High-resolution portability of 240 polygenic scores when derived and applied in the same cohort

Florian Privé,[1,*] Hugues Aschard,[2,3] Lasse Folkersen,[4] Clive Hoggart,[5] Paul F. O'Reilly,[5] and Bjarni J. Vilhjálmsson[1,6]

[1]National Centre for Register-Based Research, Aarhus University, Aarhus, 8210, Denmark.
[2]Department of Computational Biology, Institut Pasteur, Paris, 75015, France.
[3]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA.
[4]Danish National Genome Center, Copenhagen, 2300, Denmark
[5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, 10029, USA
[6]Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark.
*To whom correspondence should be addressed.

Contact: `florian.prive.21@gmail.com`

Abstract

The low portability of polygenic scores (PGS) across global populations is a major concern that must be addressed before PGS can be used for everyone in the clinic. Indeed, prediction accuracy has been shown to decay as a function of the genetic distance between the training and test cohorts. However, such cohorts differ not only in their genetic distance but also in their geographical distance and their data collection and assaying, conflating multiple factors. In this study, we examine the extent to which PGS are transferable between ancestries by deriving polygenic scores for 240 curated traits from the UK Biobank data and applying them in eight ancestry groups from the same cohort. By restricting both training and testing to the UK Biobank data, we reduce the risk of environmental and genotyping confounding from using different cohorts. We define the eight ancestry groups at a high-resolution, country-specific level, based on a simple, robust and effective method that we introduce here. We then apply two different predictive methods to derive polygenic scores for all 240 phenotypes, and show a systematic and dramatic reduction in portability of PGS trained in the inferred ancestral UK population and applied to the inferred ancestral Polish - Italian - Iranian - Indian - Chinese - Caribbean - Nigerian populations, respectively. These analyses, performed at a finer scale than the usual continental scale, demonstrate that prediction already drops off within European ancestries and reduces globally in proportion to PC distance, even when all individuals reside in the same country and are genotyped and phenotyped as part of the same cohort. Our study provides high-resolution and robust insights into the PGS portability problem and offers clues towards possible solutions.

Introduction

Ever larger genetic data is becoming more readily available. This enables researchers to derive polygenic scores (PGS), which summarize an individual's genetic component for a particular trait or disease by aggregating information from many genetic variants into a single score. In human genetics, polygenic scores are usually derived from summary statistics from a large meta-analysis of multiple Genome-Wide Association Studies (GWAS) and an ancestry-matched Linkage Disequilibrium (LD) reference panel (Choi *et al.* 2020). Polygenic scores can also be derived directly from individual-level data when available, i.e. from the genetic and phenotypic information of many individuals (De Los Campos *et al.* 2010). Individual-level data methods proved useful for deriving PGS for traits with moderately large effects, such as autoimmune diseases (Abraham *et al.* 2014; Privé *et al.* 2019). However, deriving polygenic scores using a single individual-level dataset with only moderate sample size usually results in poor prediction for most phenotypes. Fortunately, biobank datasets such as the UK Biobank now links genetic data for half a million individuals with phenotypic data for hundreds of traits and diseases (Bycroft *et al.* 2018). Thanks to the availability of these large datasets and to efficient methods recently developed to handle such data (Loh *et al.* 2018; Privé *et al.* 2019; Qian *et al.* 2020), individual-level data may be used to derive competitive PGS for hundreds of phenotypes.

A major concern about PGS is that they usually transfer poorly to other ancestries, e.g. a PGS derived from individuals of European ancestry is not likely to predict as well in individuals of African ancestry. Prediction in another ancestry has been shown to decay with genetic distance to the training population (Scutari *et al.* 2016; Wang *et al.* 2020) and with increasing proportion of admixture with a distant ancestry (Bitarello and Mathieson 2020; Cavazos and Witte 2020). This portability issue is suspected to be primarly due to differences in LD and allele frequencies between populations, and not so much about differences in effects and positions of causal variants (Shi *et al.* 2020; Wang *et al.* 2020; Cavazos and Witte 2020). Individual-level data from the UK Biobank offer an opportunity to further investigate this problem of PGS portability in a more controlled setting (Wang *et al.* 2020; Sinnott-Armstrong *et al.* 2021). Indeed, while the UKBB data contains genetic information for more than 450K British or European individuals, it also contains the same data for tens of thousands of individuals of non-British ancestry (Bycroft *et al.* 2018). Of particular interest, those individuals of diverse ancestries all live in the UK and had their genetic and phenotypic information derived in the same way as people of European ancestry. This allows to circumvent potential confounding bias that might arise in comparative analyses from independent studies, and makes the UK Biobank data very well suited for comparing and evaluating predictive performance of derived PGS in diverse ancestries and across multiple phenotypes.

To investigate portability of PGS to other ancestries, we must first define groups of different ancestries from the data. Principal Component Analysis (PCA) has been widely used to correct for population structure in association studies and has been shown to mirror geography in Europe (Price *et al.* 2006; Novembre *et al.* 2008). Due to its popularity, many methods have been developed for efficiently performing PCA (Abraham *et al.* 2017; Privé *et al.* 2018, 2020a) as well as appropriately projecting samples onto a reference PCA space (Zhang *et al.* 2020a; Privé *et al.* 2020a), making possible to perform these analyses for ever increasing datasets. Naturally, PCA has also been used for ancestry inference. However, among all studies where we

have seen PCA used for ancestry inference, there does not seem to be a consensus on what is the most appropriate method for inferring ancestry using PCA. For example, there are divergences on which distance metric to use and the number of PCs to use to compute these distances.

In this study, we examine the extent to which PGS are transferable between ancestries by deriving 240 polygenic scores from the UK biobank data and applying them in eight ancestry groups from the same cohort. We first propose a simple, robust and effective method for global ancestry inference and grouping from PCA of genetic data, and use it to define eight ancestry groups in the UK Biobank data. We then apply a computationally efficient implementation of penalized regression (Privé *et al.* 2019) to derive PGS for 240 traits using the UK Biobank genetic and phenotypic data only. As an alternative method, we also run LDpred2-auto (Privé *et al.* 2020b), for which we directly derive the summary statistics from the individual-level data available. We show a dramatically low portability of PGS from European ancestry to other ancestries. Indeed, on average, the phenotypic variance explained by the PGS is only 64.7% in India, 48.6% in China, and 18% in Nigeria compared to in individuals of UK ancestry. These results are presented at a finer scale than the usual continental level, which allows us to show that prediction already drops within Europe, e.g. for East and South Europe compared to UK. We find that this decay in variance explained by the PGS is roughly linear in the PC distance to the training population, and is remarkably consistent across most phenotypes and for both prediction methods applied. The few exceptions include traits such as hair color, tanning, and some blood measurements. We also explore using more than HapMap3 variants when fitting PGS, it proves useful when large effects are poorly tagged by HapMap3 variants, e.g. for lipoprotein(a), but not in the general case. We also explore the performance of PGS trained using a mixture of European and non-European ancestry samples, but do not observe any significant gain in prediction here.

Results

Overview of study

Here, we use the UK Biobank (UKBB) data only (Bycroft *et al.* 2018). We first infer eight ancestry groups in the UKBB. Then we use 391,124 individuals of UK ancestry to train polygenic scores (PGS) for 240 different phenotypes based on UKBB individual-level genotypes and phenotypes, and assess portability of the 240 PGS derived in the remaining individuals of diverse ancestries (Table 1). As additional analyses, we also investigate using more variants than the HapMap3 variants used in the main analyses, and train models using a mixture of multiple ancestries. To derive PGS in this study, we use two different methods, penalized regression and LDpred2-auto, and finally compare them.

Ancestry grouping

We investigate various approaches to classify individual in ancestry groups based on genome-wide genotype data. Detailed results can be found in the corresponding Supplementary Note; we recall main results here. First, we show that (squared) Euclidean distances in the PCA space of genetic data are approximately proportional to F_{ST} between populations, and therefore recommend using this simple distance. At the

Set	UK1	UK2	UK3	Poland	Italy	Iran	India	China	Caribbean	Nigeria
Training 1	367,063	24,061								
Test 1			20,000	4136	6660	1200	6331	1810	2484	3924
Training 2	367,063			4136	6660	1200	6331	1810		3924
Test 2			20,000						2484	

Table 1: In total, 437,669 unrelated individuals are used here. Main analyses use UK1 + UK2 (391,124 individuals) as training set and the other groups as test sets. Secondary analyses involve multiple ancestry training and keep only the UK3 and Caribbean groups as test sets; UK2 is removed from the training so that sample size from training 2 is the same as training 1 (391,124 individuals).

same time, we provide evidence that using only 2 PCs, or even 4 PCs, is not enough to distinguish between some less-distant populations, and recommend using all PCs visually capturing some population structure. Then, we use this PCA-based distance to infer ancestry in the UK Biobank and the POPRES datasets. We propose two solutions to do so, either relying on projection of PCs to reference populations such as the 1000 Genomes Project, or by directly using internal data. We show that these solutions are simple, robust and effective methods for inferring global ancestry and for grouping genetically homogeneous individuals. Here, we use the second solution, relying on PCs computed within the UK Biobank and individual information on the countries of birth, for inferring the eight ancestry groups presented in table 1 and represented in figure 1.

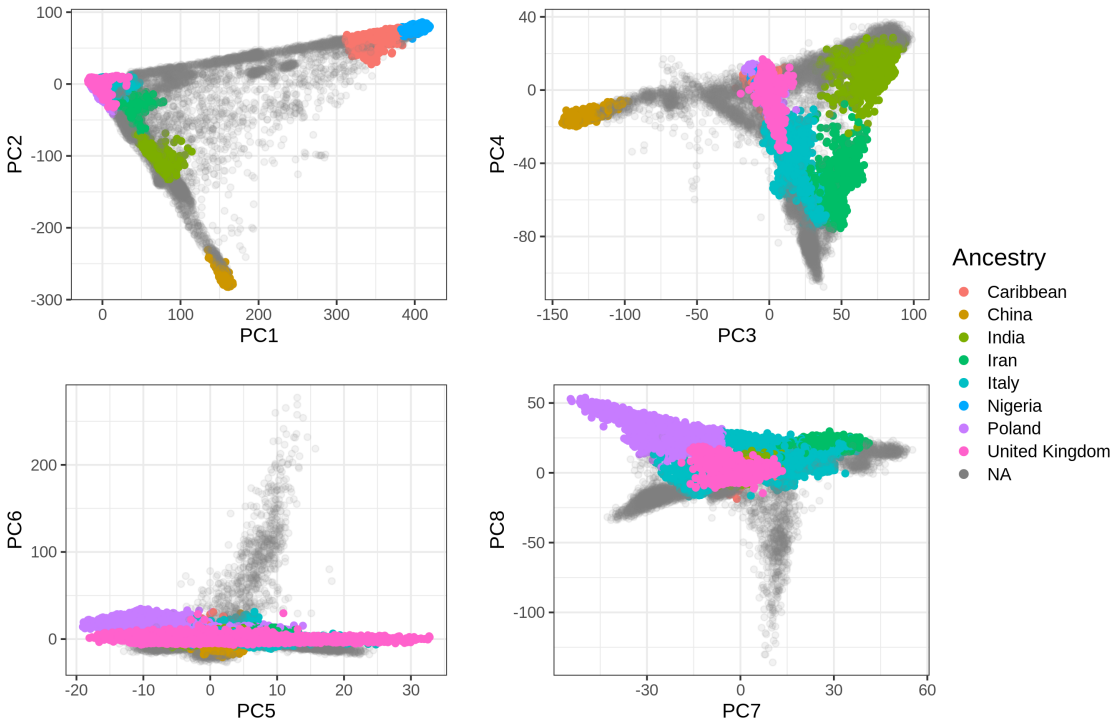

Figure 1: The first eight PC scores of the UK Biobank (Field 22009) colored by the homogeneous ancestry group we infer for these individuals. This is figure 2 in the corresponding supplementary note.

Portability of polygenic scores to other ancestries

Figure 2 presents the results when fitting penalized regression using a training set composed of UK individuals only and testing in eight different ancestry groups. Averaged over 240 phenotypes, compared to prediction performance in individuals of UK ancestry, relative predictive ability in terms of partial-r^2 (Methods) is only 93.8% in Poland, 85.6% in Italy, 72.2% in Iran, 64.7% in India, 48.6% in China, 25.2% in the Caribbean, and 18% in Nigeria. As a follow-up analysis to ensure that this drop in performance in other ancestries is not due to imputation, we perform the same analysis for 83 of the continuous phenotypes using high-quality genotyped variants only (Methods) instead of the (mostly imputed) HapMap3 variants; results are highly consistent (Figure S1). These results are also very similar when using LDpred2-auto instead of penalized regression for training predictive models for all phenotypes (Figure S2). A few phenotypes deviate from this global trend, e.g. prediction of bilirubin concentration ranges between 0.537 and 0.619 (partial-r) for all ancestries except for China for which it is 0.415 (95% CI: 0.374 - 0.453). In contrast, for e.g. hair and skin color, predictive performance decrease quickly and prediction for both China and Nigeria is not significantly different from 0, while of 0.420 (95% CI: 0.409 - 0.432) for "darker hair" (Figure 2). Overall, relative predictive performance decreases approximately linearly with PC distance to the UK (Figure 3). A similar pattern is observed when computing PCA based on more balanced ancestry groups, as recommended in Privé *et al.* (2020a) (Figure S3).

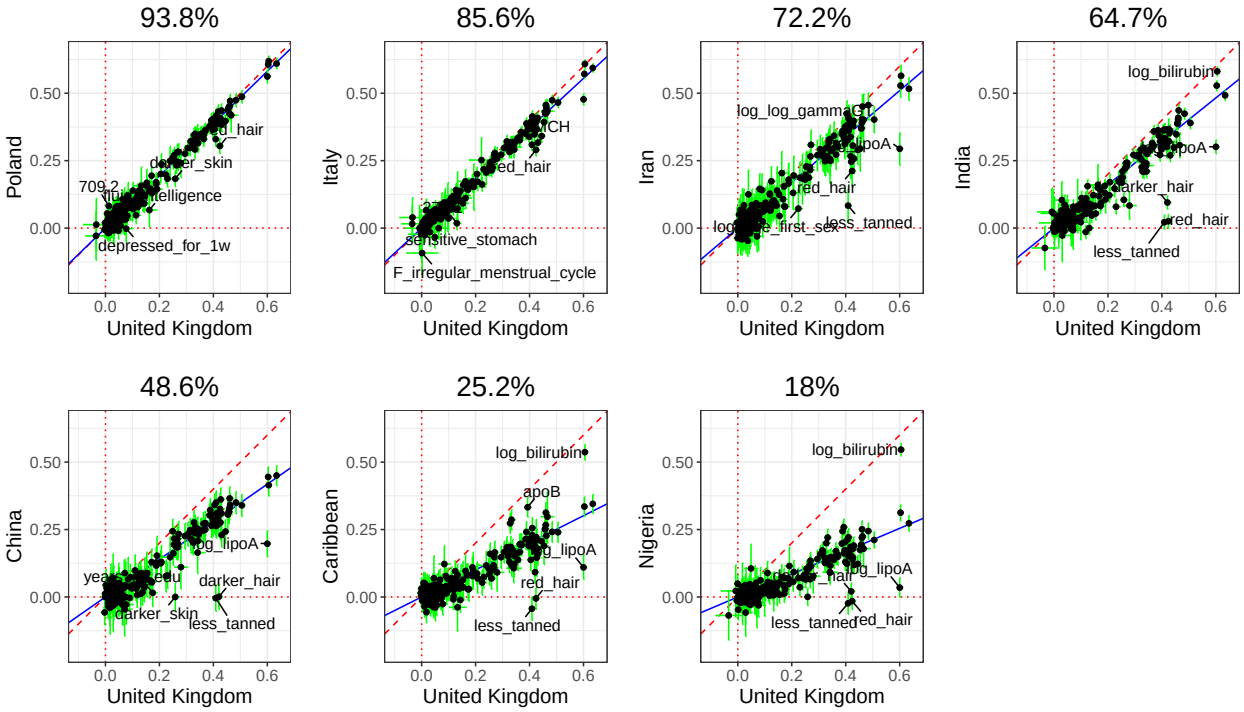

Figure 2: Partial correlation (and 95% CI) in the UK test set versus in a test set from another ancestry group (each panel). Each point represents a phenotype and training has been performed with penalized regression on UK individuals (training 1 in table 1) and HapMap3 variants. The slope (in blue) is computed using Deming regression accounting for standard errors in both x and y. The square of this slope is provided above each plot, which we report as the relative predictive performance compared to testing in UK.

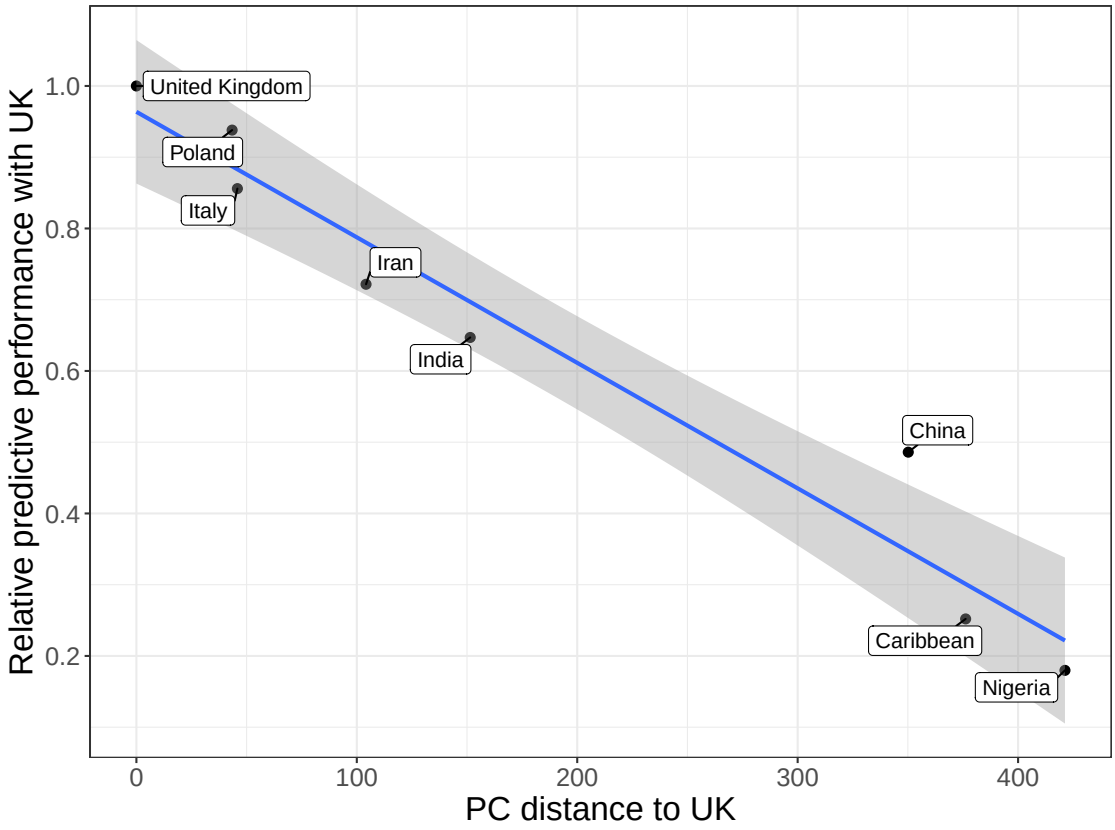

Figure 3: Relative predictive performance with UK (reported in figure 2) compared to PC distance with UK. PC distances are computed using Euclidean distance between geometric medians of the first 16 reported PC scores (Field 22009) of each ancestry group. Relative performance values are the ones reported in figure 2.

Using more than HapMap3 variants?

We investigate some of the outlier phenotypes in figure 2, especially the ones from blood biochemistry which have some variants with large effects. Using a denser set of variants could improve tagging of the causal variants with large effect sizes, resulting in an improved prediction in all ancestries. We focus on "total bilirubin", "lipoprotein(a)" (lipoA) and "apolipoprotein B" (apoB). We perform a localized GWAS which includes all variants around the most significant variant (hereinafter denoted as "top hit") from the GWAS in the training set 1 (UK individuals and HapMap3 variants only) in each of the eight ancestry groups defined here. More precisely, we include all variants with an imputation INFO score larger than 0.3 and within a window of 500Kb from the HapMap3 top hit; there are approximately 30K such variants for all three phenotypes. For bilirubin, the overall top hit is an HapMap3 variant and explains around 30% of the phenotypic variance (Figure S5). Effects from the three top hits are fairly consistent within all ancestry groups (Figure S6) explaining why genetic prediction is highly consistent in all ancestries, except for China (Figure 2), for which these variants are rarer. For lipoA, results are very different across ancestries; HapMap3 variants are far from being the top hits for the UK individuals, where the top HapMap3 variant explains 5% of phenotypic variance compared to 29% for the (non-HapMap3) top hit (Figure 4). Note that this top hit is more than 200Kb away from the HapMap3 top hit from UK. The 3 top hits for lipoA do not

have very consistent effect sizes across ancestries (Figure S7). Finally, for apoB, effects from the three top hits, which are not part of HapMap3 variants, are fairly consistent across ancestries and explain up to 8.5% of the phenotypic variance (Figures S8 & S9).

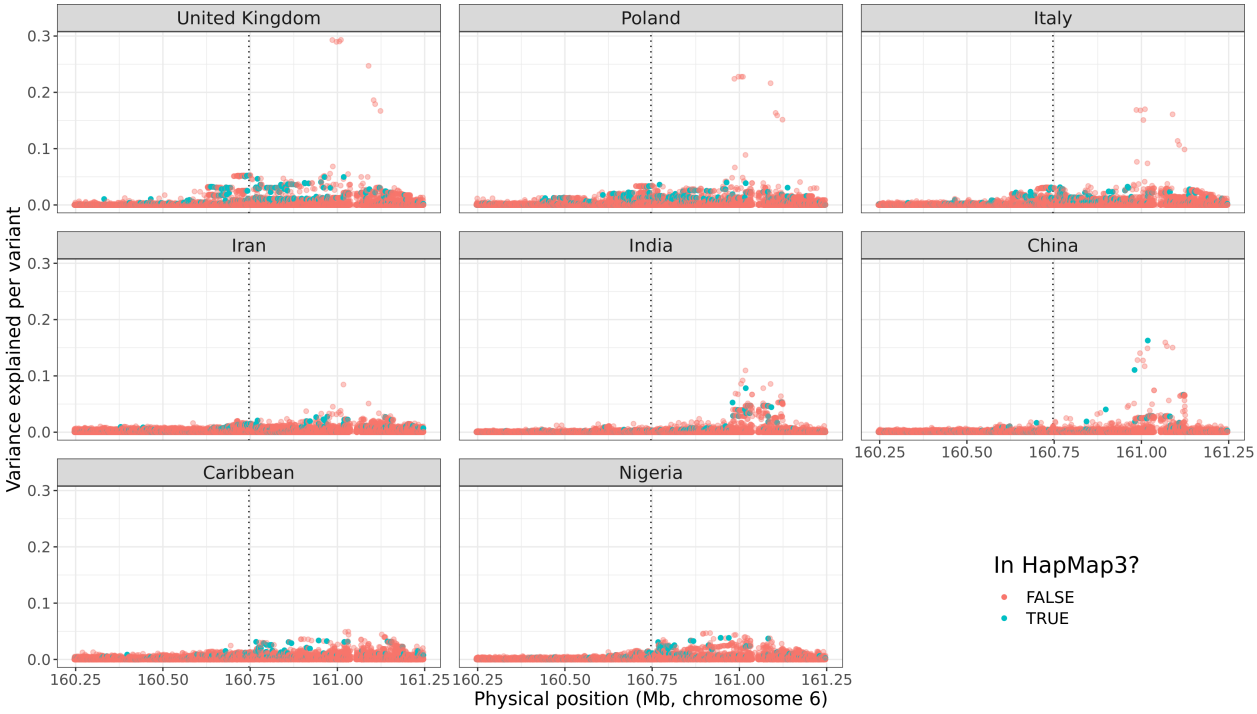

Figure 4: Zoomed Manhattan plot for lipoprotein A concentration. The phenotypic variance explained per variant is computed as $r^2 = t^2/(n + t^2)$, where t is the t-score and n is the degrees of freedom (the sample size minus the number of variables in the model, i.e. the covariates used in the GWAS, the intercept and the variant). The GWAS includes all variants with an imputation INFO score larger than 0.3 and within a 500Kb radius around the top hit from the GWAS performed in the UK training set and on the HapMap3 variants, represented by a vertical dotted line.

We then investigate using more than HapMap3 variants for training; we use more than 8M common variants (Methods) and apply LDpred2-auto thanks to p-value thresholding and winner's curse correction (Methods). Except for lipoA for which we get a large improvement in predictive accuracy compared to using HapMap3 variants only, there is no significant improvement for the other seven phenotypes analyzed here (Figure 5). Remarkably, while the partial correlation for lipoA is about 75% in the UK test set when using this prioritized set of variants, it is still not different from 0 when applied to the Nigeria group. For height and BMI, estimated SNP heritability is reduced when using this set of most significant variants only, and all these variants are estimated to be causal, i.e. estimate of p is 1 (Table S1). As height and BMI are very polygenic traits (p is estimated to be ~2% and ~4% respectively when using HapMap3 variants), contribution from less significant causal variants is missed due to this thresholding selection. For the three binary phenotypes, breast cancer (phecode: 174.1), prostate cancer (185) and coronary artery disease (411.4), while heritability estimates are larger when using this set of prioritized variants (Table S1), predictive accuracy does not improve compared to when using HapMap3 variants (Figure 5).

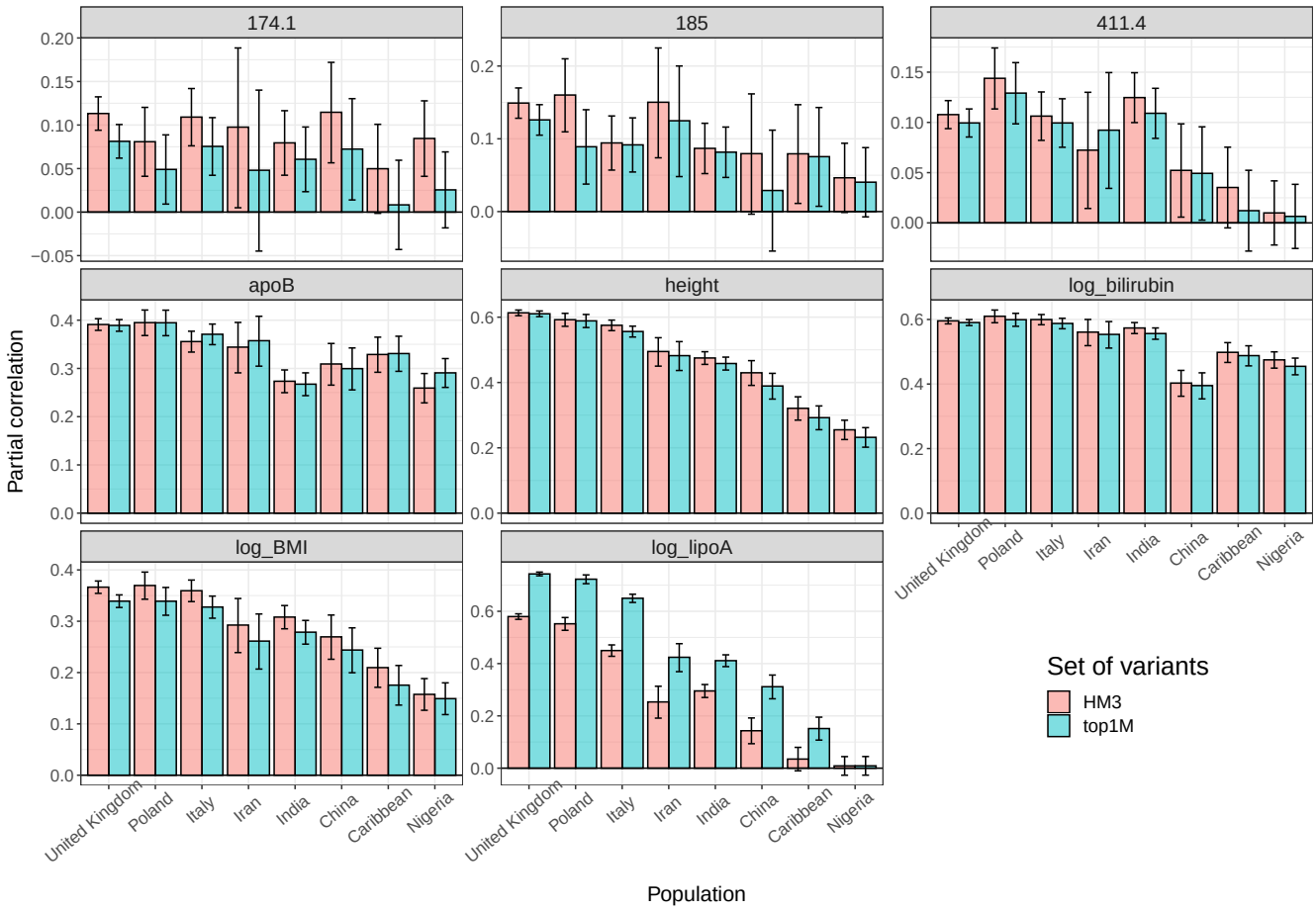

Figure 5: Predictive performance with LDpred2-auto for 8 phenotypes (each panel), when using either HapMap3 variants (HM3) or the 1M most significant variants (top1M) out of more than 8M common variants. Phecode 174.1: breast cancer; 185: prostate cancer; 411.4: coronary artery disease.

Training with a mixture of ancestries

Here we use all ancestry groups except for the Caribbean for training; we remove the same number of UK individuals to keep the same training sample size as before (training 2 in table 1). We recall that Caribbean individuals are mostly admixed between African, European and South Asian ancestries, which are all represented here in the training set 2. In figure S10, we investigate nine phenotypes of interest, either because they are highly studied diseases or are outliers in figure 2: breast cancer (phecode: 174.1), prostate cancer (186), type-2 diabetes (250.2), hypertension (401), CAD (411.4), skin tone, total bilirubin concentration, lipoprotein(a) concentration, and years of education. We predict in the test sets from the UK and the Caribbean (test 2); overall, the predictive performance is highly similar when using this multi-ancestry training compared to when using only UK individuals, in both the UK and the Caribbean target samples. Prediction is only improved for lipoprotein(a) concentration when the mixed ancestry training data is used in application to the Caribbean target data (Figure S10).

Comparison of predictive models

Penalized regression and LDpred2-auto provides approximately similar predictive performance across all traits and ancestries considered here (Figure S11); there are only four pairs of phenotype-ancestry (out of nearly 2000 pairs) for which 95% CIs for partial-r from penalized regression and LDpred2 are not overlapping: "615: Endometriosis" in China with 0.065 (0.0074 - 0.122) vs -0.051 (-0.108 - 0.0068); "hard falling asleep" in UK with -0.0349 (-0.742 - 0.0045) vs 0.071 (0.031 - 0.110); height in UK with 0.634 (0.626 - 0.643) vs 0.613 (0.605 - 0.622); log-bilirubin in Nigeria with 0.546 (0.523 - 0.569) vs 0.475 (0.449 - 0.500). For prediction in UK ancestry, penalized regression tends to provide better predictive performance than LDpred2 for traits that are well predicted (e.g. partial-r > 0.5) and LDpred2 tends to outperform penalized regression for phenotypes harder to predict (Figure S11).

Both methods allow for fitting sparse effects, i.e. some resulting effects are exactly 0. Sparse models may be beneficial because they may be more easily interpreted and implemented. The sparse option in LDpred2-auto provides similar performance as LDpred2-auto without this option (Figure S12). Sparsity of resulting effects follows a very different pattern for penalized regression compared to LDpred2-auto-sparse. Indeed, penalized regression tends not to include variants if it is uncertain that they have a non-zero effect, i.e. when effects are very small and prediction is difficult (Figure S13). In contrast, LDpred2-auto-sparse tends not to discard variants, only when h^2 is large enough it sets lots of effects to 0 if p is small (Figure S14). Finally, running each penalized regression model takes between a few minutes and a few days depending on the number of non-zero effects in the resulting model (Figure S15). In contrast, LDpred2-auto should take the same computation time for all phenotypes; it completed under seven hours for most phenotypes (Figure S16).

Discussion

In this paper, we have shown a poor portability of PGS to other populations, in agreement to what has been previously reported. Indeed, compared to three previous studies (Martin *et al.* 2019; Duncan *et al.* 2019; Wang *et al.* 2020), we show a relative predictive performance compared to Europeans of ~18% for Africans (vs. 22%, 42% and 24%), ~49% for East Asians (vs. 50%, 95% and 64%) and ~65% for South Asians (vs. 60%, 62.5% and 72%). Our results provide a significant addition to the current literature in many ways. First, we show that the portability issue remains strong even when PGS are derived and applied in the same cohort. Second, the presented results are averaged over 240 phenotypes, which is much more than what has been typically used, and should capture a broad range of the phenotypic spectrum. Third, we provide this result at a finer scale than the usual continental level by proposing a simple, robust and effective method for grouping UKBB individuals in eight ancestry groups. This allows us to show e.g. that predictive performance already decreases within Europe with only ~94% for East Europe and ~86% for South Europe of the performance reached within UK.

We showcase two methods for deriving polygenic scores when large individual-level data is available. Although LDpred2-auto is a method based on summary statistics, it provides good predictive performance compared to penalized regression, when applied to individual-level data. Moreover, portability results shown

here are similar when using either the individual-level penalized regression or the summary statistics based LDpred2 method. Fitting of penalized models is relatively fast when using 1M HapMap3 variants; we have also tried fitting penalized regression using 8M variants (>3TB of data); this was possible but took several days for the phenotypes we tried (data not shown) so we have not investigated this further. To the best of our knowledge, the implementation we use is the most efficient penalized regression implementation currently available. Recently, Qian *et al.* (2020) proposed snpnet, a new R package for fitting penalized regressions on large individual-level genetic data, but we have found it to be much less efficient than R package bigstatsr on UKBB data (Supplementary Note). As for LDpred2, it currently cannot be run using 8M variants, but we show a way to use a subset of 1M prioritized variants out of these 8M. Using this new set of variants provides a large improvement in predicting lipoprotein(a) concentration (lipoA), but not for the other seven phenotypes studied in this analysis. This improvement for lipoA is not surprising given that the top HapMap3 variant explains 5% of phenotypic variance compared to 29% for the (non-HapMap3) top hit (Figure 4).

Here we only use the UK Biobank data to fit polygenic scores. We do not use external information such as functional annotations; those could be used to improve the heritability model assumed by predictive methods in order to improve predictive performance (Zhang *et al.* 2020b). Moreover, we do not use external summary statistics. Nevertheless, Albiñana *et al.* (2020) have shown that an efficient strategy to improve predictive ability of polygenic scores consists in combining two different polygenic scores, one derived using external summary statistics, and another one derived using internal individual-level data. Therefore, the polygenic scores derived here could be combined with polygenic scores derived using external summary statistics; we will release these PGS publicly and share them in databases such as the PGS Catalog and the Cancer-PRSweb (Fritsche *et al.* 2020; Lambert *et al.* 2020).

Materials and Methods

Data

We derive polygenic scores for 240 phenotypes using the UK Biobank (UKBB) data only (Bycroft *et al.* 2018). We read dosages data from UKBB BGEN files using function `snp_readBGEN()` of R package bigsnpr (Privé *et al.* 2018). We divide the UKBB data in eight ancestry groups (Supplementary Note), and restrict to 437,669 individuals without second-degree relatives (KING kinship $< 2^{-3.5}$). For the variants, we use 1,040,096 HapMap3 variants used in the LD reference provided in Privé *et al.* (2020b) and that were also present in the iPSYCH2015 data (Bybjerg-Grauholm *et al.* 2020) with imputation INFO score larger than 0.6. Even though the iPSYCH data is not used in this study, we plan to use the PGS derived here for iPSYCH in the future.

To define phenotypes, we first map ICD10 and ICD9 codes (UKBB fields 40001, 40002, 40006, 40013, 41202, 41270 and 41271) to phecodes using R package PheWAS (Carroll *et al.* 2014; Wu *et al.* 2019). We filter down to 142 phecodes of interest that showed potential genetic signals in the PheWeb results from the SAIGE UKBB GWAS (Zhou *et al.* 2018; Taliun *et al.* 2020). There are 104 phecodes for which we can predict something using penalized regression. Second, we look closely at all 2408 UKBB fields that we have

access to and filter down to defining 111 continuous and 25 binary phenotypes based on manual curation.

Additional data: genotyped data

For the genotyped data used in some follow-up analyses, we restrict to variants that have been genotyped on both chips used by the UK Biobank, that pass quality control (QC) for all batches (cf. `https://biobank.ctsu.ox.ac.uk/crystal/crystal/auxdata/ukb_snp_qc.txt`) and QC for possible mismappings (Kunert-Graf *et al.* 2020), with a minor allele frequency (MAF) larger than 0.01 and imputation INFO score of 1. There are 586,534 such high-quality variants, which we read from the BGEN imputed data so that there is no missing value.

Additional data: 8M+ variants

We also design a larger set of imputed variants to compare against using only HapMap3 variants for prediction. We first restrict to UKBB variants with MAF > 0.01 and INFO > 0.6. We then compile frequencies and imputation INFO scores from other datasets, iPSYCH and summary statistics for breast cancer, prostate cancer, coronary artery disease and type-1 diabetes (Bybjerg-Grauholm *et al.* 2020; Michailidou *et al.* 2017; Schumacher *et al.* 2018; Nikpay *et al.* 2015; Censin *et al.* 2017). We restrict to variants with a mean INFO > 0.5 in these other datasets, and also compute the median frequency. To exclude potential mismappings in the genotyped data (Kunert-Graf *et al.* 2020) that might have propagated to imputed data, we compare frequencies in this data to the ones in UKBB (Figure S17). As we expect these potential errors to be localized around errors in the genotype data (confirmed in figure S18), we apply a moving-average smoothing on the frequency differences to increase power to detect these errors and also reduce false positives. This is the same method we have previously applied to PC loadings to detect long-range LD regions when computing PCA (Privé *et al.* 2018, 2020a). This results in a set of 8,238,692 variants.

Penalized regression

To derive polygenic scores based on individual-level data from the UKBB, we use the fast implementation of penalized linear and logistic regressions from R package bigstatsr (Privé *et al.* 2019). We have also considered the recently developed R package snpnet for fitting penalized regressions on large genetic data; however, we provide theoretical and empirical evidence that bigstatsr is much faster than snpnet (Supplementary Note). Our implementation allows for lasso and elastic-net penalizations; yet, for the sake of simplicity and because the UKBB data is very large, we have decided to only use the lasso penalty (Privé *et al.* 2019). We recall that fitting a penalized linear regression with lasso penalty corresponds to finding the vector of effects β (also μ and γ) that minimizes

$$L(\lambda) = \underbrace{||y - (\mu + G\beta + X\gamma)\,||_2^2}_{\text{Loss function}} + \underbrace{\lambda\|\beta\|_1}_{\text{Penalisation}} \quad ,$$

where μ is an intercept, G is the genotype matrix, X is the matrix of covariates, y is the (quantitative) phenotype of interest and λ is a hyper-parameter that needs to be chosen. We use sex (Field 22001), age

(Field 21022), birth date (Fields 34 & 52), Townsend deprivation index (Field 189) and the first 16 principal components (Field 22009, Privé *et al.* (2020a)) as unpenalized covariates when fitting the lasso models.

We have extended our implementation in two ways by allowing for using different penalties for the variants (i.e. having $\sum_j \lambda_j |\beta_j|$ instead of $\lambda \|\beta\|_1$). First, this enables us to use a different scaling for genotypes. By default, variants in G are implicitly scaled. By using $\lambda_j \propto (\text{SD}_j)^{(\xi-1)}$, this effectively scales variant j by dividing it by $(\text{SD}_j)^{\xi}$ in our implementation. The default uses $\xi = 1$ but we also test $\xi = 0$ (no scaling) and $\xi = 0.5$ (Pareto scaling). We introduce a new parameter `power_scale` for which the user can provide a vector of values to test; the best value is chosen within the Cross-Model Selection and Averaging (CMSA) procedure (Privé *et al.* 2019). We also introduce a second parameter, `power_adaptive`, which can be used to put less penalizition on variants with the largest marginal effects (Zou 2006); we try 3 values here (0 the default, 0.5 and 1.5) and the best one is also chosen within the CMSA procedure.

LDpred2-auto

Using the individual-level data from the training set in the UK biobank, we run a linear regression GWAS using function `big_univLinReg` of R package bigstatsr (Privé *et al.* 2018), accounting for the same covariates as in the penalized regression above. As LD reference, we use the one provided in Privé *et al.* (2020b) based on UKBB data for European ancestry. We use these summary statistics and this LD reference as input for LDpred2-auto. LDpred2 assumes a point-normal mixture distribution for effect sizes, where only a proportion of causal variants p contributes to the SNP heritability h^2. In LDpred2-auto, these two parameters are directly estimated from the data (Privé *et al.* 2020b). We use the `sparse` option in LDpred2-auto to also obtain a vector of effects that is potentially sparse, i.e. effects of some variants are exactly 0. Also note that, as we use linear regression for all phenotypes, we use the total sample size instead of the effective sample size $(4/\left(1/n_{\text{case}} + 1/n_{\text{control}}\right))$ for binary phenotypes as input to LDpred2. This means that heritability estimates from both LD score regression and LDpred2-auto must be transformed to the liability scale using both the prevalence in the GWAS and in the population; this can be performed using function `coef_to_liab` from R package bigsnpr. For simplicity, we assume here that the prevalence in the population is the same as the prevalence in the training set.

New formula used in LDpred2

We also slightly modify the formula used in Privé *et al.* (2020b); we have previously used

$$(\text{se}(\hat{\gamma}_j))^2 = \frac{(\breve{\boldsymbol{y}} - \hat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}})^T (\breve{\boldsymbol{y}} - \hat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}})}{(n-K-1)\, \breve{\boldsymbol{G}}_{\boldsymbol{j}}{}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}} \approx \frac{\breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}}}{n\, \breve{\boldsymbol{G}}_{\boldsymbol{j}}{}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}}} \approx \frac{\text{var}(\boldsymbol{y})}{n\, \text{var}(\boldsymbol{G}_{\boldsymbol{j}})}\ ,$$

where $\hat{\gamma}_j$ is the marginal effect of variant j, and where $\breve{\boldsymbol{y}}$ and $\breve{\boldsymbol{G}}_{\boldsymbol{j}}$ are the vectors of phenotypes and genotypes for variant j residualized from K covariates, e.g. centering them. The first approximation expects $\hat{\gamma}_j$ to be small, while the second approximation assumes the effects from covariates are small. However, we found here that some variants can have very large effects, e.g. one variant explains about 30% of the variance in

bilirubin log-concentration. Then, instead we compute

$$(\breve{\boldsymbol{y}} - \hat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}})^T (\breve{\boldsymbol{y}} - \hat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}) = \breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}} - 2\hat{\gamma}_j \breve{\boldsymbol{G}}_{\boldsymbol{j}}{}^T \breve{\boldsymbol{y}} + \hat{\gamma}_j^2 \breve{\boldsymbol{G}}_{\boldsymbol{j}}{}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}} = \breve{\boldsymbol{y}}^T \breve{\boldsymbol{y}} - \hat{\gamma}_j^2 \breve{\boldsymbol{G}}_{\boldsymbol{j}}{}^T \breve{\boldsymbol{G}}_{\boldsymbol{j}} \ ,$$

which now gives (note the added term $\hat{\gamma}_j^2$)

$$\mathrm{sd}(\boldsymbol{G}_{\boldsymbol{j}}) \approx \frac{\mathrm{sd}(\breve{\boldsymbol{y}})}{\sqrt{n \ \mathrm{se}(\hat{\gamma}_j)^2 + \hat{\gamma}_j^2}} \ . \tag{1}$$

Figure S19 shows that the updated formula (1) is better; we now use it in the code of LDpred2, and also recommend using it for the QC procedure proposed in Privé *et al.* (2020b).

Using more than HapMap3 variants in LDpred2

Here we also run LDpred2 using more than HapMap3 variants, based on a set of 8M+ variants (see above). However, LDpred2 cannot be run on 8M variants because the implementation is quadratic with the number of variants in terms of time and memory requirements. Thus, we employ another strategy consisting in keeping only the 1M most significant variants. To correct for winner's curse, we employ the maximum likelihood estimator used in Zhong and Prentice (2008) and Shi *et al.* (2016):

$$Z = Z^* + \frac{\phi\left(Z^* - Z_{thr}\right) - \phi\left(-Z^* - Z_{thr}\right)}{\Phi\left(Z^* - Z_{thr}\right) + \Phi\left(-Z^* - Z_{thr}\right)} \ ,$$

where ϕ is the standard normal density function, Φ is the standard normal cumulative density function, Z is the Z-score obtained from the GWAS, Z_{thr} is the threshold used on (absolute) Z-scores for filtering, and Z^* is the corrected Z-score that we estimate and use. As input for LDpred2, instead of using β (along with SE(β) and N), we use $\beta^* = \beta \cdot Z^*/Z$ where $Z = \beta/\mathrm{SE}(\beta)$. This is now implemented in function `snp_thr_correct` of package bigsnpr.

Performance metric

Here we use the partial correlation as the performance metric, which is the correlation between the PGS and the phenotype after they have been both residualized using the covariates used in this paper, i.e. sex, age, birth date, deprivation index and 16 PCs. To derive 95% confidence intervals for these correlations, we use Fisher's Z-transformation. We implement this in function `pcor` of R package bigstatsr and use it here.

Software and code availability

All code used for this paper is available at `https://github.com/privefl/UKBB-PGS/tree/master/code`. Links to the code used for the two supplementary notes are provided there. We have extensively used R packages bigstatsr and bigsnpr (Privé *et al.* 2018) for analyzing large genetic data, packages from the future framework (Bengtsson 2020) for easy scheduling and parallelization of analyses

on the HPC cluster, and packages from the tidyverse suite (Wickham *et al.* 2019) for shaping and visualizing results. We have also used R package deming to fit Deming regressions.

R packages bigstatsr and bigsnpr can be installed from GitHub and CRAN. A tutorial on fitting penalized regressions with R package bigstatsr is available at `https://privefl.github.io/bigstatsr/articles/penalized-regressions.html`. A tutorial on running LDpred2 with R package bigsnpr is available at `https://privefl.github.io/bigsnpr/articles/LDpred2.html`.

Acknowledgements

Authors thank GenomeDK and Aarhus University for providing computational resources and support that contributed to these research results. This research has been conducted using the UK Biobank Resource under Application Number 58024.

Funding

F.P. and B.J.V. are supported by the Danish National Research Foundation (Niels Bohr Professorship to Prof. John McGrath), and also acknowledge the Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH (R248-2017-2003). B.J.V. is also supported by a Lundbeck Foundation Fellowship (R335-2019-2339)

Declaration of Interests

The authors declare no competing interests.

References

Abraham, G., Tye-Din, J. A., Bhalala, O. G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS genetics*, **10**(2), e1004137.

Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*.

Albiñana, C., Grove, J., McGrath, J. J., Agerbo, E., Wray, N. R., Werge, T., Børglum, A. D., Mortensen, P. B., Privé, F., and Vilhjálmsson, B. J. (2020). Leveraging both individual-level genetic data and gwas summary statistics increases polygenic prediction. *bioRxiv*.

Bengtsson, H. (2020). A unifying framework for parallel and distributed processing in R using futures. *arXiv preprint arXiv:2008.00553*.

Bitarello, B. D. and Mathieson, I. (2020). Polygenic scores for height in admixed populations. *bioRxiv*.

Bybjerg-Grauholm, J., Pedersen, C. B., Baekvad-Hansen, M., Pedersen, M. G., Adamsen, D., Hansen, C. S., Agerbo, E., Grove, J., Als, T. D., Schork, A. J., *et al.* (2020). The ipsych2015 case-cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *medRxiv*.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Carroll, R. J., Bastarache, L., and Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, **30**(16), 2375–2376.

Cavazos, T. B. and Witte, J. S. (2020). Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *bioRxiv*.

Censin, J., Nowak, C., Cooper, N., Bergsten, P., Todd, J. A., and Fall, T. (2017). Childhood adiposity and risk of type 1 diabetes: A mendelian randomization study. *PLoS Medicine*, **14**(8), e1002362.

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, pages 1–14.

De Los Campos, G., Gianola, D., and Allison, D. B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, **11**(12), 880–886.

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications*, **10**(1), 1–9.

Fritsche, L. G., Patil, S., Beesley, L. J., VandeHaar, P., Salvatore, M., Ma, Y., Peng, R. B., Taliun, D., Zhou, X., and Mukherjee, B. (2020). Cancer prsweb: An online repository with polygenic risk scores for major cancer traits and their evaluation in two independent biobanks. *The American Journal of Human Genetics*, **107**(5), 815–836.

Kim, M. S., Patel, K. P., Teng, A. K., Berens, A. J., and Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. *Genome biology*, **19**(1), 1–14.

Kunert-Graf, J. M., Sakhanenko, N. M., and Galas, D. J. (2020). Allele frequency mismatches and apparent mismappings in uk biobank snp data. *bioRxiv*.

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., Abraham, G., Chapman, M., Parkinson, H., Danesh, J., *et al.* (2020). The polygenic score catalog: an open database for reproducibility and systematic evaluation. *medRxiv*.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature genetics*, **50**(7), 906–908.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, **51**(4), 584–591.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., *et al.* (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**(7678), 92.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10), 1121.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., *et al.* (2008). Genes mirror geography within Europe. *Nature*, **456**(7218), 98–101.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**(8), 904–909.

Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, **34**(16), 2781–2787.

Privé, F., Aschard, H., and Blum, M. G. B. (2019). Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, **212**(1), 65–74.

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

Privé, F., Arbel, J., and Vilhjálmsson, B. J. (2020b). LDpred2: better, faster, stronger. *BioRxiv*.

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M. A., and Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genetics*, **16**(10), 1–31.

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., *et al.* (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, **50**(7), 928.

Scutari, M., Mackay, I., and Balding, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. *PLoS genetics*, **12**(9), e1006288.

Shi, H., Burch, K. S., Johnson, R., Freund, M. K., Kichaev, G., Mancuso, N., Manuel, A. M., Dong, N., and Pasaniuc, B. (2020). Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *The American Journal of Human Genetics*.

Shi, J., Park, J.-H., Duan, J., Berndt, S. T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., *et al.* (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS genetics*, **12**(12), e1006493.

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G. R., Wainberg, M., Ollila, H. M., Kiiskinen, T., *et al.* (2021). Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature Genetics*, pages 1–10.

Taliun, S. A. G., VandeHaar, P., Boughton, A. P., Welch, R. P., Taliun, D., Schmidt, E. M., Zhou, W., Nielsen, J. B., Willer, C. J., Lee, S., *et al.* (2020). Exploring and visualizing large-scale genetic associations by using pheweb. *Nature Genetics*, **52**(6), 550–552.

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature communications*, **11**(1), 1–9.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., *et al.* (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686.

Wu, P., Gifford, A., Meng, X., Li, X., Campbell, H., Varley, T., Zhao, J., Carroll, R., Bastarache, L., Denny, J. C., *et al.* (2019). Mapping icd-10 and icd-10-cm codes to phecodes: workflow development and initial evaluation. *JMIR Medical Informatics*, **7**(4), e14325.

Zhang, D., Dey, R., and Lee, S. (2020a). Fast and robust ancestry prediction using principal component analysis. *Bioinformatics*, **36**(11), 3439–3446.

Zhang, Q., Privé, F., Vilhjalmsson, B. J., and Speed, D. (2020b). Improved genetic prediction of complex traits from individual-level data or summary statistics. *bioRxiv*.

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, **9**(4), 621–634.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., *et al.* (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, **50**(9), 1335–1341.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.