The squared prediction error, the measure of individual-level prediction accuracy we refer to throughout, is the unstandardized squared prediction error divided by the baseline value.

Spline fits. For both the individual-level and group-level analysis, **Figs. 2,S2-S9** show cubic spline fits. We fitted these splines using 8 knots. The knot positions were chosen based on the density of the individual genetic distances, such that there is an equal number of samples between any two knots. This resulted in knots at genetic distances of 2.39, 3.54, 5.99, 11.19, 23.76, 45.40, 65.49, and 167.53.

Mean trends in individual-level prediction accuracy. In the **Results** section of the main text, we discuss various individual predictors of squared prediction error. In addition to genetic distance, we considered 2 measures of individual-level prediction error: The Townsend Deprivation Index (data field 189) and average yearly total household income before tax (data field 738). For genetic distance and the Townsend Deprivation Index, we considered five uniformly-spaced bins, and computed the mean squared prediction error and the standard error of this mean (**Figs. 3A,S14-S17**). For household income, which the UK Biobank provides as categorical data conferring to ranges in British Pounds, we converted the categories into an ordinal variable coded as 1, 2, 3, 4 and 5, and computed the mean

squared prediction error, and standard error of the mean, in each. This also allowed us to use the income categories directly as measures in the regression models used for comparison of variance in prediction error explained that we discuss below.

Comparison of variance in prediction accuracy explained across measures. For this analysis, we used all the individuals in the prediction set and did not filter for the individuals with a genetic distance greater than 97.5^{th} percentile of genetic distance from among GWAS sample individuals. We compared the variance in squared prediction error explained for 8 raw measures: genetic distance, Townsend Deprivation Index (data field 189), average yearly total household income before tax (data field 738), educational attainment (data field 6138), which we converted into years of education, minor allele counts for SNPs with different with different magnitudes of effects (three equally-sized bins of small, medium, and large squared effect sizes, see **Fig. S23**), and minor allele counts of all SNPs. "Minor" here is with respect to the GWAS sample, and the count is the total sum of minor alleles across index SNPs of the magnitude category. Namely, for each measure, we independently fit three different models:

- A linear predictor, fit using Ordinary Least Squares (OLS).
- A discretized predictor, using one predicted value per each of the 5 bins where all five bins had identical widths.
- A cubic spline. 16 knots were placed based on the density of data points, such that there was an equal number of data points between each pair of consecutive knots.

After fitting the models, we calculated the R^2 values to determine the variance explained by each measure-method combination. We then computed 95% central confidence intervals for these R^2 values to assess the reliability of the estimates (**Fig. 3B,Fig. S18-S21**).

Additional analyses on lymphocyte count

Comparing allelic estimates across three GWASs. To test whether allelic effect estimates are similar across genetic ancestry, we performed two additional GWASs for each

trait in two subsets of the prediction sample: "close" group (genetic distance ≤ 10, with 96,457 individuals) and "far" (genetic distance > 10, with 32,822 individuals). For both groups, we adjusted for 20 PCs of the genotype matrix of the respective set of individuals, using the `--pca approx 20` flag in *PLINK 2.0*[35,5]. After running GWAS independently in the two groups, for each index SNP of the original PGS, we divided allelic effect estimates in the original GWAS / close / far set by the allelic effect estimate in the original GWAS. **Fig. 4A** shows the mean $\pm$ standard deviation across PGS index SNPs for each of three traits, highlighting the poorer agreement of the allelic effect estimates for lymphocyte count.

Heterozygosity at index SNPs as a function of genetic distance. For each PGS, we calculated the heterozygosity of each index SNP in each bin from allele counts using the `--freq` flag from *PLINK 1.9*[34,5]. We stratified index SNPs into three equally-sized bins based on their squared effect sizes (**Fig. S23**). **Figs. 4B,S22** show the mean heterozygosity (across stratum SNPs) for each stratum of in each genetic distance bin.

Variance of PGS as a function of genetic distance. For each phenotype, we calculated the variance of PGS in each bin relative to the mean of the variance of PGS in the 25 bins close to the GWAS set. In **Figs. 4C**,we plotted the values in each bin as well as a linear fit for lymphocyte count. For other traits, we only plotted the linear fit.

Heritability associated with each index SNP. We estimated the heritability explained by each index SNP as

$$\hat{h}^2_{index} = 2p(1-p)\hat{\beta}^2,$$

where $\hat{\beta}$ is the estimated allelic effect and p is the allele frequency. In **Fig. S24**, we compared the distribution of index SNP heritability across traits and with allelic effect estimates and heterozygosities calculated both in the original GWAS sample, the "close" prediction sample and the "far" prediction sample. For each trait, the SNPs used are also stratified into three equal-sized strata (small, medium, and large) based on their squared effect sizes, as discussed above.

Code availability

The scripts for the analyses and figures are available at `https://github.com/harpak-lab/Portability_Questions`.

References

[1] Abramowitz, S. A., Boulier, K., Keat, K., Cardone, K. M., Shivakumar, M., *et al.*, 2024. Population Performance and Individual Agreement of Coronary Artery Disease Polygenic Risk Scores. *medRxiv*, pages 2024–07.

[2] Añorve-Garibay, V., Huerta-Sanchez, E., Sohail, M., and Ortega-Del Vecchyo, D., 2025. Natural selection acting on complex traits hampers the predictive accuracy of polygenic scores in ancient samples. *American Journal of Human Genetics*.

[3] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., *et al.*, 10 2015. A global reference for human genetic variation. *Nature*, 526:68–74.

[4] Carlson, M. O., Rice, D. P., Berg, J. J., and Steinrücken, M., 5 2022. Polygenic score accuracy in ancient samples: Quantifying the effects of allelic turnover. *PLOS Genetics*, 18:e1010170.

[5] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., *et al.*, 02 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):s13742–015–0047–8.

[6] Cox, S. L., Moots, H. M., Stock, J. T., Shbat, A., Bitarello, B. D., *et al.*, 1 2022. Predicting skeletal stature using ancient DNA. *American Journal of Biological Anthropology*, 177:162–174.

[7] Ding, Y., Hou, K., Burch, K. S., Lapinska, S., Privé, F., *et al.*, 1 2022. Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nature Genetics*, 54:30–39.

[8] Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., *et al.*, 6 2023. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618:774–781.

[9] Durvasula, A. and Lohmueller, K. E., 4 2021. Negative selection on complex traits limits phenotype prediction accuracy between populations. *The American Journal of Human Genetics*, 108:620–631.

[10] Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W., 4 2019. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10:1776.

[11] Habtewold, T. D., Wijesiriwardhana, P., Biedrzycki, R. J., and Tekola-Ayele, F., 7 2024. Genetic distance and ancestry proportion modify the association between maternal genetic risk score of type 2 diabetes and fetal growth. *Human Genomics*, 18:81.

[12] Hou, K., Xu, Z., Ding, Y., Mandla, R., Shi, Z., *et al.*, 7 2024. Calibrated prediction intervals for polygenic scores across diverse contexts. *Nature Genetics*, 56:1386–1396.

[13] Hu, S., Ferreira, L. A. F., Shi, S., Hellenthal, G., Marchini, J., *et al.*, 2023. Leveraging fine-scale population structure reveals conservation in genetic effect sizes between human populations across a range of human phenotypes. *bioRxiv*.

[14] Kamiza, A. B., Toure, S. M., Vujkovic, M., Machipisa, T., Soremekun, O. S., *et al.*, 6 2022. Transferability of genetic risk scores in African populations. *Nature Medicine*, 28:1163–1166.

[15] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., *et al.*, 9 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50:1219–1224.

[16] Kullo, I. J., 8 2024. Promoting equity in polygenic risk assessment through global collaboration. *Nature Genetics*.

[17] Lewis, A. C. F., Perez, E. F., Prince, A. E. R., Flaxman, H. R., Gomez, L., *et al.*, 10 2022. Patient and provider perspectives on polygenic risk scores: implications for clinical reporting and utilization. *Genome Medicine*, 14:114.

[18] Lewontin, R. C., 5 1974. Annotation: the analysis of variance and the analysis of causes. *American journal of human genetics*, 26:400–11.

[19] Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., *et al.*, 4 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51:584–591.

[20] Martin, G. and Lenormand, T., 12 2006. The fitness effect of mutations across environments: a survey in light of fitness landscape models. *Evolution; international journal of organic evolution*, 60:2413–27.

[21] Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., *et al.*, 7 2021. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, 53:1097–1103.

[22] Mefford, J., Smullen, M., Zhang, F., Sadowski, M., Border, R., *et al.*, 2025. Beyond predictive R2: Quantile regression and non-equivalence tests reveal complex relationships of traits and polygenic scores. *The American Journal of Human Genetics*, 112(6):1363–1375.

[23] Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., *et al.*, 1 2020. Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, 9.

[24] Mostafavi, H., Spence, J. P., Naqvi, S., and Pritchard, J. K., 2023. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature genetics*, 55(11):1866–1875.

[25] Nagpal, S. and Gibson, G., 2024. Dual exposure-by-polygenic score interactions highlight disparities across social groups in the proportion needed to benefit. *medRxiv*, pages 2024–07.

[26] Neale Lab, 10. UK Biobank. URL `http://www.nealelab.is/uk-biobank`.

[27] O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., *et al.*, 9 2019. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *The American Journal of Human Genetics*, 105:456–476.

[28] Patel, R. A., Weiß, C. L., Zhu, H., Mostafavi, H., Simons, Y. B., *et al.*, 2024. Conditional frequency spectra as a tool for studying selection on complex traits in biobanks. *bioRxiv*.

[29] Peter, B. M., 6 2022. A geometric relationship of F 2 , F 3 and F 4 -statistics with principal component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377.

[30] Polygenic Risk Score Task Force of the International Common Disease Alliance, Adeyemo, A., Balaconis, M. K., Darnes, D. R., Fatumo, S., *et al.*, 11 2021. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature Medicine*, 27:1876–1884.

[31] Pritchard, J. K. and Przeworski, M., 7 2001. Linkage Disequilibrium in Humans: Models and Data. *The American Journal of Human Genetics*, 69:1–14.

[32] Privé, F., Luu, K., Blum, M. G., McGrath, J. J., and Vilhjálmsson, B. J., 2020. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, 36(16):4449–4457.

[33] Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., *et al.*, 1 2022. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, 109:12–23.

[34] Purcell, S. and Chang, C. PLINK 1.9. URL `www.cog-genomics.org/plink/1.9`.

[35] Purcell, S. and Chang, C. PLINK 2.0. URL `www.cog-genomics.org/plink/2.0`.

[36] Ragsdale, A. P., Nelson, D., Gravel, S., and Kelleher, J., 10 2020. Lessons Learned from Bugs in Models of Human History. *The American Journal of Human Genetics*, 107:583–588.

[37] Ried, T., 9 1998. Chromosome painting: a useful art. *Human Molecular Genetics*, 7:1619–1626.

[38] Saitou, M., Dahl, A., Wang, Q., and Liu, X., 2022. Allele frequency differences of causal variants have a major impact on low cross-ancestry portability of PRS. *medRxiv*.

[39] Sella, G. and Barton, N. H., 8 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics*, 20:461–493.

[40] Shalizi, C. R., 2024. Advanced Data Analysis from an Elementary Point of View. URL `www.stat.cmu.edu/~cshalizi/ADAfaEPoV`.

[41] Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G., 3 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLOS Biology*, 16:e2002985.

[42] Smith, S. P., Smith, O. S., Mostafavi, H., Peng, D., Berg, J. J., *et al.*, 2025. A Litmus Test for Confounding in Polygenic Scores. *bioRxiv*, pages 2025–02.

[43] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., *et al.*, 3 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12:e1001779.

[44] Townsend, P., 4 1987. Deprivation. *Journal of Social Policy*, 16:125–146.

[45] Tsuo, K., Shi, Z., Ge, T., Mandla, R., Hou, K., *et al.*, 2024. All of Us diversity and scale improve polygenic prediction contextually with greatest improvements for underrepresented populations. *bioRxiv*.

[46] Tukey, J. W., 2 1969. Analyzing data: Sanctification or detective work? *American Psychologist*, 24:83–91.

[47] Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., *et al.*, 7 2020. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, 11:3865.

[48] Weir, B. S. and Cockerham, C. C., 11 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38:1358.

[49] Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W. J., *et al.*, 4 2022. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics*, 54:450–458.

[50] Westerman, K. E. and Sofer, T., 4 2024. Many roads to a gene-environment interaction. *The American Journal of Human Genetics*, 111:626–635.

[51] Wood, S. N., 2024. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. URL `https://CRAN.R-project.org/package=mgcv`. R package version 1.9-1.

[52] Yair, S. and Coop, G., 6 2022. Population differentiation of polygenic score predictions under stabilizing selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377.

[53] Young, A. I., Benonisdottir, S., Przeworski, M., and Kong, A., 2019. Deconstructing the sources of genotype-phenotype associations in humans. *Science*, 365(6460):1396–1400.