How robust are cross-population signatures of polygenic adaptation in humans?

Alba Refoyo-Martínez[1,*], Siyang Liu[2,3], Anja Moltke Jørgensen[4], Xin Jin[3], Anders Albrechtsen[2], Alicia R. Martin[5,6,7], Fernando Racimo[1]

[1]Lundbeck GeoGenetics Centre, GLOBE Institute, University of Copenhagen, 1350, Copenhagen, Denmark.

[2]Department of Biology, Section for Computational and RNA Biology, University of Copenhagen, 2200, Copenhagen, Denmark.

[3]BGI-Shenzhen, Shenzhen 518083, Guangdong, China.

[4]Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, 2200, Copenhagen, Denmark.

[5]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.

[6]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.

[7]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.

*Corresponding author.

July 13, 2020

Abstract

Over the past decade, summary statistics from genome-wide association studies (GWAS) have been used to detect and quantify polygenic adaptation in humans. Several studies have reported signatures of natural selection at sets of SNPs associated with complex traits, like height and body mass index. However, more recent studies suggest that some of these signals may be caused by biases from uncorrected population stratification in the GWAS data with which these tests are performed. Moreover, past studies have predominantly relied on SNP effect size estimates obtained from GWAS panels of European ancestries, which are known to be poor predictors of phenotypes in non-European populations. Here, we collated GWAS data from multiple anthropometric and metabolic traits that have been measured in more than one cohort around the world, including the UK Biobank, FINRISK, Chinese NIPT, Biobank Japan, APCDR and PAGE. We then evaluated how robust signals of polygenic adaptation are to the choice of GWAS cohort used to identify associated variants and their effect size estimates, while using the same panel to obtain population allele frequencies (The 1000 Genomes Project). We observe many discrepancies across tests performed on the same phenotype and find that GWAS meta-analyses produce scores with strong overdispersion across populations. This results in apparent signatures of polygenic adaptation which are not observed when using effect size estimates from biobank-based GWAS of homogeneous ancestries. Indeed, we were able to artificially create score overdispersion when taking a homogeneous cohort like the UK Biobank, and simulating a meta-analysis on multiple subsets of the cohort. This suggests that extreme caution should be taken in the execution and interpretation of future tests of polygenic adaptation based on population differentiation, especially when using summary statistics from GWAS meta-analyses.

Introduction

Most human phenotypes are polygenic: the genetic component of trait variation across individuals is caused by differences in genotypes between individuals at a large number of variants, each with a relatively small contribution to the trait (Fisher *et al.* , 1918; Turelli, 2017). This applies to traits as diverse as a person's height, their risk of schizophrenia or their risk of developing arthritis. The study of complex traits spans more than a century but only in the last two decades has it become possible to systematically explore the genetic variation underlying these traits (Sella & Barton, 2019). The advent of genome-wide association studies (GWAS) has led to the identification of thousands of variants that are associated with such traits, either due to true biological mechanisms or because of linkage with causal variants (Visscher *et al.* , 2012).

However, most research into the genetic aetiology of complex traits is based on GWAS data from populations of European ancestries (Popejoy & Fullerton, 2016). This bias in representation contributes to existing disparities in medical genetics and healthcare around the world (Martin *et al.* , 2019). The low portability of European GWAS results - and, in particular, polygenic scores - to non-European populations is particularly concerning (Martin *et al.* , 2017, 2019) (but see (Ragsdale *et al.* , 2020)). Important trait-associated variants in non-European populations may be missed if they have low frequencies or are absent in European populations. Moreover, effect size estimates for an associated variant derived from a European-ancestry GWAS may not accurately reflect the effect of the same variant on the trait in other populations (Wojcik *et al.* , 2019). This could be due to differences in epistasis, differences in linkage disequilibrium between causal and ascertained variants, or gene-by-environment interactions, to name a few causes (Guo *et al.* , 2018). Additionally, negative selection and demographic history may cause differences in genetic architectures between populations (Durvasula & Lohmueller, 2019).

During the last decade, GWAS summary statistics have also been used to look for evidence of directional selection pushing a trait to a new phenotypic optimum, via allele-frequency shifts occurring across a large number of associated variants - a phenomenon known as polygenic adaptation (Pritchard *et al.* , 2010; Hayward & Sella, 2019). For example, several studies have consistently found evidence for polygenic adaptation operating on height-associated variants across a south-to-north gradient in Europe (Turchin *et al.* , 2012; Berg & Coop, 2014; Robinson *et al.* , 2015). To test for selection, these studies primarily relied on summary statistics from the GIANT consortium dataset, which is a meta-analysis of anthrompometric GWAS from multiple European cohorts (Allen *et al.* , 2010; Turchin *et al.* , 2012; Wood *et al.* , 2014; Racimo *et al.* , 2018; Berg *et al.* , 2017). They looked for overdispersion in the frequencies of trait-associated variants across populations, relative to a neutral null model. To account for potential confounding due to population stratification, some have tried to replicate this signal using family-based association studies (Allison *et al.* , 1999; Robinson *et al.* , 2015). Berg *et al.* (2019) and Sohail *et al.* (2019) showed that this signal of polygenic adaptation on height-associated variants in Europe (and possibly on other trait-associated variants) is attenuated and in some cases no longer significant when using effect size estimates from GWAS performed on the UK Biobank - a large cohort composed primarily of individuals of British ancestry (Bycroft *et al.* , 2018). There is no single explanation yet for these contradictory findings, but the most plausible one is that previous studies were impacted by very subtle confounding due to uncorrected population stratification in GIANT, and that data from family-based studies was not analyzed properly (Berg *et al.* , 2019; Sohail *et al.* , 2019).

It is as yet unclear how the choice of GWAS cohort affects tests of polygenic adaptation based on allele frequency differences between populations. Each cohort differs in ancestries of participants, inclusion criteria of individuals, SNP ascertainment scheme and association method. Given the poor portability of polygenic scores across populations, is it also true that GWAS

performed on different cohorts will result in inconsistent signals of selection? Can we narrow down on the reason for the inconsistencies in previous studies of polygenic adaptation by looking at a larger number of cohorts? Here, we collated GWAS summary statistics from multiple complex traits that have been measured in more than one cohort around the world. We then evaluated how robust signals of polygenic adaptation are to the choice of cohort used to obtain effect size estimates. Across all comparisons, we used the same population genomic panel to obtain population allele frequency estimates: The 1000 Genomes Project phase 3 (The 1000 Genomes Project Consortium, 2015). We observe many discrepancies across tests performed on the same phenotype and attempt to understand what may be causing these discrepancies. Although we compare results for several traits, we pay special attention to height, as it is the most well-characterized and studied complex trait in the human genetics literature, as well as a trait for which we have summary statistics from the largest number of GWAS cohorts.

Methods

GWAS summary statistics

We obtained GWAS summary statistics from five large-scale biobanks and two GWAS meta-analyses. Since we aim to make comparisons among them, our interest is focused on traits that were measured in at least two different studies. This resulted in a total of 30 traits being included in our analysis.

Below, we provide a brief summary of each of the GWASs we focused on. For an overview of the type of arrays and association methods used in each of these, see Table 1.

UKBB: Summary statistics from the GWAS performed on all UK Biobank traits (Bycroft *et al.* , 2018). These were released by the Neale lab (round 2: `http://www.nealelab.is/uk-biobank/`), after filtering for individuals with European ancestries. The UK Biobank includes genetic and phenotypic data from participants from across the United Kingdom, aged between 40 and 69. The traits measured include a wide range of lifestyle factors, physical measurements, and other phenotypic information gained from blood, urine and saliva samples. The Neale lab performed association testing in $\sim$ 340,000 unrelated individuals.

FINRISK: Summary statistics from GWASs carried out using the National FINRISK 1992-2012 collection from Finland. The FINRISK study is coordinated by the National Institute for Health and Welfare (THL) in Finland and its target population is sampled from six different geographical areas in Northern Finland. The FINRISK cohort was conducted as a cross-sectional population survey every 5 years from 1972 to assess the risk factors of chronic diseases and health behavior in the working age population. Blood samples were collected from 1992 to 2012. Anthropometric measures and other lifestyle information were also collected. The number of samples used for the GWAS results varies among the different traits ($\sim$25,000 to $\sim$5,000) (Borodulin *et al.* , 2018).

PAGE: Summary statistics from a meta-analysis GWAS performed by the PAGE (Population Architecture using Genomics and Epidemiology) consortium (http://www.pagestudy.org). This is a multi-ethnic association project developed by the National Human Genome Research Institute and the National Institute on Minority Health and Health Disparities, to characterize population-level disease risks in various populations from the Americas (Matise *et al.* , 2011; Carlson, 2016). The meta-analysis was assembled from four different cohorts: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), the Women's Health Initiative (WHI), the Multiethnic Cohort (MEC) and the Icahn School of Medicine at Mount Sinai BioMe biobank in New York City (BioMe). The authors performed GWAS on 26 clinical and behavioural phenotypes. The study includes samples from 49,839 non-European-descent individuals. Genotyped

individuals self-reported as Hispanic/Latino (n = 22,216), African American (n = 17,299), Asian (n = 4,680), Native Hawaiian (n = 3,940), Native American (n = 652) or Other (n = 1,052). The number of variants analyzed varies from 22 to 25 million for continuous phenotypes and 11 to 28 million for case/control traits. Sample sizes ranged from 9,066 to 49,796 individuals (Wojcik *et al.* , 2019).

BBJ: Summary statistics from GWASs performed using the Biobank Japan Project, which enrolled 200,000 patients from 12 medical institutions located throughout Japan between 2003-2008. The authors collected biological samples and other clinical information related to 47 diseases and self-reported anthropometric measures. GWAS were then conducted on approximately 162,000 individuals to identify genetic variants associated with disease susceptibility and drug responses. Around 6 million variants were included in this GWAS (Nagai *et al.* , 2017; Hirata *et al.* , 2017; Kanai *et al.* , 2018).

Chinese NIPT: Summary statistics from a GWAS performed in China using non-invasive prenatal testing (NIPT) samples from $\sim$ 141,431 pregnant women. The participants were recruited from 31 administrative divisions across the country. The study aimed to investigate genetic associations with maternal and infectious traits, as well as two antropometric traits: height and BMI (Liu *et al.* , 2018a). It included $\sim$ 60,000 individuals. The number of imputed variants used was around 2 million.

APCDR: Summary statistics performed using the African Partnership for Chronic Disease Research cohort, which was assembled to conduct epidemiological and genomic research of non-communicable diseases across sub-Saharan Africa. The dataset includes 4,956 samples from Uganda (Baganda, Banyarwanda, Burundi, and others). The authors performed GWAS on 34 phenotypes, including anthropometric traits, blood factors, glycemic control, blood pressure, lipid tests, and liver function tests (Heckerman *et al.* , 2016).

GIANT: Summary statistics published by the Genetic Investigation of Anthropometric Traits consortium (2012-2015 version, before including UK Biobank individuals) (Wood *et al.* , 2014; Locke *et al.* , 2015). GIANT is a meta-analysis of summary association statistics for various anthropometric traits, and includes information from more than 250,000 individuals of European descent. The meta-analysis was performed on 2.5 million autosomal SNPs, after imputation.

Population genomic panel

We used the 1000 Genomes Project phase 3 release data (Consortium *et al.* , 2015) to retrieve population-specific allele frequencies of trait-associated variants, so as to compute polygenic scores for each population, using autosomal SNPs only. The dataset contains samples from 2,504 people from 26 present-day population panels, whose abbreviations and descriptions are listed in Table S1.

Identifying trait-associated SNPs

We used summary statistics for a set of 30 traits that were measured in at least two of the previously-listed GWAS datasets. Table S2 shows the full list of the traits included in this analysis and the number of variants and individuals per trait. For each trait, we excluded triallelic variants, variants with a minor allele frequency lower than 0.01 and those classified as low confident variants whenever this information was available in the summary statistics file. We selected a set of trait-associated SNPs based on a P-value threshold, and the effect size estimates of these variants were used to construct a set of polygenic scores. To only include approximately independent trait-associated variants in our scores, we use a published set of 1,700 non-overlapping and approximately independent linkage-disequilibrium (LD) blocks to divide

the genome (Berisa & Pickrell, 2016). We extracted the SNP within each block with the lowest association P-value. To investigate the robustness of signals to different filtering schemes, we used two P-value thresholds to extract significantly associated variants: 1) $P < 1e{-}5$ and 2) the standard genome-wide significant cutoff, $P < 5e{-}8$. As an example, Figure S1 shows the distribution of effect size estimates of 1700 approximately independent SNPs for height ($P < 1e{-}5$). In turn, Figure S2 shows the distribution of their effect size estimates scaled by the square root of the study's sample size, which serves as a fairer comparison among studies. In order to build an empirical genome-wide covariance matrix (F-matrix) with non-associated SNPs, we extracted all SNPs with a P-value larger than 5e−8 and then sampled every 20th "non-associated" SNP across the entire genome. We also used the LD score regression approach (Heckerman *et al.* , 2016) to obtain an LD score regression intercept, LD score regression ratio, and a SNP heritability estimate for each GWAS that we looked at.

Neutrality test for polygenic scores

Polygenic risk scores aim to predict the genetic risk of a disease, or the genetic value of a trait, by combining the additive effect of a large number of trait-associated loci across the genome. For each trait, we obtained polygenic scores by computing the sum of allele frequencies at each of the top trait-associated SNPs from each block, weighted by their effect size estimates for that trait. The allele frequencies for these SNPs were retrieved from The 1000 Genomes Project population panels using *glactools* (Renaud, 2017). We then built a polygenic score vector for a given trait, $\vec{Z}$, that contains the polygenic scores of all populations for that trait. Let $\vec{p_l} \in [0,1]^M$ be the vector of derived allele frequencies at locus l, where $p_{l,m}$ is the derived allele frequency at locus l in population m, while α_l is the effect size estimate of the derived allele at locus l. Then, the vector of the polygenic scores, $\vec{Z}$, has length M equal to the number of populations (M = 26) and each element Z_m is the polygenic score for population m

$$Z_m = \sum_{l=1}^{L} 2\alpha_l p_{l,m} \tag{1}$$

Here, L is the total number of trait-associated loci.

Berg & Coop (2014) introduced a model designed for comparing polygenic scores across populations, in order to test for deviations from neutrality, which could perhaps be driven by adaptive divergence between populations. The test works by looking for overdispersion from a multivariate normal distribution, which would fit the distribution of scores if this was determined purely by genetic drift.

Under neutral genetic drift, Berg & Coop (2014) showed that the joint distribution of $\vec{Z}$ across closely-related populations should be approximately multivariate normal under a purely neutral model:

$$\vec{Z} \sim MVN(\mu\vec{1}, 2V_A\mathbf{F}) \tag{2}$$

where $\vec{1}$ is a vector of ones and:

$$\mu = 2\sum_{l=1}^{L} \alpha_l \overline{p}_l \tag{3}$$

$$V_A = 2\sum_{l=1}^{L} \alpha_l^2 \overline{p}_l (1 - \overline{p}_l) \tag{4}$$

and $\overline{p}_l$ is the average allele frequency of locus l across all populations. The matrix $\mathbf{F}$ is a genome-wide covariance matrix that captures the co-ancestry among each pair of populations (Berg & Coop, 2014). Based on this null model, we can measure the Mahalanobis distance of the observed distribution of $\vec{Z}$ from the distribution under neutral genetic drift by computing Q_x

$$Q_X = \frac{(\vec{Z} - \mu)^T \mathbf{F}^{-1} (\vec{Z} - \mu)}{2V_A} \tag{5}$$

Under neutrality, the Q_X statistic is expected to follow a chi-squared distribution with M -1 degrees of freedom, χ^2_{M-1} (Berg & Coop, 2014). A significantly large value of Q_X indicates that there is an excess of variance in $\vec{Z}$ that cannot be explained by drift alone.

P-values via randomization schemes

To avoid relying on the assumption that $\vec{Z}$ follows a multivariate normal distribution under neutrality, we also obtained P-values via two alternative methods (Berg & Coop, 2014; Berg *et al.* , 2017; Racimo *et al.* , 2018). The first one relies on obtaining neutral pseudo-samples by randomizing the sign of the effect size estimates of all trait-associated SNPs, and then recomputing Q_X. The second one involves obtaining pseudo-samples by sampling random SNPs across the genome with the same allele frequency distribution in a particular (target) population as the SNPs used to computed Q_X. For each trait-associated SNP, we thus sampled a new SNP from a subset of the non-associated SNPs whose frequencies lie in the range $[0.01 - p, p + 0.01]$ where p is the derived allele frequency of the trait-associated SNP. Then, we obtained a new P-value by computing the Q_X statistic on each of the pseudo-samples i:

$$P = \frac{1 + \sum_i^N I(Q_X^i > Q_X^A)}{1 + N} \tag{6}$$

Here, Q_X^i is the Q_X statistic computed on pseudo-sample i, Q_X^A is the Q_X statistic computed on the true set of trait-associated SNPs, $I()$ is an indicator function and N is the number of pseudo-samples used, which was set to 1,000. We tested the effect of using different population panels as our 'target' population for the frequency-matching scheme. Since we are utilizing seven GWAS cohorts that are composed of Latin American individuals, Asian (Japanese and Chinese), sub-Saharan African and European (Finnish and British) individuals, we decided to use population panels from the 1000 Genomes that roughly matched these ancestries: PUR, CHB, JPT, LWK, FIN and GBR.

Assessing different association methods

We were also interested in evaluating the effects of different types of association methods on the significance of the Q_X statistic. We used the UKBB cohort to perform different types of association studies on height. Starting from 805,426 genotyped variants across the genome, we restricted to SNPs with a minor allele frequency (MAF) $< 5\%$ globally, and performed associations on three different sets of individuals from the UKBB cohort: 1) self-reported white British individuals ("British"), 2) self-reported "white" individuals, and 3) "all ethnicities", i.e. a UKBB set including all self-reported ethnicity categories. We applied the following quality filters in each of the resulting sets: 1) removed variants with $P < 1e - 10$ from the Hardy-Weinberg equilibrium test, 2) removed variants with $MAF < 0.1\%$ in the set, 3) removed variants with an INFO score less than 0.8 4) removed variants outside the autosomes, and 5) removed individuals that were 7 standard deviations away from the first six PCs in a PCA of the set. We then

performed a GWAS via a linear model (LM) with PLINK 1.9 (Chang *et al.* , 2015) and 2) a GWAS via a linear mixed model (LMM) using BOLT-LMM (Loh *et al.* , 2018), on each of the three sets (Table 2). We used sex, age, age^2, sex*age, sex*age^2 and the first 20 PCs as covariates.

We also aimed to test whether a meta-analysis approach could lead to overdispersion of polygenic scores, and consequently, an inflated Q_X statistics. Therefore, we created an artificial meta-analysis on the entire UKBB cohort, approximately emulating the number of individual sub-cohorts that were included in GIANT. We divided the "all ethnicities" set of individuals into 75 sets, using two different approaches. In one approach, we obtained 75 clusters from a K-means clustering of the first three principal components from a PCA of the the UKBB individuals. In the other approach, we created 75 groups of equal size, randomly assigning individuals to each group, regardless of their PCA placement. We used PLINK 1.9 to perform a linear association model in each of the 75 clusters or groups. Afterwards, we integrated all summary statistics into a meta-analysis, using two different meta-analysis methods (Table 2): an inverse variance method and a sample size-based method, both implemented in METAL (Willer *et al.* , 2010). As before, we used sex, age, age^2, sex*age, sex*age^2 and the first 20 PCs as covariates.

Source code

The code used to perform the analyses in this manuscript is available at: `https://github.com/albarema/GWAS_choice/`

Results

Robustness of signal of selection and population-level differences

We obtained sets of trait-associated SNPs for GWAS performed on seven different cohorts: UK Biobank, FINRISK, Chinese NIPT, Biobank Japan, APCDR and PAGE. Using the effect size estimates from each GWAS, we calculated population-wide polygenic scores for each of the 26 population panels from the 1000 Genome Project (The 1000 Genomes Project Consortium, 2015), using allele frequencies from each population panel. We then tested for overdispersion of these scores using the Q_X statistic, which was designed to detect deviations from neutral genetic drift affecting a set of trait-associated SNPs (Berg & Coop, 2014). We focused on 30 traits that were phenotyped in two or more cohorts, so that we could compare the P-value of this statistic using effect size estimates from at least two different cohorts (see Methods).

We applied the Q_X statistic to each of the 30 traits by selecting SNPs we deemed to be significantly associated with each trait . We used two different P-value cutoffs to select these SNPs: 1) a lenient cutoff, $P < 1e{-}5$ and 2) the standard genome-wide significance cutoff $P < 5e{-}8$. To verify that significant P-values of the Q_X statistics were not due to violations of the chi-squared distributional assumption, we also computed P-values using two randomization schemes: one is based on randomizing the effect size estimates of the trait-associated SNPs, while the other is based on using frequency-matched non-associated SNPs (see Methods). In general, we observe little notable differences in P-values when using the three schemes, although the sign-randomization scheme is sometimes inconsistent with the other two (Figures S3 and S4). The number of significant SNPs for each of the traits under the two cutoffs is shown in Figures S5 and S6.

After controlling for multiple testing ($P < 0.05/n$, where n is the number of assessed GWAS), we only find few traits with significant overdispersion in Q_X. Under the $P < 1e{-}5$ SNP-association cutoff, the only traits with significant overdispersion in at least one cohort are height and white blood cells (WBC) (Figures 1, S3). Potassium urine levels and mean corpuscular

hemoglobin (MCH) also result in significant values of Q_X when using the $P < 5e{-}8$ SNP-association cutoff (Figures S4,S7).

Figure 2 shows polygenic scores computed for each of the 1000 Genomes populations for height. In agreement with previous studies (Berg *et al.* , 2019; Sohail *et al.* , 2019), we observe that differences in polygenic height scores when using effect size estimates from the UKBB are greatly attenuated relative to differences in scores built when using estimates from GIANT. Extending this analysis across all datasets, we observe that PAGE scores are also over-dispersed, though in different directions than GIANT scores (Figure 2). Additionally, the observation that Europeans have very high polygenic scores when using GIANT effect size estimates cannot be replicated using any of the other GWAS estimates. Indeed, we only obtain significant Q_X P-values only for PAGE and GIANT, after multiple testing correction (using either association P-value threshold to select SNPs). The number of SNPs used for polygenic scores are shown in Table 3. The LD score regression ratio is substantially higher for PAGE and FINRISK than for the other cohorts (Tables 4, S3).

We also tested how the choice of the SNP association P-value threshold influenced the results. Sohail *et al.* (2019) showed that between-population differences in polygenic height scores grow stronger when using more lenient SNP association P-value cutoffs. However, one then runs the risk of including more variants that may be significantly associated due to uncorrected population stratification. We see there is a smaller score overdispersion when using the genome-wide significant SNPs, than when using the more lenient P-value cutoff (right column, Table 3 and Figure S8). We also looked in closer detail at other traits with evidence for significant overdispersion via the Q_X test. White blood cell counts (WBC), for example, shows strong overdispersion when using PAGE, but not when using the UKBB or BBJ effect size estimates (Figures S9, S10). We also observe a similar pattern when looking at mean corpuscular hemoglobin (MCH) scores (Figures S11, S12). In the case of urine levels of potassium, larger between-population differences are found in UKBB than in BBJ, when we use the stringent threshold (Figures S13, S14). In general, we observe that between-population differences in scores tend to be more similar between studies when using the stricter SNP-association P-value threshold, than when using the more lenient threshold.

Relationship between GWAS effect size estimates

To better understand where the differences in overdispersion of Q_X could stem from, we performed pairwise comparisons of the effect size estimates from the different GWAS. Since the UKBB GWAS is the GWAS with the largest number of individuals, we decided to compare the estimates from each of the other studies to the UKBB estimates. Here, we only focused on the 1,700 approximately-independent SNPs (the best tag SNP within each LD block). We began by only using SNPs that were classified as significant in UKBB using the lenient cutoff ($P < 1e{-}5$) (Figure 3). We observe that effect size estimates are correlated, as expected, but the strength of this correlation varies strongly across comparisons. UKBB vs. GIANT shows the highest correlation, while UKBB vs. APCDR shows the lowest. Importantly, those SNPs that also have a significant P-value in the non-UKBB GWAS in each comparison (colored in red in Figure 3) show a higher correlation than the rest of the SNPs.

The same analysis was carried out with SNPs classified as significant in each of the non-UKBB studies. The correlation of effect size estimates is generally lower (Figure S15), and a high percentage of SNPs deemed to be significant in the non-UKBB GWAS have effect size estimates approximately equal to zero in the UKBB GWAS (Figure S15). This pattern is stronger when we do not filter the 1,700 approximately independent SNPs by a particular SNP-association P-value cutoff (Figures S16 and S17).

We computed pairwise Pearson correlation coefficients between estimated effect sizes in the UKBB GWAS and each of the other GWAS (Table S4 when using SNPs that are significant in UKBB and Table S5 when using SNPs that are significant in the other GWAS). We observe that GWAS performed on individuals living geographically close to Britain have higher correlations to UKBB estimates than those that are performed on distant individuals. For instance, GIANT and FINRISK, both European-based GWAS, show high correlation in effect size estimates with UKBB (0.9958 and 0.790, respectively). In contrast, the GWAS based on an African panel, APCDR, shows an extremely low correlation in effect size estimates with UKBB (correlation coefficient = 0.087). This cohort has by far the smallest sample size of all the cohorts we analyzed (n = 4,778), which may explain the low correlation.

We also observe higher correlations when filtering for significantly trait-associated SNPs using either of the two SNP significance thresholds ($P < 1\text{e}{-5}$ and $P < 5\text{e}{-8}$) from the 1700 LD blocks. The sample size of the GWAS also affects the correlation in effect size estimates. We can see in Figure S18 that there is a positive relationship between the number of samples included in the non-UKBB GWAS and the Pearson correlation coefficients of the estimated effect sizes to the UKBB GWAS, which has the largest sample size.

Evidence for population stratification

Berg *et al.* (2019) looked for latent population stratification by studying the relation between allele frequency differences in two GWAS and their difference in effect size estimates. Presumably, if neither GWAS is affected by population stratification, there should not be a relation between these two variables. We plotted SNP differences in allele frequency between northern European and East Asian, African, and southern European samples (GBR, CHB, LWK and TSI subsets of 1000 Genomes, respectively) against the difference in effect size between a pair of GWAS. When comparing the UKBB and GIANT, we replicate the signal of correlation in differences between northern and southern European from Berg *et al.* (2019) ($P < 1e-5$, see Figure S19). This pattern is also observed in the GBR vs. CHB and GBR vs. LWK comparisons (Figures S20 and S21, panel A and B). However, these differences are not observed for any other pairwise GWAS comparisons (Figures S19, S22, S23 and S24).

Assessing different association methods

We find strong differences in the amount of polygenic score overdispersion across GWASs, but the GWASs we assessed were carried out using different association methods. We wanted to evaluate the effect of different association methods on the overdispersion of polygenic scores, while using the same underlying association cohort. We chose the UKBB cohort for this assessment, as it is the largest cohort among the ones we tested. We first split the UKBB cohort into three increasingly more expansive sets: 1) "British", 2) "White", and 3) "all ethnicities", based on a self-identified ethnicity classification carried out by the UKBB consortium. We then performed linear model (LM) and linear mixed model (LMM) association methods on each of the three sets of individuals (Table 2, see Methods). We also wanted to see if we could replicate the strong overdispersion in polygenic scores we saw in the meta-analysis cohorts, like GIANT, by partitioning the entire UKBB cohort into 75 cohorts (approximately emulating the number of cohorts in GIANT), and then performing a meta-analysis on the summary statistics obtained from individual GWASs performed separately on each of these cohorts (Table 2, see Methods).

The population-wide polygenic scores and the Q_X scores obtained using effect size estimates from each of these different methods are in Figures 4 and 5, respectively. There is an increased power to detect height-associated SNPs when we used the mixed model. Regardless of whether

one uses a linear or a linear mixed model, GWASs performed on a more expansive category of people ("all-ethnicities") lead to increased overdispersion of polygenic scores than when using more restrictive categories ("British" or "White"). Additionally, our artificial meta-analysis on the UKBB data resulted in even stronger overdispersion of the scores and, consequently, an even more strongly inflated Q_X statistic. Indeed, the most extreme Q_X-derived P-values across settings were those from the meta-analyses (Figure 5). This increased overdispersion is particularly evident when looking at the scores of European (FIN, CEU, GBR) and Latin American (PEL, CLM, MXL) populations in the meta-analysis setting (Figure 4).

Discussion

When looking for signals of polygenic adaptation based on population differentiation, we observe highly inconsistent signals depending on the GWAS cohort from which we obtained effect size estimates. Because we are using the exact same population panels to obtain population allele frequencies in all tests, the source of the inconsistencies must necessarily come from differences in the effect size estimates in the different GWAS. These inconsistencies are not limited to tests involving height-associated SNPs: they also appear in tests involving SNPs associated with other phenotypes, like white blood cell count, mean corpuscular hemoglobin and potassium levels in urine.

For those phenotypes for which we have estimates from more than two different cohorts, the meta-analysis GWASs - GIANT and PAGE - show strong overdispersion in genetic scores, and, consequently, strong evidence for selection when relying on the Q_X statistic. In biobank-based GWAS with cohorts made up of panels with relatively homogeneous ancestries, the signals of selection are generally (but not always) more attenuated. Furthermore, in the case of height, the distribution of genetic scores when using GIANT estimates and when using PAGE estimates are not consistent, suggesting differences in scores are likely not driven by a biological signal that is not being picked up by the biobank-based tests.

Furthermore, when we performed an artificial meta-analysis on the UKBB data, emulating the methodology of GIANT, we observed increased dispersion of polygenic scores among populations than when using single GWAS cohorts of more homogeneous ancestries, echoing findings by Kerminen *et al.* (2019) at a more localized geographic scale. This increase in score dispersion in turn causes an inflation of the Q_X statistic. Overall, this suggests that uncorrected population structure in GWAS meta-analyses may be a strong confounder for tests of polygenic adaptation based on patterns of population differentiation in polygenic scores.

Another possible cause for these inconsistencies could be differences in the number of individuals included in each GWAS, leading to differences in power to detect polygenic adaptation on trait-associated variants. Indeed, in some of the smaller cohorts (FINRISK and APCDR) we observe little to no evidence for strong deviations from neutrality in the distribution of genetic scores across populations.

We also find that the type of test performed to obtain Q_X P-values does not yield strong differences in such P-values, at least not of the magnitude observed when using effect size estimates from different GWAS cohorts. Those phenotypes and GWAS cohorts for which we find significant overdispersion via the chi-squared distributional assumption for the Q_X statistic also tend to be the ones for which we find significant overdispersion when not relying on it. This suggests that this assumption - while not entirely accurate (Berg & Coop, 2014) - is still reasonably valid, across all the phenotypes we looked at, assuming the effect size estimates are not affected by stratification.

When focusing on height, which is the phenotype for which we have summary statistics from

the largest number of GWAS cohorts, we were only able to detect evidence for uncorrected European stratification in GIANT (Berg *et al.* , 2019), i.e. a strong correlation between cross-population allele frequencies and differences in effect size estimates between cohorts. However Haworth *et al.* (2019), Novembre & Barton (2018) and Rosenberg *et al.* (2019) encourage caution about the interpretation of signals of polygenic adaptations due to the presence of residual stratification even in GWAS panels with no clear evidence for stratification.

In future studies of polygenic adaptation, we recommend the use of large homogeneous data sets and the verification of signals of polygenic adaptation in multiple GWAS cohorts (e.g. (Chen *et al.* , 2019)). We also recommend caution even when finding that statistics testing against neutrality are significant in multiple GWAS cohorts: it is still possible that all the GWAS cohorts may be affected by subtle stratification, possibly along different axes of population structure. To try to avoid stratification issues, recent studies have proposed to look for evidence for polygenic adaptation within the same panel that was used to obtain SNP effect size estimates, i.e. avoiding comparisons between populations that might be made up of individuals outside of the GWAS used to obtain effect size estimates, e.g. (Liu *et al.* , 2018b). However, Mostafavi *et al.* (2019) recently showed that the accuracy of polygenic scores often depends on the age and sex composition of the GWAS study participants, even when studying individuals of roughly similar ancestries within a single cohort, due to heritability differences along these axes of variation. This implies that ancestry-based stratification is not the only confounder that researchers of polygenic adaptation should be aware of when trying to detect polygenic adaptation. Approaches based on tree sequence reconstructions along the genome (Speidel *et al.* , 2019; Stern *et al.* , 2020) appear to be a fruitful avenue of research towards the development of methods that can properly control for some of these confounders.

Overall, we generally urge caution in the interpretation of signals of polygenic adaptation based on human GWAS data, at least until we have robust generative models that can explain exactly how stratification is creeping into these tests (Young *et al.* , 2019). Due to the high risk of misappropriation of this type of results by hate groups (Harmon, 2018), we also recommend that researchers make an effort to explain the caveats and problems associated with these tests in their publications (Novembre & Barton, 2018; Rosenberg *et al.* , 2019), as well as the strong sensitivity of their performance to the input datasets that we choose to feed into them.

Acknowledgments

We thank Jeremy Berg for helpful comments on the manuscript, as well as Mark Daly, Samuli Ripatti, Jukka Koskela, Masahiro Kanai, Yukinori Okada, Yoichiro Kamatani, Evan Irving-Pease and Graham Gower for general advice and technical assistance in obtaining and handling association data. We would also like to thank the members of the Racimo group for feedback throughout the duration of the project. Finally, we thank all participants of the biobanks and association studies included in this work, for their valuable contribution to the study of genetic variation and disease. Access to the UK Biobank data was provided to AA via application number 32683 and to ARM via application number 31063. FR is supported by a Villum Fonden Young Investigator award (project no. 00025300). AR-M is supported by the Lundbeck Foundation GeoGenetics Centre grant (R302-2018-2155) and a Novo Nordisk grant to the GeoGenetics Centre (NNF18SA0035006). ARM is supported by funding from the National Institutes of Health (K99MH117229).

References

Allen, Hana Lango, Estrada, Karol, Lettre, Guillaume, Berndt, Sonja I, Weedon, Michael N, Rivadeneira, Fernando, Willer, Cristen J, Jackson, Anne U, Vedantam, Sailaja, Raychaudhuri, Soumya, *et al.* . 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**(7317), 832.

Allison, David B, Heo, Moonseong, Kaplan, Norman, & Martin, Eden R. 1999. Sibling-based tests of linkage and association for quantitative traits. *The American Journal of Human Genetics*, **64**(6), 1754–1764.

Berg, Jeremy J, & Coop, Graham. 2014. A population genetic signal of polygenic adaptation. *PLoS genetics*, **10**(8), e1004412.

Berg, Jeremy J, Zhang, Xinjun, & Coop, Graham. 2017. Polygenic adaptation has impacted multiple anthropometric traits. *BioRxiv*, 167551.

Berg, Jeremy J, Harpak, Arbel, Sinnott-Armstrong, Nasa, Joergensen, Anja Moltke, Mostafavi, Hakhamanesh, Field, Yair, Boyle, Evan August, Zhang, Xinjun, Racimo, Fernando, Pritchard, Jonathan K, *et al.* . 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*, **8**, e39725.

Berisa, Tomaz, & Pickrell, Joseph K. 2016. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**(2), 283.

Borodulin, Katja, Tolonen, Hanna, Jousilahti, Pekka, Jula, Antti, Juolevi, Anne, Koskinen, Seppo, Kuulasmaa, Kari, Laatikainen, Tiina, Männistö, Satu, Peltonen, Markku, *et al.* . 2018. Cohort profile: the National FINRISK study. *International journal of epidemiology*, **47**(3), 696–696i.

Bycroft, Clare, Freeman, Colin, Petkova, Desislava, Band, Gavin, Elliott, Lloyd T, Sharp, Kevin, Motyer, Allan, Vukcevic, Damjan, Delaneau, Olivier, O'Connell, Jared, *et al.* . 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203.

Carlson, Christopher S. 2016. Diversity is future for genetic analysis. *Nature*, **540**(7633), 341–341.

Chang, Christopher C, Chow, Carson C, Tellier, Laurent CAM, Vattikuti, Shashaank, Purcell, Shaun M, & Lee, James J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**(1), s13742–015.

Chen, Minhui, Sidore, Carlo, Akiyama, Masato, Ishigaki, Kazuyoshi, Kamatani, Yoichiro, Schlessinger, David, Cucca, Francesco, Okada, Yukinori, & Chiang, Charleston WK. 2019. Evidence of polygenic adaptation at height-associated loci in mainland Europeans and Sardinians. *bioRxiv*, 776377.

Consortium, Genomes Project, Auton, A, & Brooks, LD. 2015. A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Durvasula, Arun, & Lohmueller, Kirk E. 2019. Negative selection on complex traits limits genetic risk prediction accuracy between populations. *bioRxiv*, 721936.

Fisher, Ronald Aylmer, *et al.* . 1918. The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399–433.

Guo, Jing, Yang, Jian, & Visscher, Peter M. 2018. Leveraging GWAS for complex traits to detect signatures of natural selection in humans. *Current opinion in genetics & development*, **53**, 9–14.

Harmon, Amy. 2018. Why White Supremacists Are Chugging Milk (and Why Geneticists Are Alarmed). *New York Times*, **17**.

Haworth, Simon, Mitchell, Ruth, Corbin, Laura, Wade, Kaitlin H, Dudding, Tom, Budu-Aggrey, Ashley, Carslake, David, Hemani, Gibran, Paternoster, Lavinia, Smith, George Davey, *et al.* . 2019. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature communications*, **10**(1), 333.

Hayward, Laura K, & Sella, Guy. 2019. Polygenic adaptation after a sudden change in environment. *bioRxiv*, 792952.

Heckerman, David, Gurdasani, Deepti, Kadie, Carl, Pomilla, Cristina, Carstensen, Tommy, Martin, Hilary, Ekoru, Kenneth, Nsubuga, Rebecca N, Ssenyomo, Gerald, Kamali, Anatoli, *et al.* . 2016. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, **113**(27), 7377–7382.

Hirata, Makoto, Kamatani, Yoichiro, Nagai, Akiko, Kiyohara, Yutaka, Ninomiya, Toshiharu, Tamakoshi, Akiko, Yamagata, Zentaro, Kubo, Michiaki, Muto, Kaori, Mushiroda, Taisei, *et al.* . 2017. Cross-sectional analysis of BioBank Japan clinical data: a large cohort of 200,000 patients with 47 common diseases. *Journal of epidemiology*, **27**(Supplement_III), S9–S21.

Kanai, Masahiro, Akiyama, Masato, Takahashi, Atsushi, Matoba, Nana, Momozawa, Yukihide, Ikeda, Masashi, Iwata, Nakao, Ikegawa, Shiro, Hirata, Makoto, Matsuda, Koichi, *et al.* . 2018. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature genetics*, **50**(3), 390–400.

Kerminen, Sini, Martin, Alicia R, Koskela, Jukka, Ruotsalainen, Sanni E, Havulinna, Aki S, Surakka, Ida, Palotie, Aarno, Perola, Markus, Salomaa, Veikko, Daly, Mark J, *et al.* . 2019. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *The American Journal of Human Genetics*, **104**(6), 1169–1181.

Liu, Siyang, Huang, Shujia, Chen, Fang, Zhao, Lijian, Yuan, Yuying, Francis, Stephen Starko, Fang, Lin, Li, Zilong, Lin, Long, Liu, Rong, *et al.* . 2018a. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell*, **175**(2), 347–359.

Liu, Xuanyao, Loh, Po-Ru, O'Connor, Luke J, Gazal, Steven, Schoech, Armin, Maier, Robert M, Patterson, Nick, & Price, Alkes L. 2018b. Quantification of genetic components of population differentiation in UK Biobank traits reveals signals of polygenic selection. *bioRxiv*, 357483.

Locke, Adam E, Kahali, Bratati, Berndt, Sonja I, Justice, Anne E, Pers, Tune H, Day, Felix R, Powell, Corey, Vedantam, Sailaja, Buchkovich, Martin L, Yang, Jian, *et al.* . 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**(7538), 197.

Loh, Po-Ru, Kichaev, Gleb, Gazal, Steven, Schoech, Armin P, & Price, Alkes L. 2018. Mixed-model association for biobank-scale datasets. *Nature genetics*, **50**(7), 906–908.

Martin, Alicia R, Gignoux, Christopher R, Walters, Raymond K, Wojcik, Genevieve L, Neale, Benjamin M, Gravel, Simon, Daly, Mark J, Bustamante, Carlos D, & Kenny, Eimear E. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, **100**(4), 635–649.

Martin, Alicia R, Kanai, Masahiro, Kamatani, Yoichiro, Okada, Yukinori, Neale, Benjamin M, & Daly, Mark J. 2019. Current clinical use of polygenic scores will risk exacerbating health disparities. *BioRxiv*, 441261.

Matise, Tara C, Ambite, Jose Luis, Buyske, Steven, Carlson, Christopher S, Cole, Shelley A, Crawford, Dana C, Haiman, Christopher A, Heiss, Gerardo, Kooperberg, Charles, Marchand, Loic Le, *et al.* . 2011. The Next PAGE in understanding complex traits: design for the analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *American journal of epidemiology*, **174**(7), 849–859.

Mostafavi, Hakhamanesh, Harpak, Arbel, Conley, Dalton, Pritchard, Jonathan K, & Przeworski, Molly. 2019. Variable prediction accuracy of polygenic scores within an ancestry group. *BioRxiv*, 629949.

Nagai, Akiko, Hirata, Makoto, Kamatani, Yoichiro, Muto, Kaori, Matsuda, Koichi, Kiyohara, Yutaka, Ninomiya, Toshiharu, Tamakoshi, Akiko, Yamagata, Zentaro, Mushiroda, Taisei, *et al.* . 2017. Overview of the BioBank Japan Project: study design and profile. *Journal of epidemiology*, **27**(Supplement_III), S2–S8.

Novembre, John, & Barton, Nicholas H. 2018. Tread lightly interpreting polygenic tests of selection. *Genetics*, **208**(4), 1351.

Popejoy, A. B., & Fullerton, S. M. 2016. Genomics is failing on diversity. *Nature*, **538**(7624), 632–637.

Pritchard, Jonathan K, Pickrell, Joseph K, & Coop, Graham. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology*, **20**(4), R208–R215.

Racimo, Fernando, Berg, Jeremy J, & Pickrell, Joseph K. 2018. Detecting polygenic adaptation in admixture graphs. *Genetics*, genetics–300489.

Ragsdale, Aaron P, Nelson, Dominic, Gravel, Simon, & Kelleher, Jerome. 2020. Lessons learned from bugs in models of human history. *bioRxiv*.

Renaud, Gabriel. 2017. glactools: a command-line toolset for the management of genotype likelihoods and allele counts. *Bioinformatics*, **34**(8), 1398–1400.

Robinson, Matthew R, Hemani, Gibran, Medina-Gomez, Carolina, Mezzavilla, Massimo, Esko, Tonu, Shakhbazov, Konstantin, Powell, Joseph E, Vinkhuyzen, Anna, Berndt, Sonja I, Gustafsson, Stefan, *et al.* . 2015. Population genetic differentiation of height and body mass index across Europe. *Nature genetics*, **47**(11), 1357.

Rosenberg, Noah A, Edge, Michael D, Pritchard, Jonathan K, & Feldman, Marcus W. 2019. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. *Evolution, medicine, and public health*, **2019**(1), 26–34.

Sella, G, & Barton, NH. 2019. Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies. *Annual review of genomics and human genetics*.

Sohail, Mashaal, Maier, Robert M, Ganna, Andrea, Bloemendal, Alex, Martin, Alicia R, Turchin, Michael C, Chiang, Charleston WK, Hirschhorn, Joel, Daly, Mark J, Patterson, Nick, *et al.* . 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, **8**, e39702.

Speidel, Leo, Forest, Marie, Shi, Sinan, & Myers, Simon R. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, **51**(9), 1321–1329.

Stern, Aaron J, Speidel, Leo, Zaitlen, Noah A, & Nielsen, Rasmus. 2020. Disentangling selection on genetically correlated polygenic traits using whole-genome genealogies. *bioRxiv*.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*, **526**(7571), 68.

Turchin, Michael C, Chiang, Charleston WK, Palmer, Cameron D, Sankararaman, Sriram, Reich, David, Hirschhorn, Joel N, of ANthropometric Traits (GIANT) Consortium, Genetic Investigation, *et al.* . 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature genetics*, **44**(9), 1015.

Turelli, Michael. 2017. Commentary: Fisher's infinitesimal model: A story for the ages. *Theoretical population biology*, **118**, 46–49.

Visscher, Peter M, Brown, Matthew A, McCarthy, Mark I, & Yang, Jian. 2012. Five years of GWAS discovery. *The American Journal of Human Genetics*, **90**(1), 7–24.

Willer, Cristen J, Li, Yun, & Abecasis, Gonçalo R. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**(17), 2190–2191.

Wojcik, Genevieve L, Graff, Mariaelisa, Nishimura, Katherine K, Tao, Ran, Haessler, Jeffrey, Gignoux, Christopher R, Highland, Heather M, Patel, Yesha M, Sorokin, Elena P, Avery, Christy L, *et al.* . 2019. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 1.

Wood, Andrew R, Esko, Tonu, Yang, Jian, Vedantam, Sailaja, Pers, Tune H, Gustafsson, Stefan, Chu, Audrey Y, Estrada, Karol, Luan, Jian'an, Kutalik, Zoltán, *et al.* . 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, **46**(11), 1173.

Young, Alexander I, Benonisdottir, Stefania, Przeworski, Molly, & Kong, Augustine. 2019. Deconstructing the sources of genotype-phenotype associations in humans. *Science*, **365**(6460), 1396–1400.

Figures

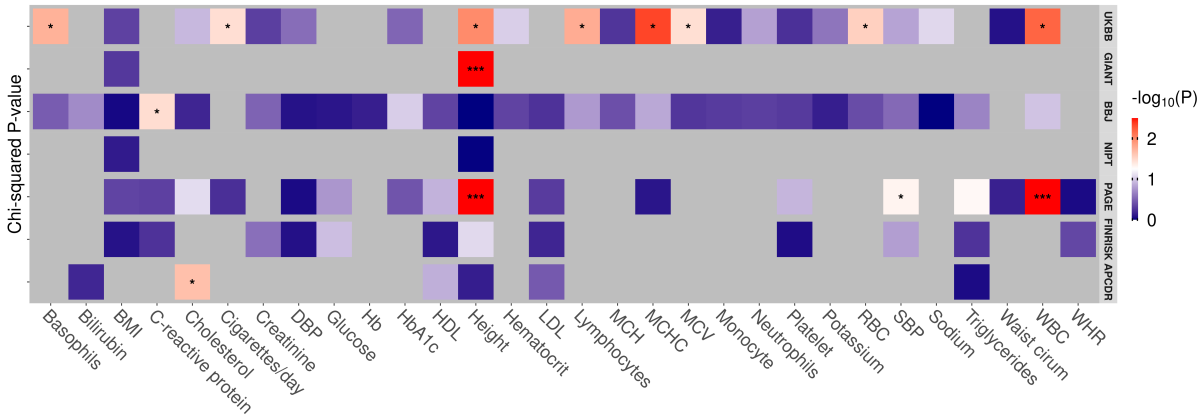

Figure 1: Each cell of this heatmap is the $-log_{10}(P-value)$ of a Q_X test using SNPs associated with different traits and effect size estimates obtained from different GWASs. Each row corresponds to one of the six GWASs we are evaluating and the columns correspond to the different traits for each GWAS that have SNPs with a lower P-value than $1e-5$. BMI, body mass index; DBP, Diastolic blood pressure; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; SBP, systolic blood pressure; WBC, white blood cell count; WHR, waist-to-hip ratio. Significance thresholds after Bonferroni correction: *** $<= 0.05/\text{m}$, ** $<= 0.05/\text{n}$, * $<= 0.05$, where n is the number of traits in each GWAS (row-dependent) and m is the total number of tests calculated, across all GWASs.

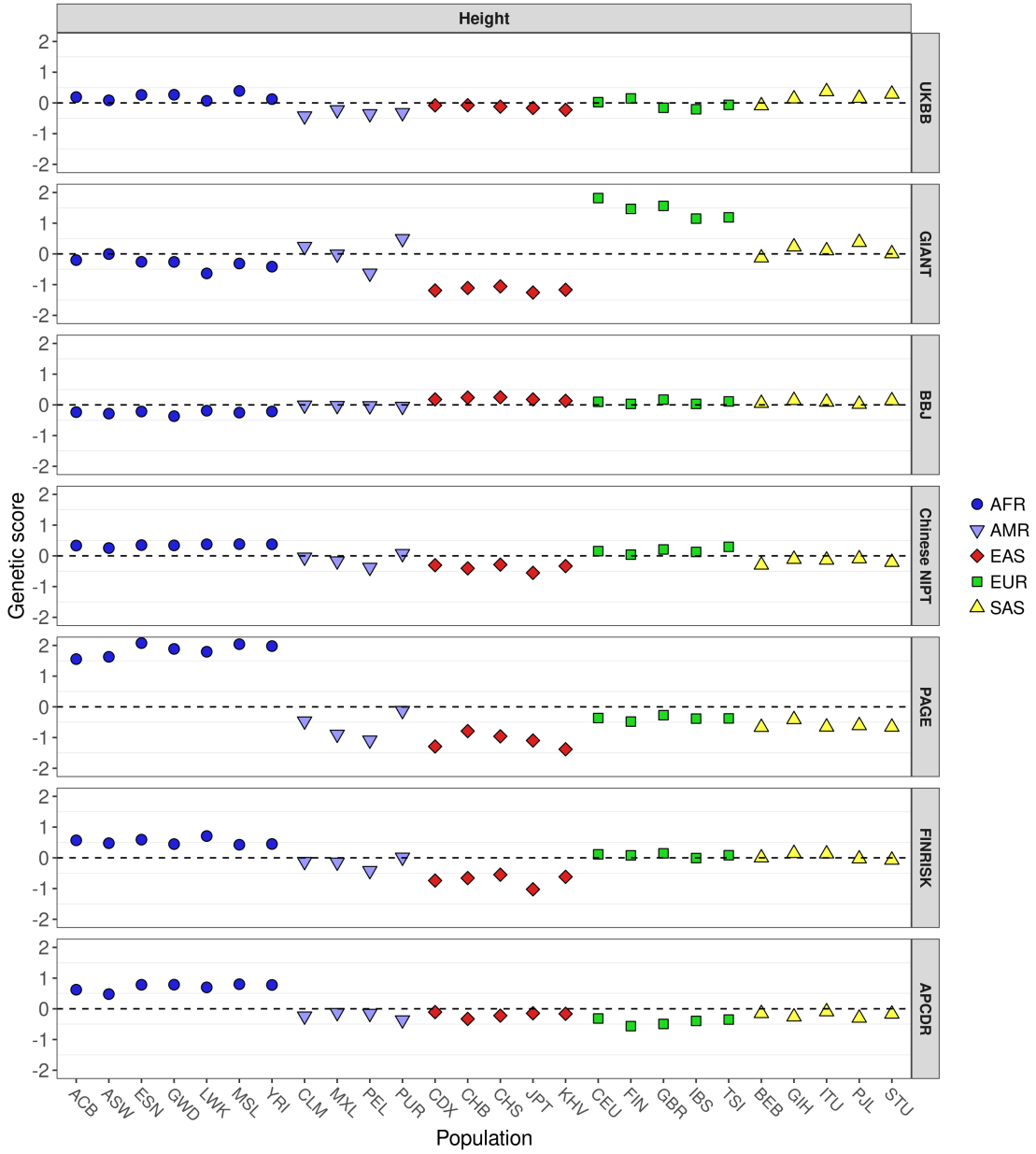

Figure 2: Polygenic scores for height using candidate SNPs with $P < 1e{-}5$. The 1000 Genomes Project populations colored by their super-population code. The corresponding number of trait-associated SNPs and the Q_X P-value for each GWAS are shown in top row, Table 3.

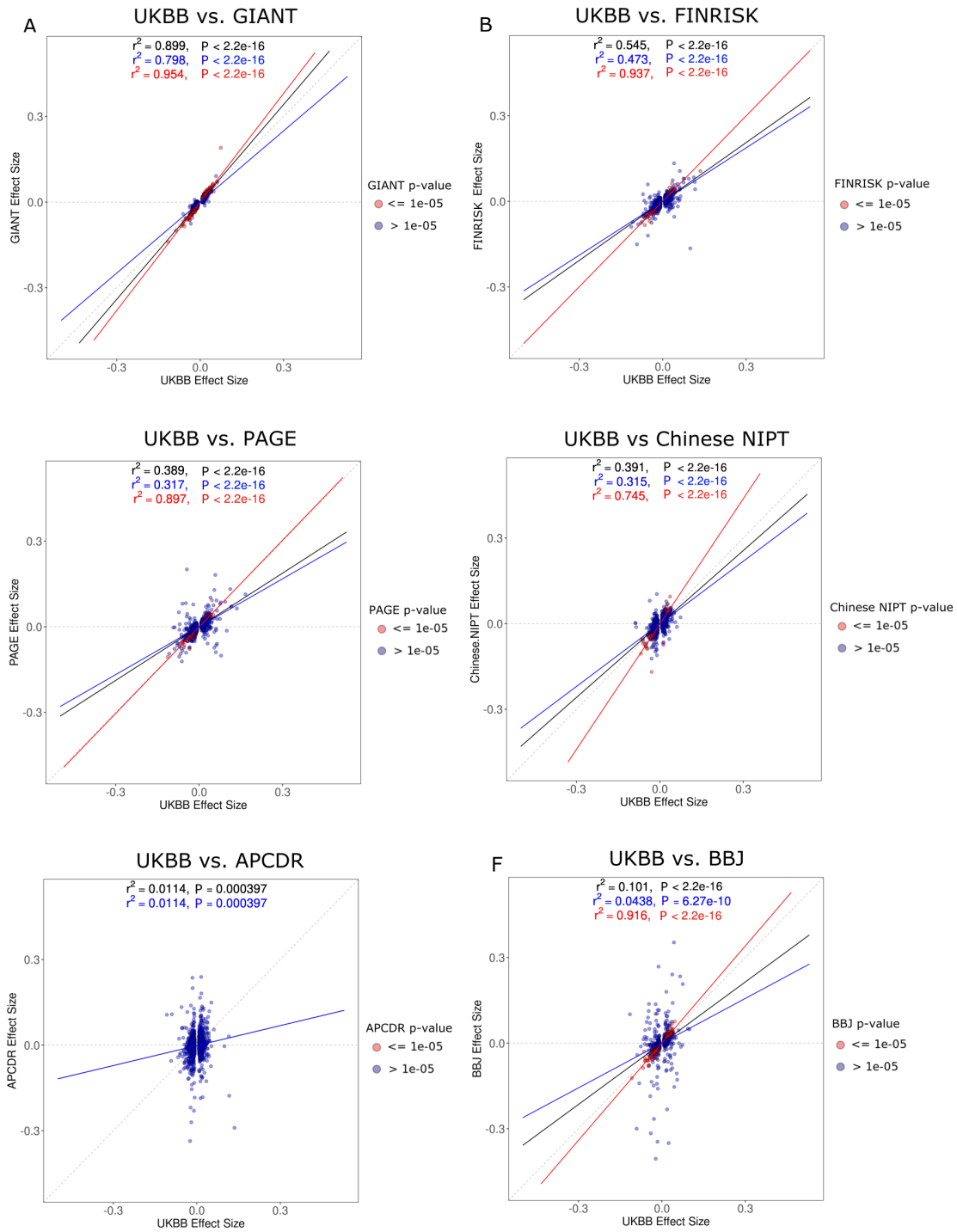

Figure 3: Regression of UKBB effect size estimates against non-UKBB effect size estimates, after filtering for SNPs with $P < 1e{-}5$ in the UKBB GWAS. The SNPs are colored based on their P-value in the non-UKBB GWAS, as are the corresponding regression lines (red: $P \leq 1e-5$; blue: $P > 1e-5$). The black regression line was obtained using all SNPs, regardless of their P-value in the non-UKBB GWAS.

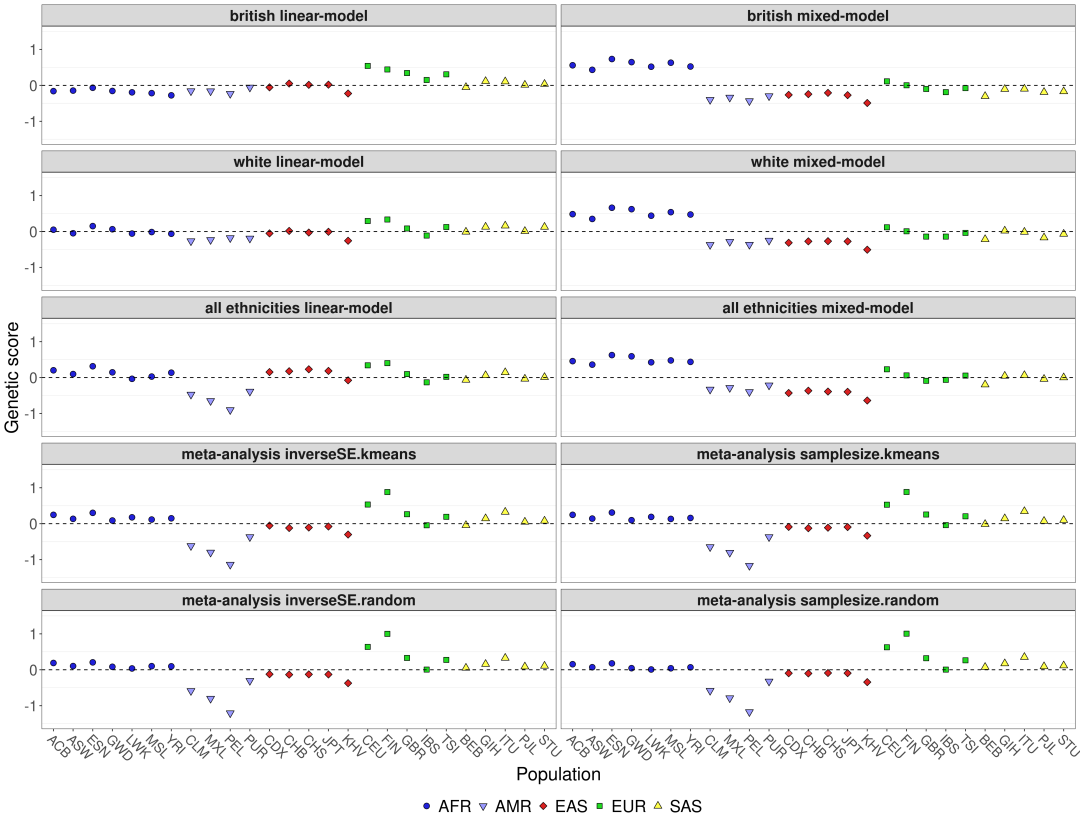

Figure 4: Polygenic scores for height in each of the 1000 Genomes population panels, using effect size estimates from UKBB, obtained via different types of association methods. The types of GWAS methods used are described in Table 2.

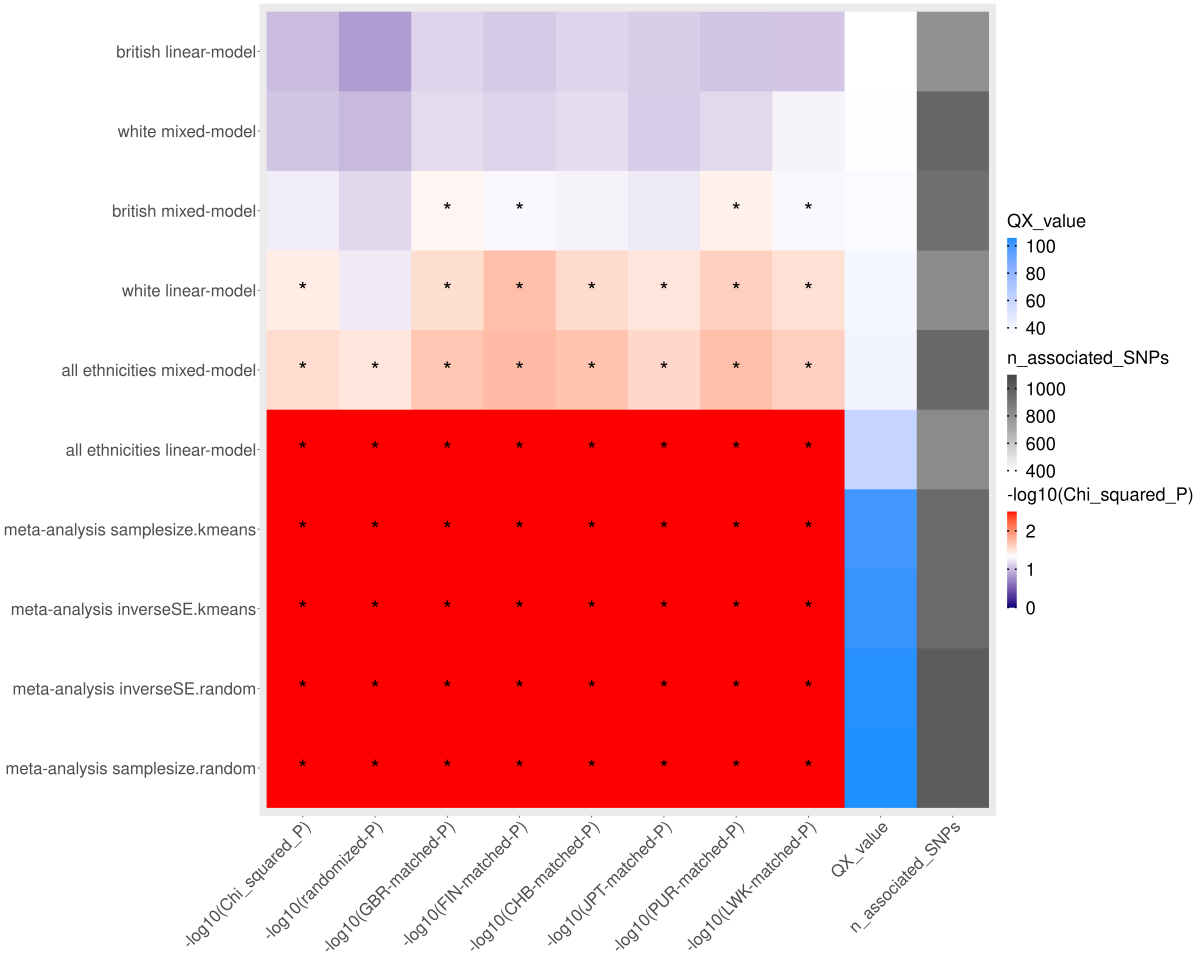

Figure 5: Q_X statistics and P-values for the trait height, obtained by using different types of association methods on the UKBB data. The asterisk denotes a significance cutoff for Q_X of $P < 0.05$. "-log10(Chi squared P)" = -log10(P), obtained assuming a chi-squared distribution for the Q_X statistic. "-log10(randomized P)" = -log10(P-value), obtained using the effect sign-randomization scheme. All other P-values were obtained by sampling random SNPs from the genome using the allele frequency matching scheme in different populations, as described in the Methods section. The different types of GWAS methods along the y-axis are described in Table 2.

Tables

Table 1: Information on genotyping, imputation and association methods for each GWAS analyzed in this study. LM+PCs = linear model with principal components as covariates. LMM = linear mixed model.

GWAS cohort	No. individuals	No. genotyped SNPs	No. imputed SNPs	Association method	GWAS array	Filters
UKBB	360388	820967	10,800,000	LM+PCs	UK Biobank Axiom Array	INFO > 0.8 MAF > 1e-4 HWE P-value > 1e-10
GIANT	253288	[meta-analysis]	2,834,208	LM+PCs	Affymetrix 500K platform, Illumina genotyping arrays and custom Perlegen arrays	Removed low quality SNPs
BBJ	159,095	958,497	27,896,057	LM+PCs	Illumina HumanOmniExpressExome	INFO >= 0.4 MAF > 1.0% HWE P-value >= 1e6
Chinese NIPT	61,321	NA	2,130,000	LM+PCs	Ultra low depth shotgun sequencing	INFO > 0.4 HWE P-value > 1e6
PAGE	49,781	1,748,250	39,723,562	LMM	Illumina Multi-Ethnic Genotyping Array (MEGA)	INFO > 0.4
FINRISK	24,725	551,004	11,670,715	LM+PCs	Illumina CoreExome	INFO > 0.7 MAF > 1.0% HWE P-value > 1e6
APCDR	4,778	2,382,209	16,477,797	LMM	Illumina Omni2.5	INFO >= 0.3 MAF > 0.5%

Table 2: Description of types of methods used to obtain SNP associations from UKBB.

Code	Cohort composition	Description of association method
british linear-model	Self-reported "British" individuals from UKBB	linear model implemented in PLINK 1.9
british mixed-model	Self-reported "British" individuals from UKBB	linear mixed model implemented in BOLT-LMM
white linear-model	Self-reported "White" individuals from UKBB	linear model implemented in PLINK 1.9
white mixed-model	Self-reported "White" individuals from UKBB	linear mixed model implemented in BOLT-LMM
all ethnicities linear-model	All self-reported ethnicities from UKBB	linear model implemented in PLINK 1.9
all ethnicities mixed-model	All self-reported ethnicities from UKBB	linear mixed model implemented in BOLT-LMM
meta-analysis.inverseSE	75 sub-cohorts of equal size, composed of randomly sampled individuals from UKBB (all ethnicities)	inverse-variance-based meta-analysis implemented in METAL
meta-analysis.inverseSE	75 sub-cohorts obtained from K-means clustering using 1st 3 PCs of UKBB individuals (all ethnicities)	inverse-variance-based meta-analysis implemented in METAL
meta-analysis.samplesize	75 sub-cohorts of equal size, composed of randomly sampled individuals from UKBB (all ethnicities)	sample-size-based meta-analysis implemented in METAL
meta-analysis.samplesize	75 sub-cohorts obtained from K-means clustering using 1st 3 PCs of UKBB individuals (all ethnicities)	sample-size-based meta-analysis implemented in METAL

Table 3: Difference in height Q_X scores within and between GWAS. The left panel corresponds to scores obtained when we use P-value $P < 1e{-}5$ threshold. The right panel corresponds to genome-wide significant threshold ($P < 5e{-}8$). The number of trait-associated SNPs used to compute the scores are shown for both cutoffs.

	Filter by LD block lenient P-value threshold (1e-5)		Filter by LD block lenient P-value threshold (5e−8)	
GWAS cohort	No. trait associated SNPs	Qx	No. trait associated SNPs	Qx
UKBB	1140	45,57 (P = 0.01)	810	39,72(P = 0.04)
GIANT	739	102,73 (P = 4e-3)	468	69,18 (P = 8e-8)
BBJ	759	9,255 (P = 0.99)	431	35,00 (P = 0.11)
Chinese NIPT	164	12,82 (P = 0.99)	60	24,68(P = 0.54)
PAGE	528	77,31 (P = 5e-7)	84	49,52 (P = 4e-3)
FINRISK	209	36,43 (P = 0.08)	47	41,48 (P = 0.03)
APCDR	61	21,15 (P = 0.73)	0	NA

Table 4: SNP-based heritability and LD Score regression ratio and intercept estimates (with standard errors in parentheses) for height measured in different cohorts. LD scores were computed for the super-population panels in the 1000 Genomes Project. The APCDR heritability estimate is not shown because it was estimated to be negative, due to the small sample size of the cohort.

GWAS cohort	Genome-wide significant SNPs (P 5e−8)	Observed scale heritability (SE)	LD regression ratio (SE)	LD regression intercept (SE)	Super Population
UKBB	30891	0.4205 (0.0183)	0.1184 (0.0098)	1.4384 (0.0362)	EUR
GIANT	11625	0.3159 (0.0146)	0.1419 (0.0094)	1.2727 (0.0181)	EUR
BBJ	9976	0.4168 (0.019)	0.1142 (0.0128)	1.1829 (0.0204)	EAS
Chinese NIPT	1573	0.329 (0.0201)	0.0746 (0.0353)	1.0425 (0.0201)	EAS
PAGE	564	0.2578 (0.0237)	0.371 (0.0336)	1.1762 (0.016)	AMR
FINRISK	415	0.4557 (0.0446)	0.3775 (0.035)	1.1358 (0.0126)	EUR
APCDR	0	NA	NA	NA	AFR

Supplementary Figures

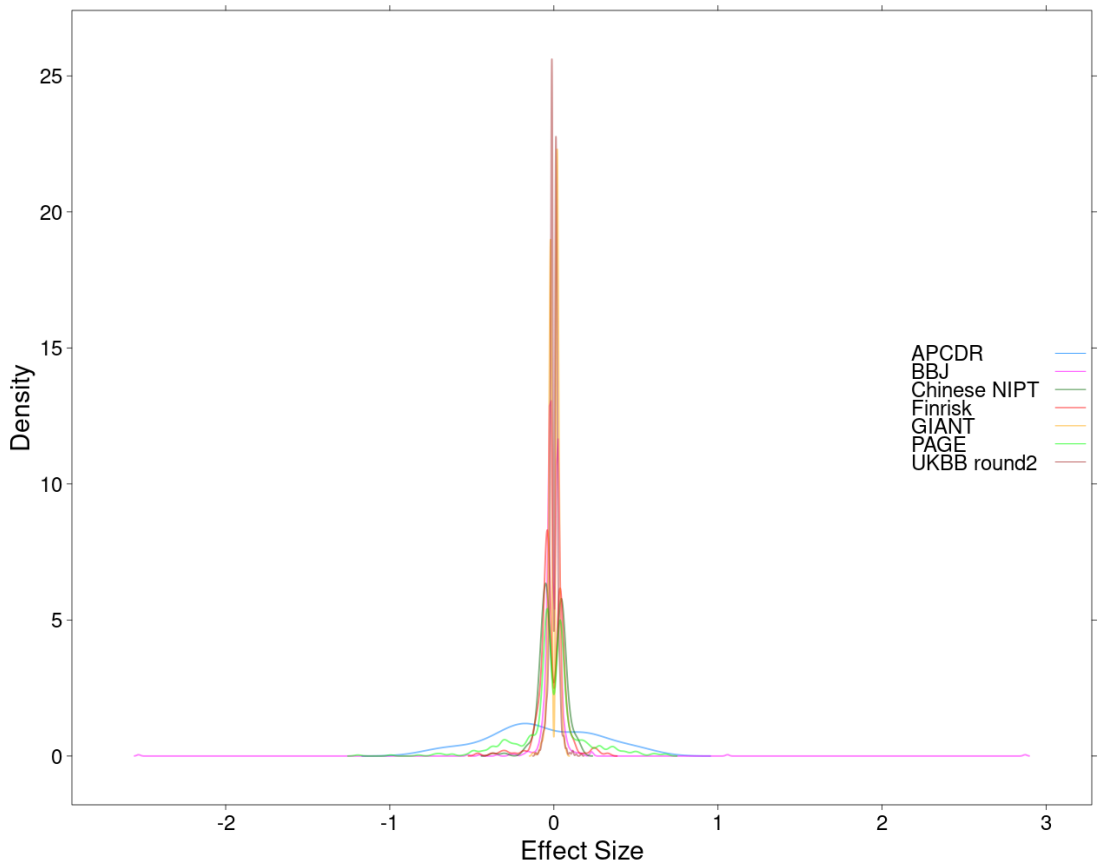

Figure S1: Distribution of the derived-allele effect size estimates for approximately independent height-associated SNPs with $P < 1e{-}5$.

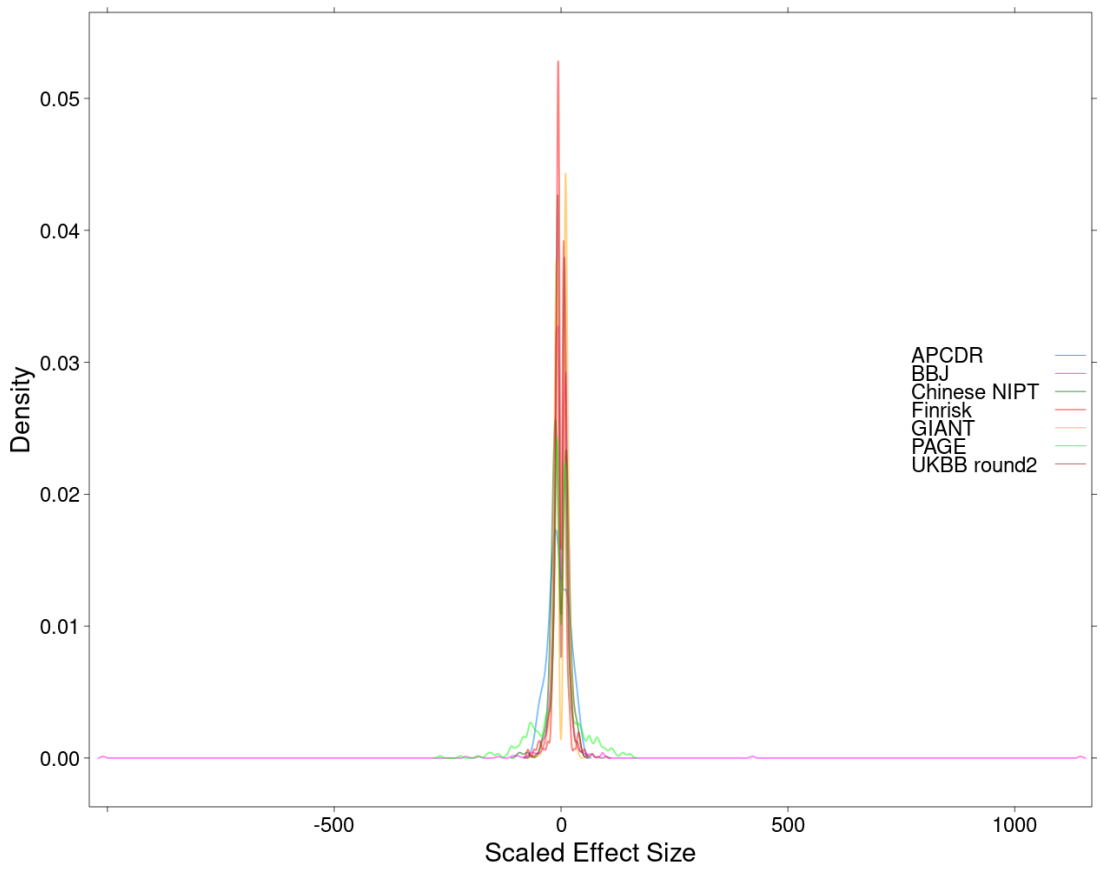

Figure S2: Distribution of the derived-allele effect size estimates scaled by the square root of the sample size of the study for approximately independent height-associated SNPs with $P < 1e{-}5$.

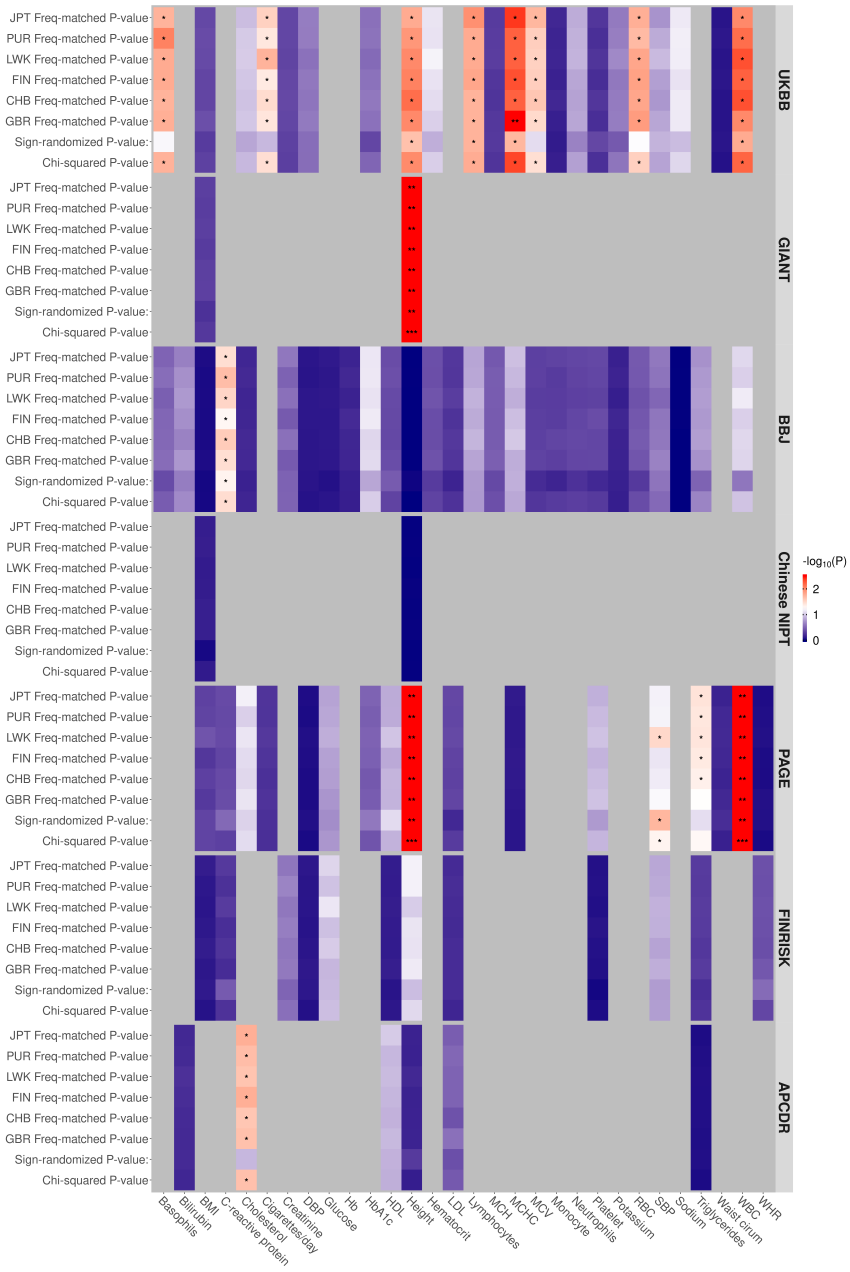

Figure S3: Evidence for polygenic adaptation on trait-associated SNPs ($log_{10}(P-value)$), using the Q_X test statistic. Each row in the heatmap corresponds to a specific GWAS cohort and a specific type of scheme to determine the significance of the Q_X statistic. The columns correspond to the different traits for each GWAS cohort that have SNPs with a P-value lower than 1e−5. "Freq-matched P-value" = P-value obtained by sampling SNPs with matching frequencies to the trait-associated SNPs in a particular cohort. "Sign-randomized P-value" = P-value obtained by randomizing the signs of the effect size estimates. "Chi-squared P-value" = P-value obtained by assuming the Q_X statistics has a chi-squared distribution.

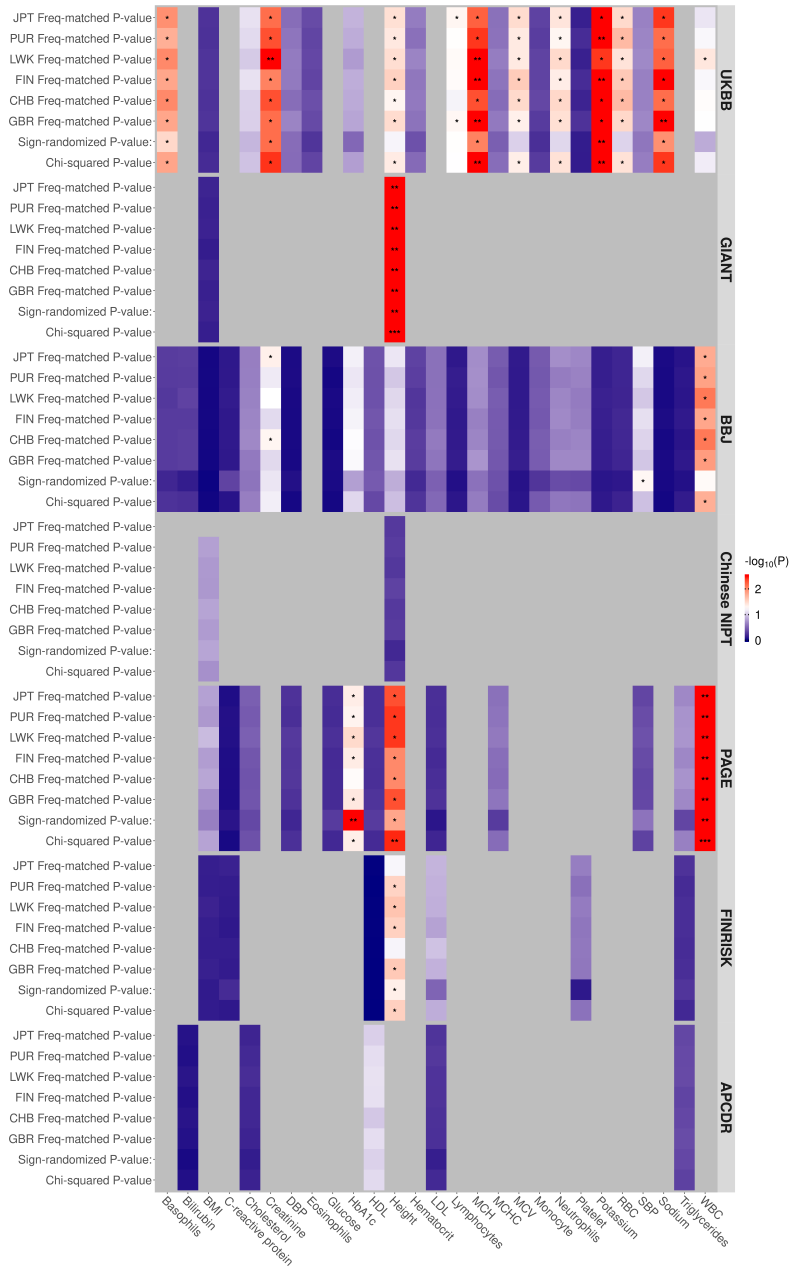

Figure S4: Evidence for polygenic adaptation on trait-associated SNPs ($log_{10}(P-value)$), using the Q_X test statistic. Each row in the heatmap corresponds to a specific GWAS cohort and a specific type of scheme to determine the significance of the Q_X statistic. The columns correspond to the different traits for each GWAS cohort that have SNPs with a P-value lower than 5e−8. "Freq-matched P-value" = P-value obtained by sampling SNPs with matching frequencies to the trait-associated SNPs in a particular cohort. "Sign-randomized P-value" = P-value obtained by randomizing the signs of the effect size estimates. "Chi-squared P-value" = P-value obtained by assuming the Q_X statistics has a chi-squared distribution.

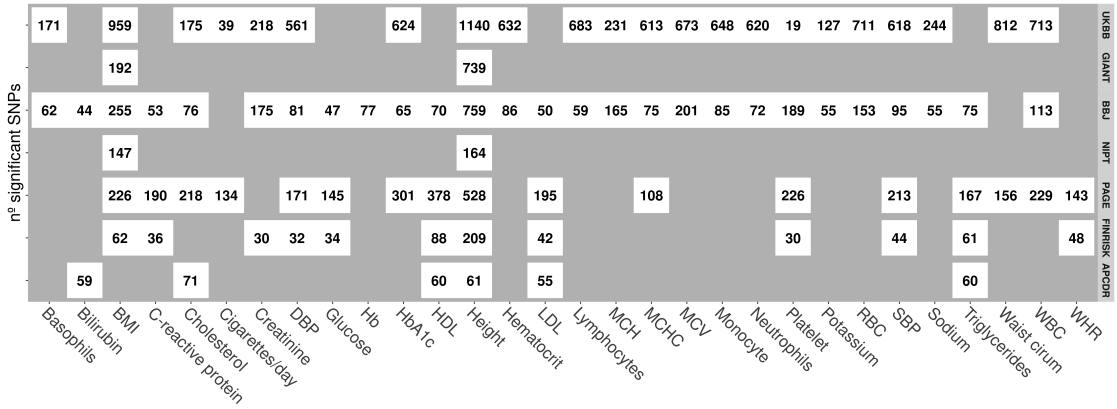

Figure S5: Number of significant SNPs used to compute the Q_X statistics for each cell of Figure 1.

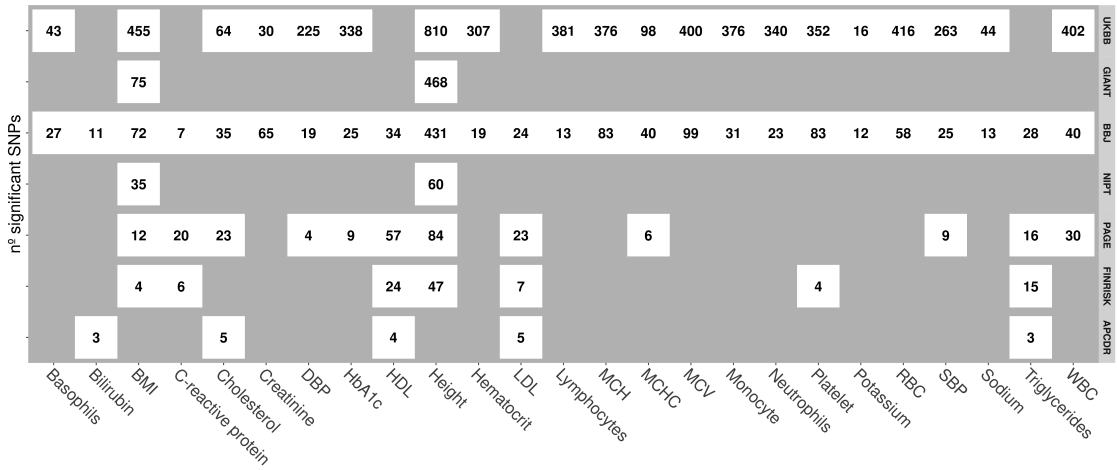

Figure S6: Number of significant SNPs for each trait in each GWAS with $P < 5e{-}8$.

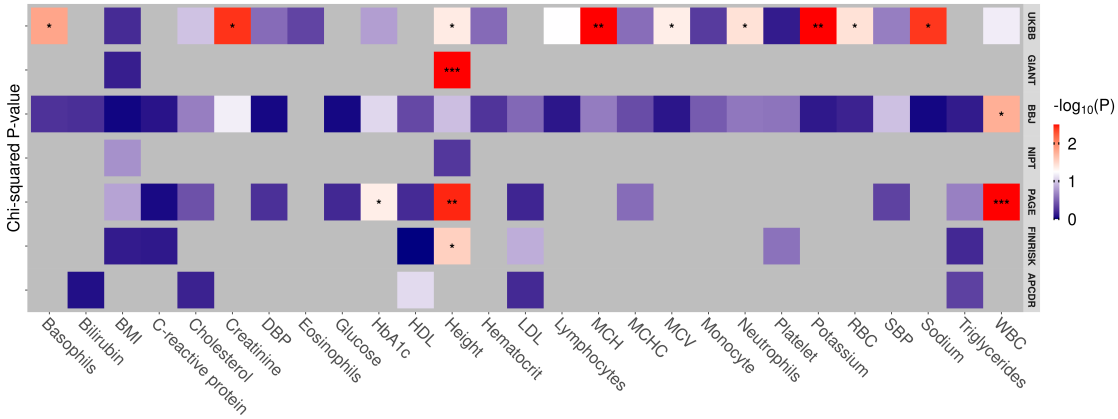

Figure S7: Evidence for polygenic adaptation on trait-associated SNPs ($log_{10}(P-value)$), using the Q_X test statistic. Each row in the heatmap corresponds to one of the six GWAS cohorts we are evaluating and the columns correspond to the different traits for each GWAS cohort that have SNPs with a P-value lower than 5e−8. BMI, body mass index; DBP, Diastolic blood pressure; HbA1c, glycated hemoglobin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MCV, mean corpuscular volume; RBC, red blood cell count; SBP, systolic blood pressure; WBC, white blood cell count; WHR, waist-to-hip ratio. Significance thresholds after Bonferroni corrections: *** denotes P <= 0.05/m, ** denotes P <= 0.05/n, * denotes P <= 0.05 where n is the number of traits measured in each GWAS (row-dependent) and m is the total number of tests calculated, across all GWAS.

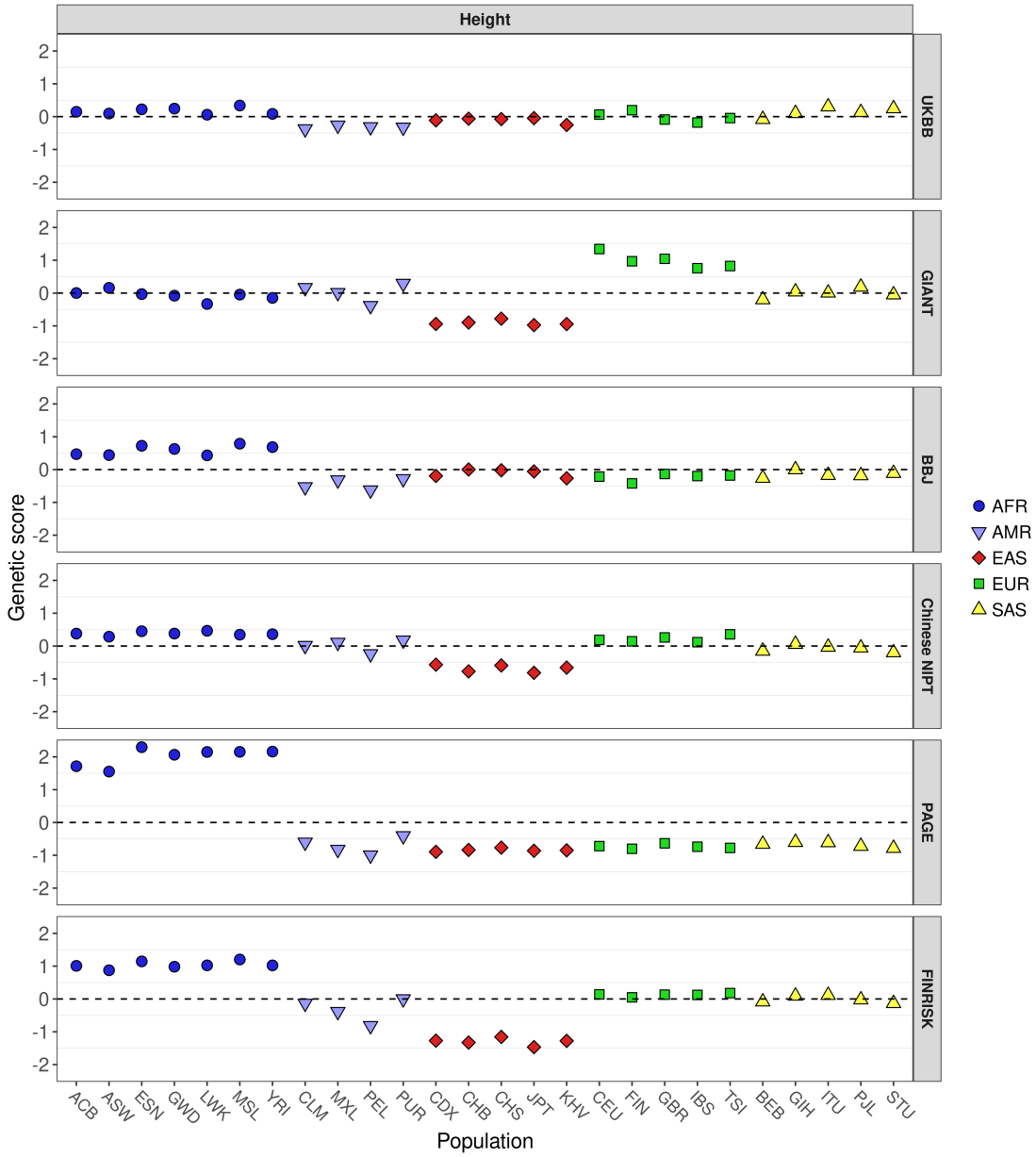

Figure S8: Polygenic scores for height using candidate SNPs with $P < 5e{-}8$ in 1000 Genome populations colored by their super-population code. The corresponding number of trait-associated SNPs and the Q_X *P-value* for each GWAS are shown in bottom row, Table 3.

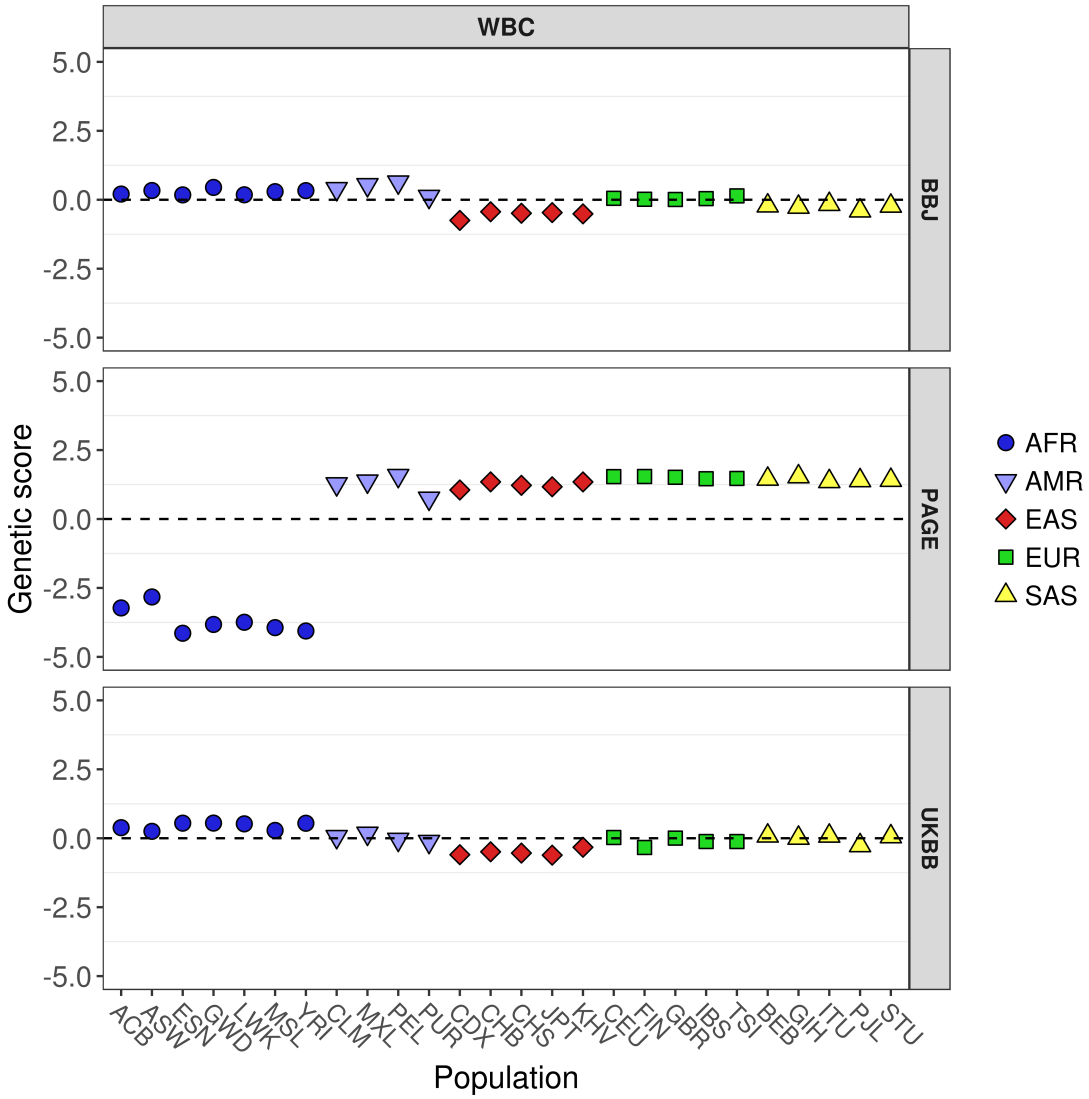

Figure S9: Polygenic scores for white blood cell counts using candidate SNPs with $P < 1e{-}5$. The 1000 Genomes Project populations colored by their super-population code.

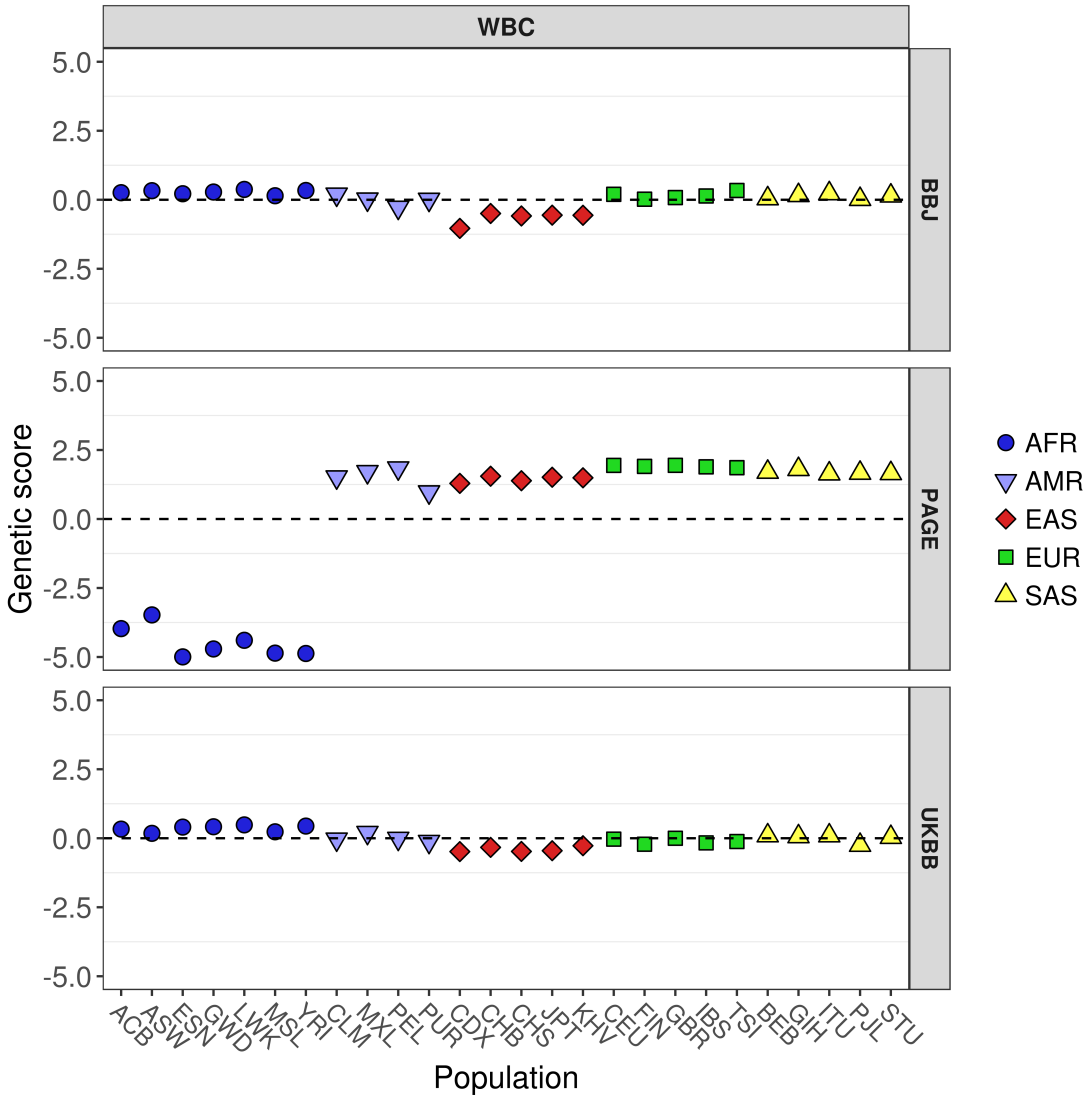

Figure S10: Polygenic scores for white blood cell counts using candidate SNPs with $P < 5e{-}8$ in 1000 Genome populations colored by their super-population code.

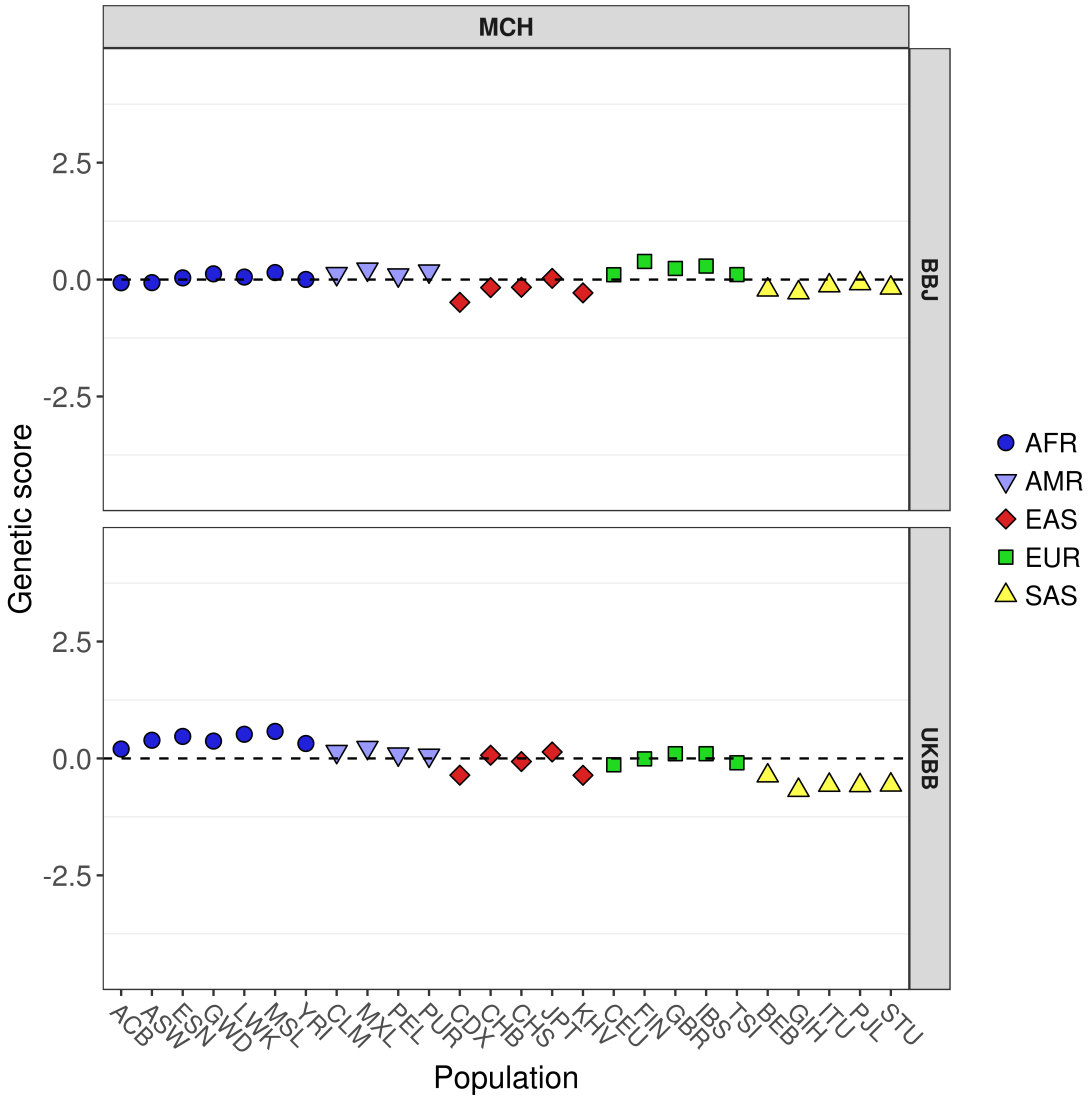

Figure S11: Polygenic scores for mean corpuscular hemoglobin using candidate SNPs with $P <$ 1e$-$5 in 1000 Genome populations colored by their super-population code. X trait-associated SNPs.

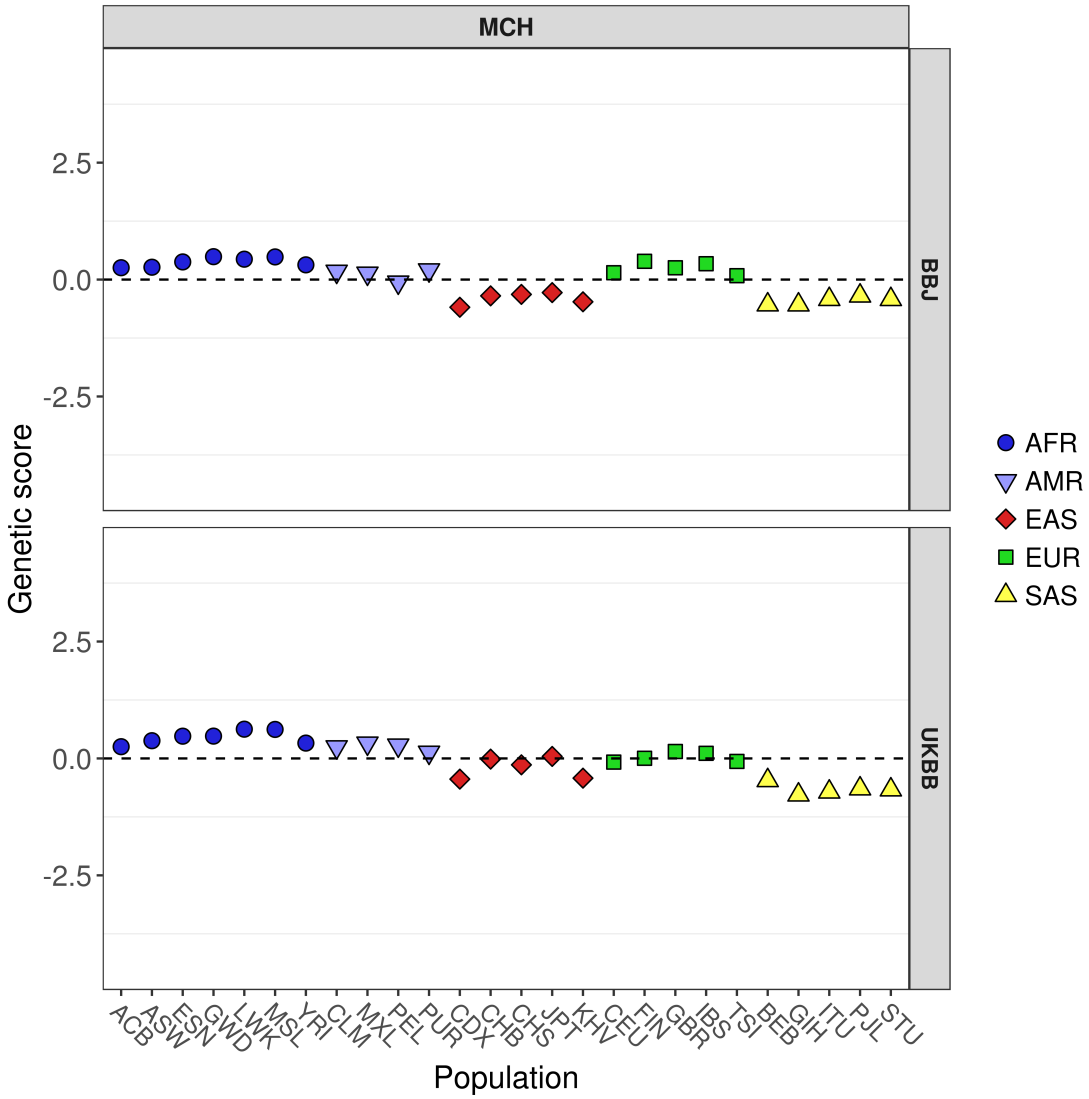

Figure S12: Polygenic scores for mean corpuscular hemoglobin using candidate SNPs with $P <$ 5e−8 in 1000 Genome populations colored by their super-population code.

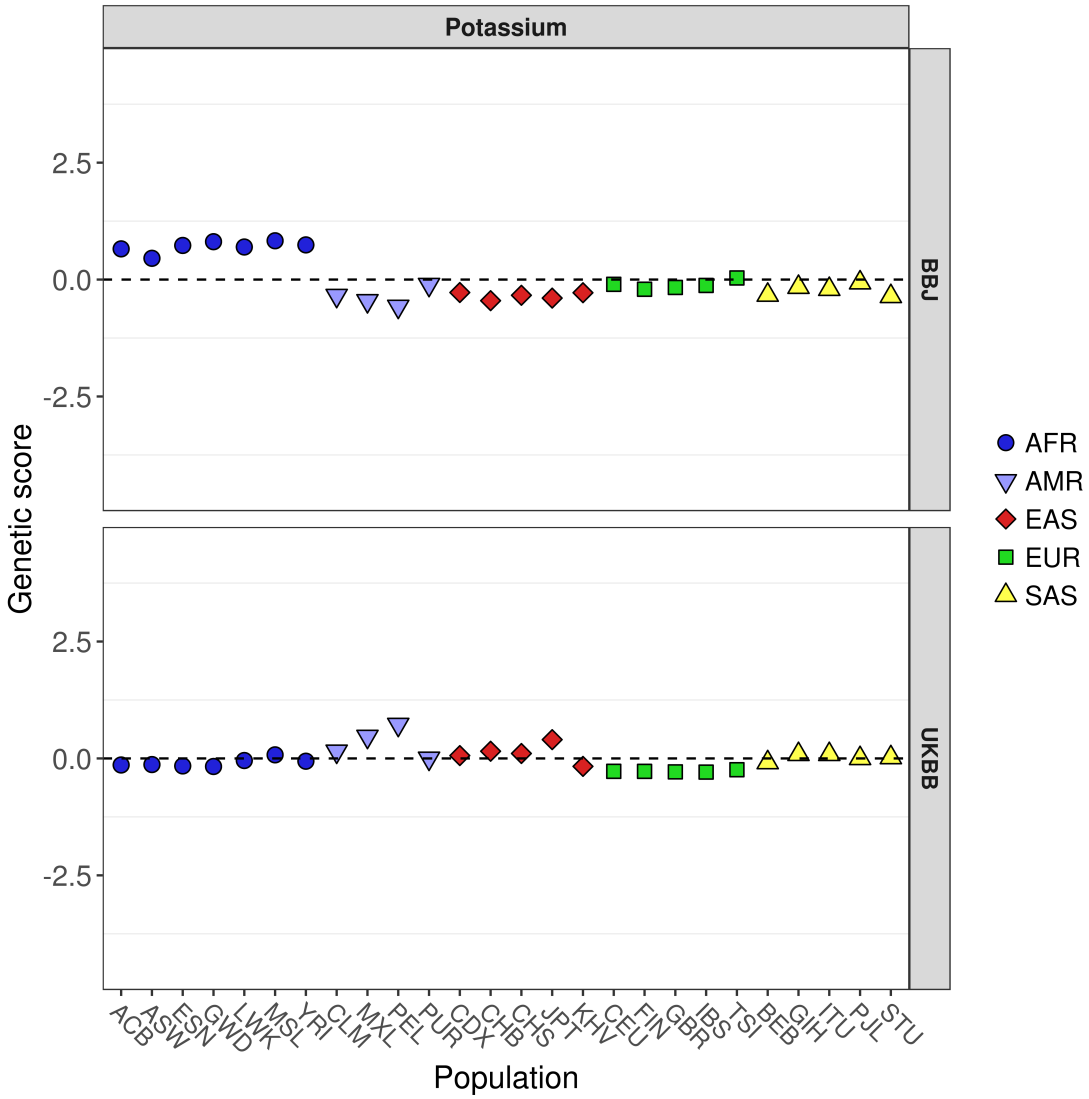

Figure S13: Polygenic scores for potassium level using candidate SNPs with $P < 1e{-}5$ in 1000 Genome populations colored by their super-population code.

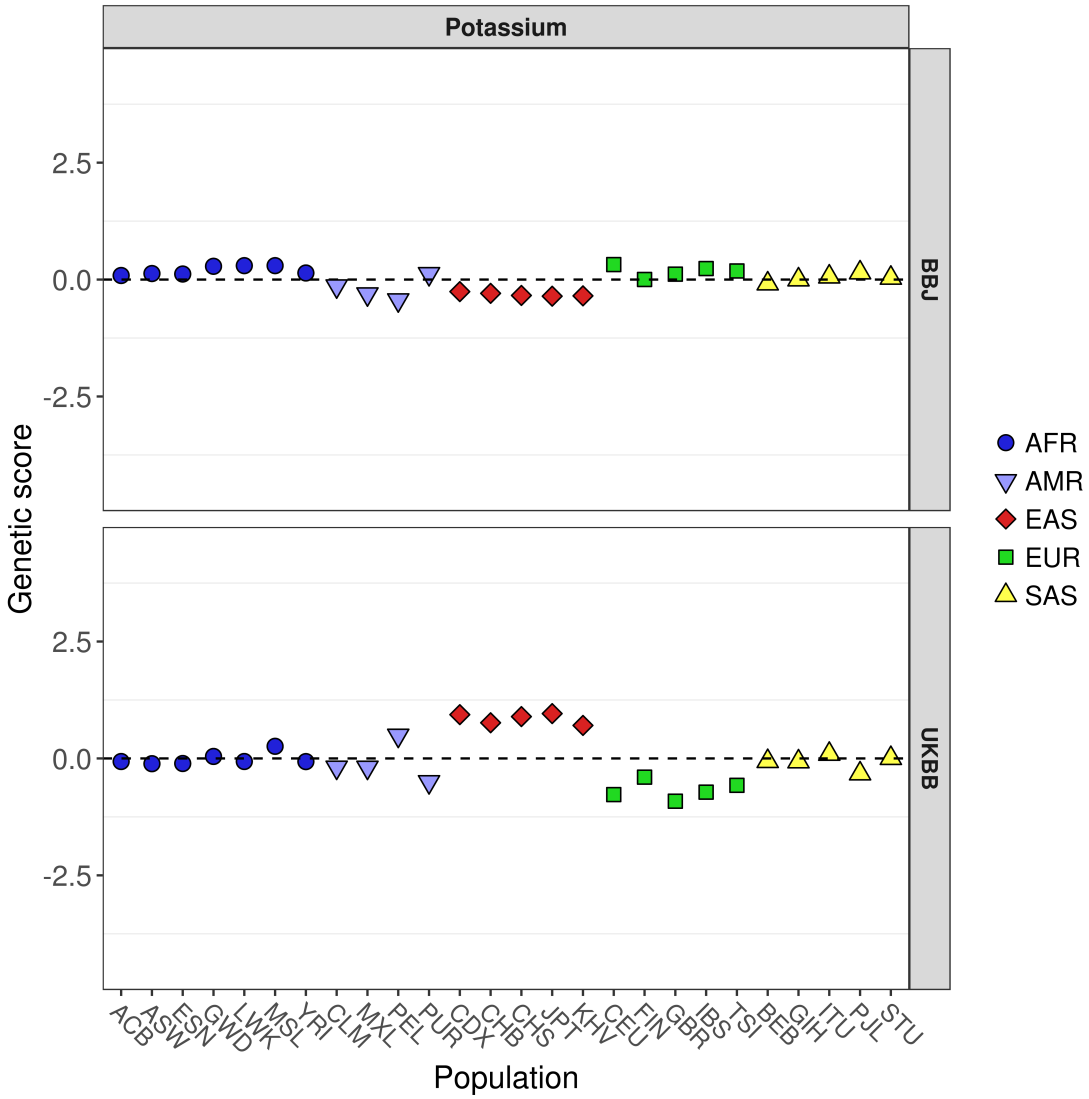

Figure S14: Polygenic scores for potassium level using candidate SNPs with $P < 5e{-}8$ in 1000 Genome populations colored by their super-population code.

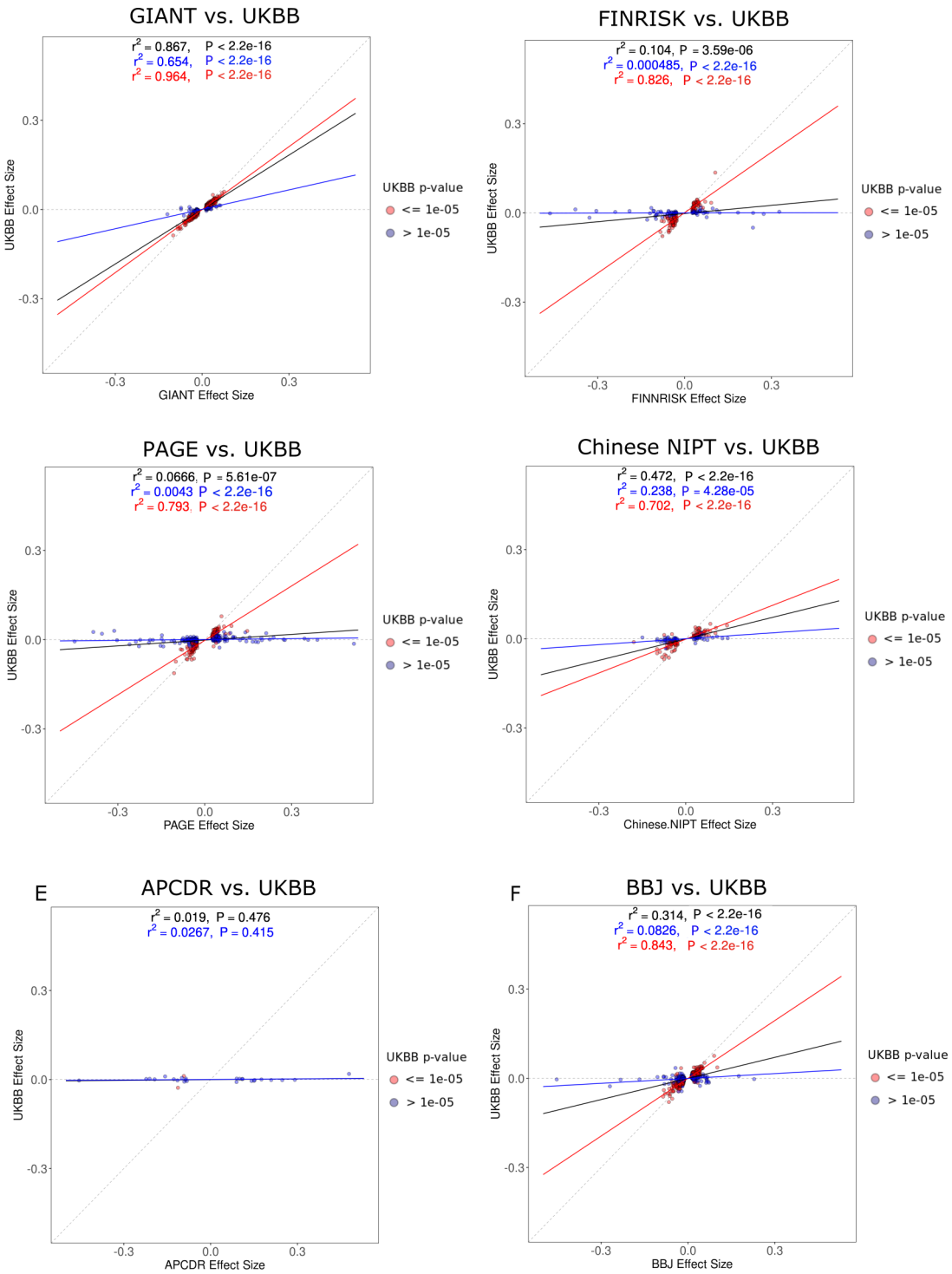

Figure S15: Regression of UKBB effect size estimates against non-UKBB effect size estimates, after filtering for SNPs with $P < 1e{-}5$ in the non-UKBB GWAS. The SNPs are colored based on their P-value in the UKBB, as are the corresponding regression lines (red: $P \leq 1e-5$; blue: $P > 1e-5$). The black regression line was obtained using all SNPs, regardless of their P-value in the UKBB GWAS.

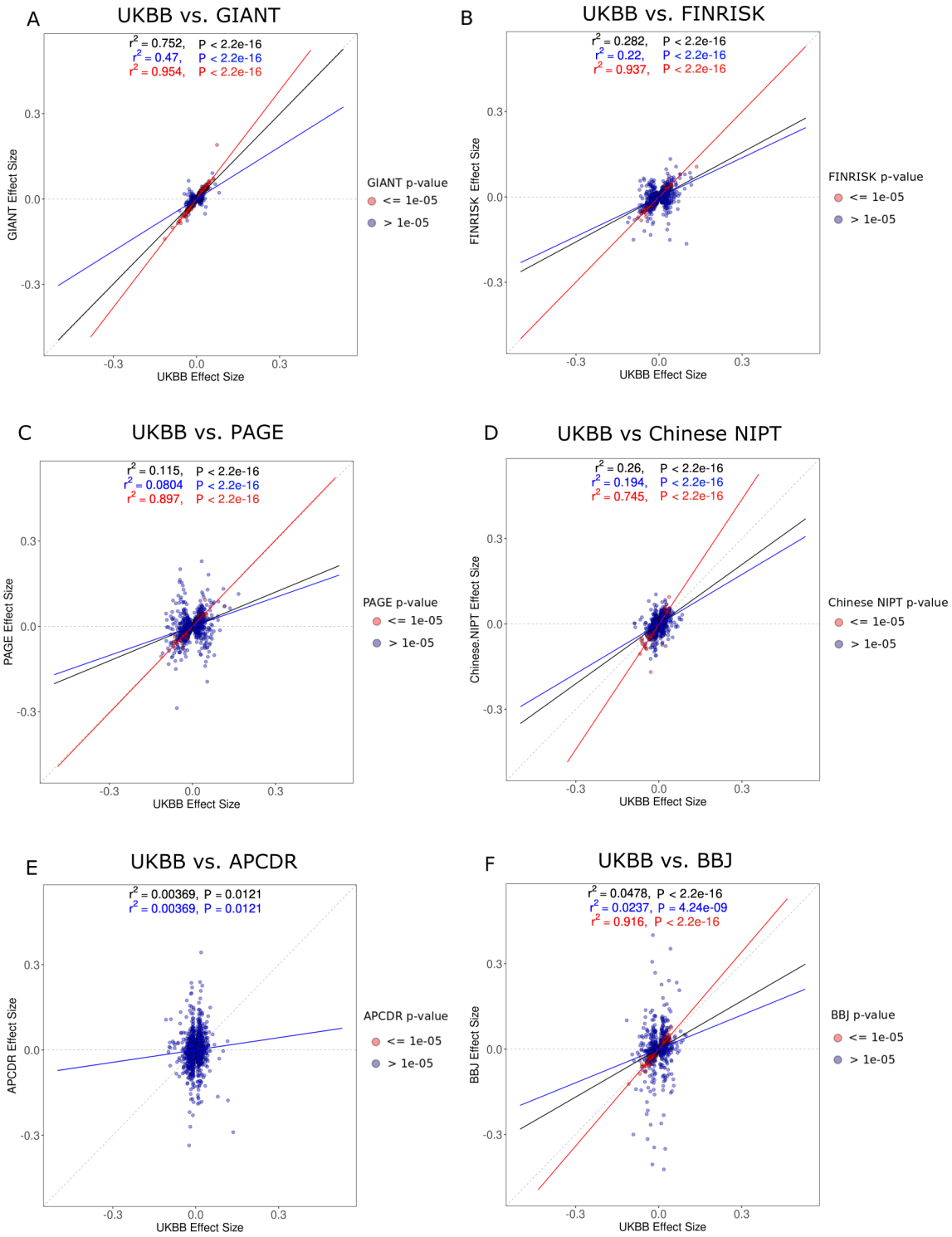

Figure S16: Regression of non-UKBB effect size estimates against UKBB effect size estimates, after filtering for the SNPs with the lowest UKBB P-values in 1,700 approximately-independent LD blocks. The SNPs are colored based on their P-value in the non-UKBB GWAS, as are the corresponding regression lines (red: $P \leq 1e-5$; blue: $P > 1e-5$). The black regression line was obtained using all SNPs, regardless of their P-value in the non-UKBB GWAS.

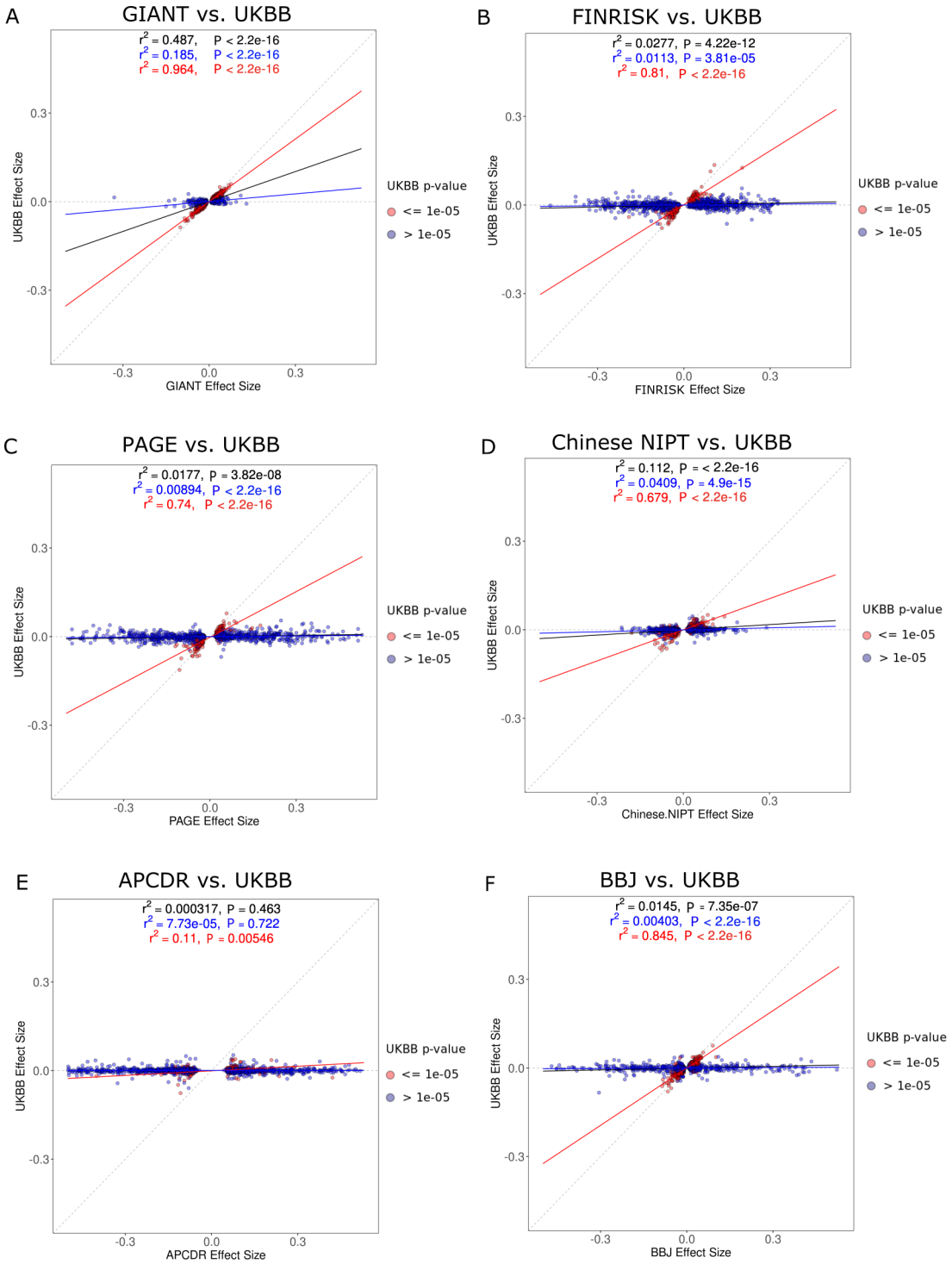

Figure S17: Regression of UKBB effect size estimates against non-UKBB effect size estimates, after filtering for the SNPs with the lowest non-UKBB P-values in 1,700 approximately-independent LD blocks. The SNPs are colored based on their P-value in the UKBB GWAS, as are the corresponding regression lines (red: $P \leq 1e-5$; blue: $P > 1e-5$). The black regression line was obtained using all SNPs, regardless of their P-value in the UKBB GWAS.

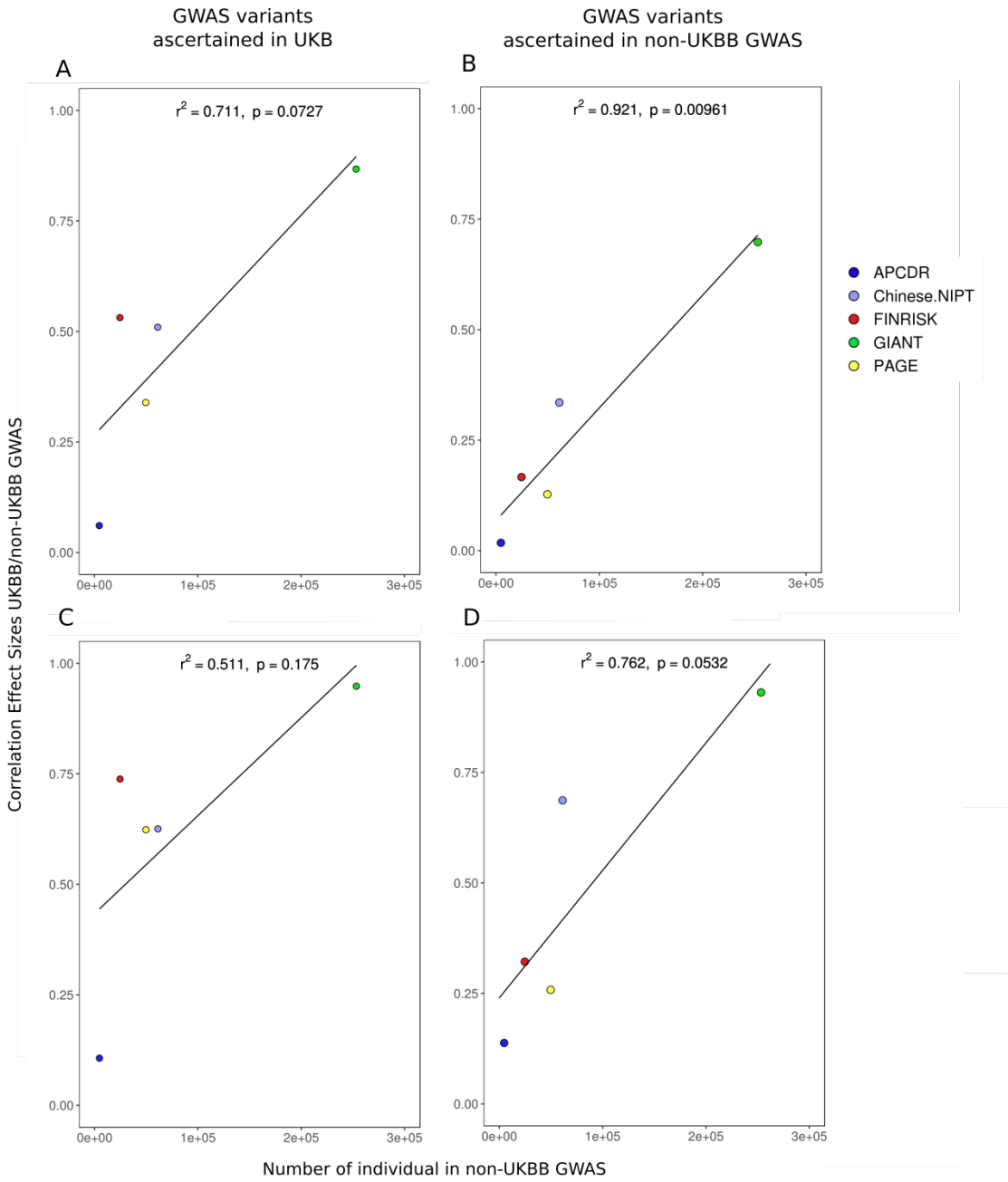

Figure S18: Correlation between the number of individuals in a GWAS and the correlation of effect size estimates between two different GWAS. In all panels, we compare UKBB against one of the other studies. **A and B**. Coefficients obtained when we include the SNPs from the 1,700 LD blocks. **C and D**. Coefficients obtained when we filter out those SNPs with a P-value above the threshold ($P > 1e-5$). There is a strong correlation between the number of samples and the correlation of the effect size when they are replicated in an alternative GWAS.

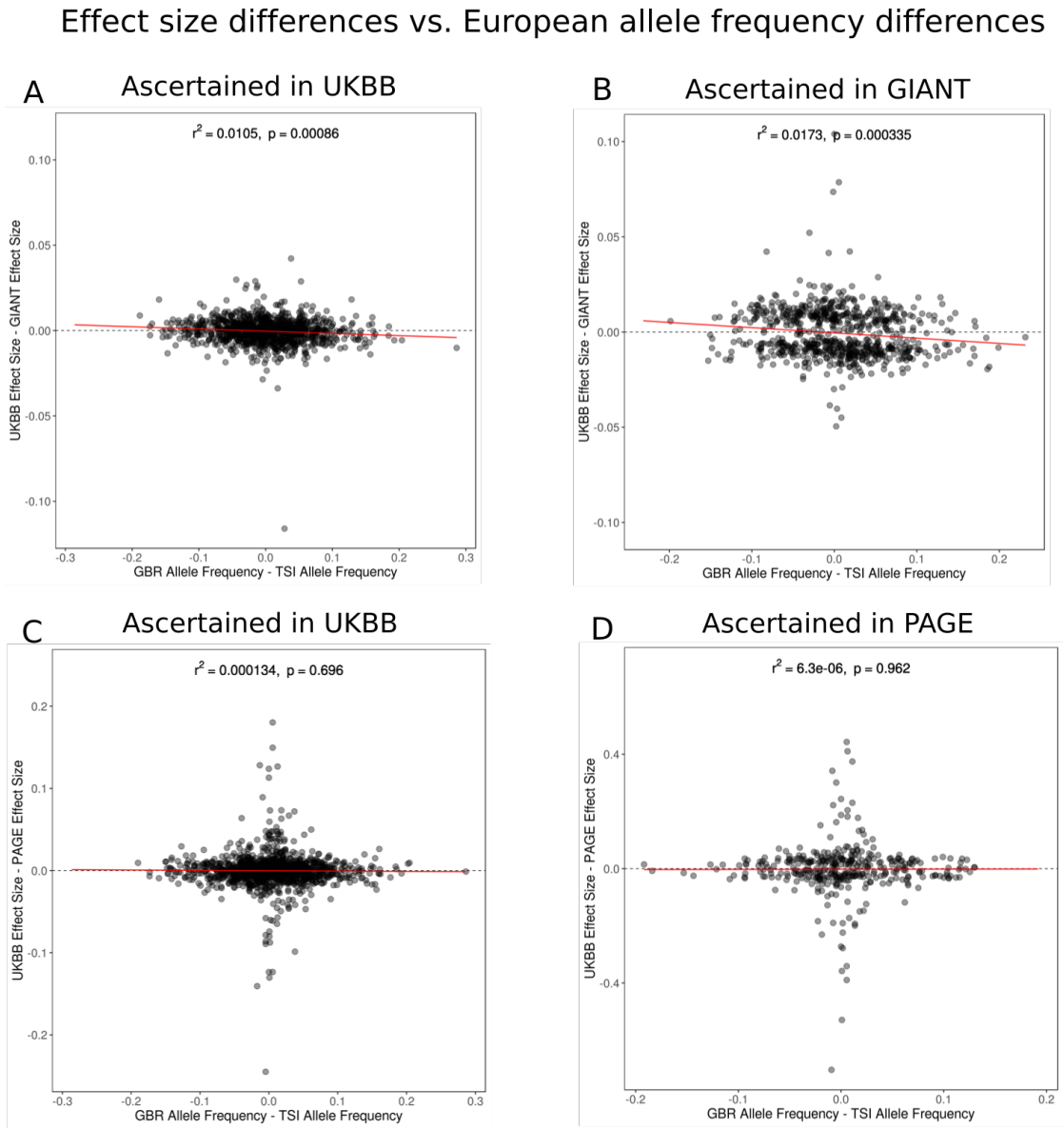

Figure S19: Regression of GWAS effect size differences against allele frequency differences between northern and southern European population panels (GBR and TSI). We selected SNPs for this analysis that had $P < 1e{-}5$.

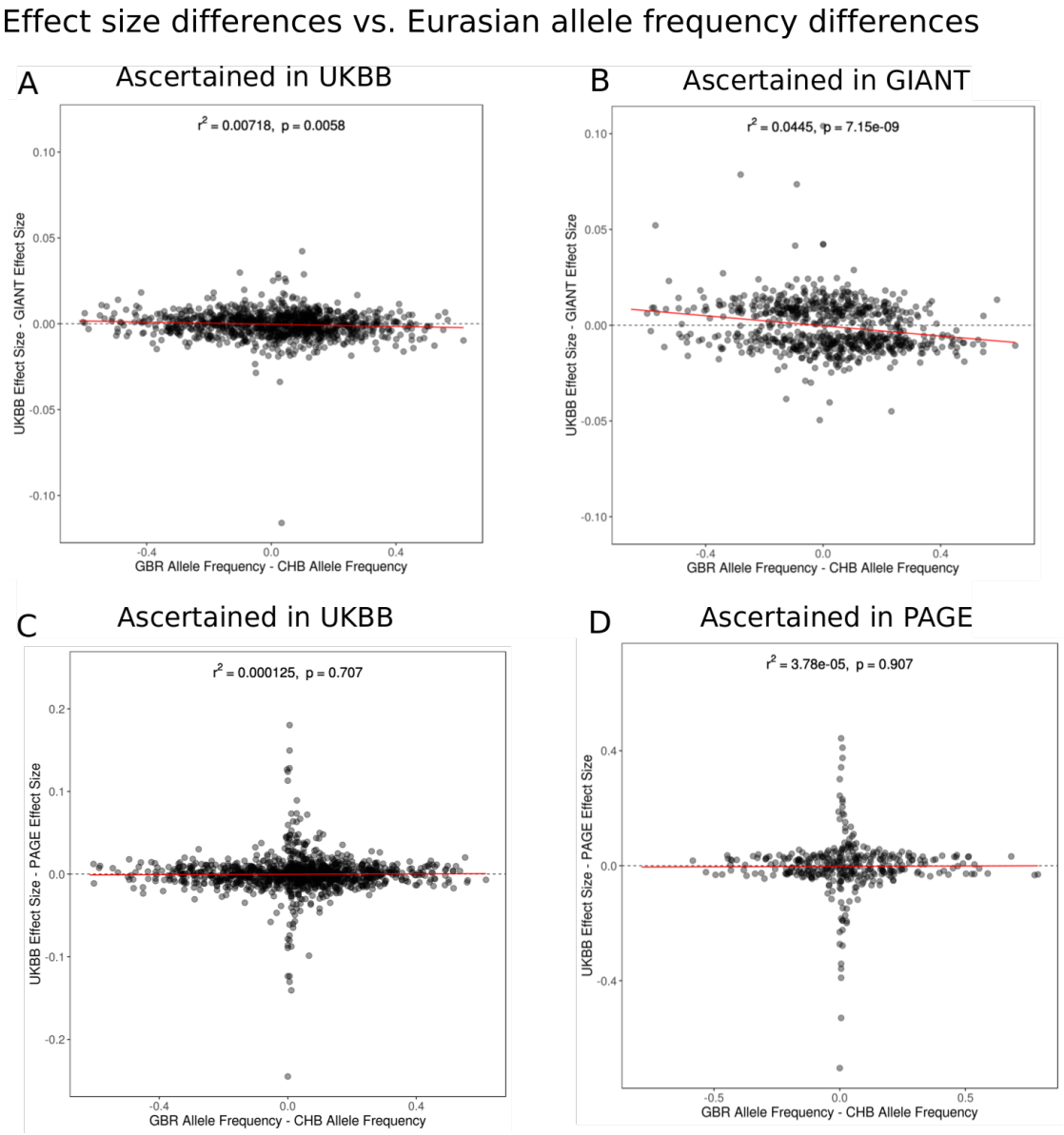

Figure S20: Each panel in this figure shows the same regressions as Figure S19, except that the difference in allele frequency is between a European population panel and an Asian one (GBR and CHB).

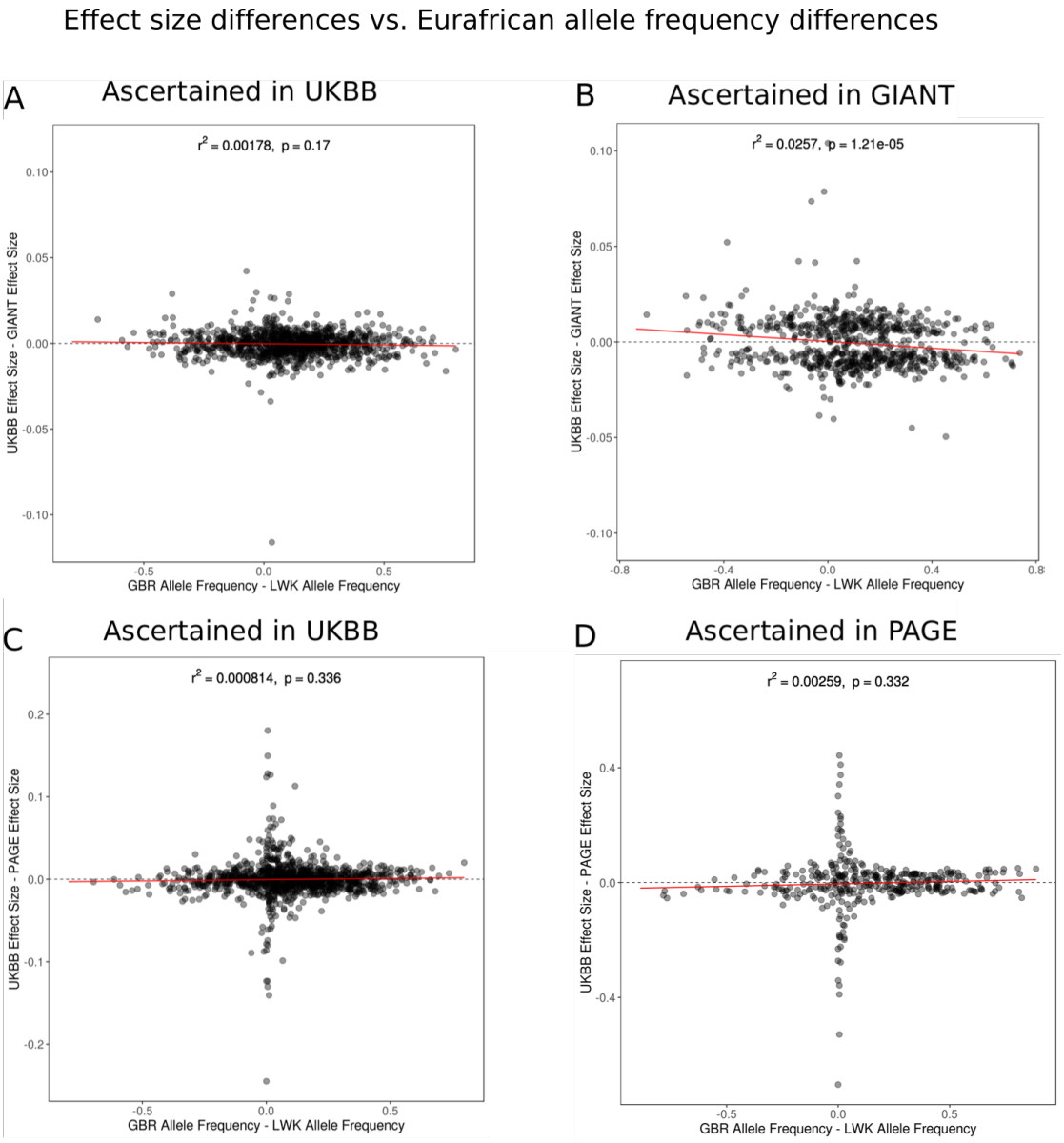

Figure S21: Each panel in this figure shows the same regressions as Figure S19, except that the difference in allele frequency is between the European and African populations panels (GBR and LWK).

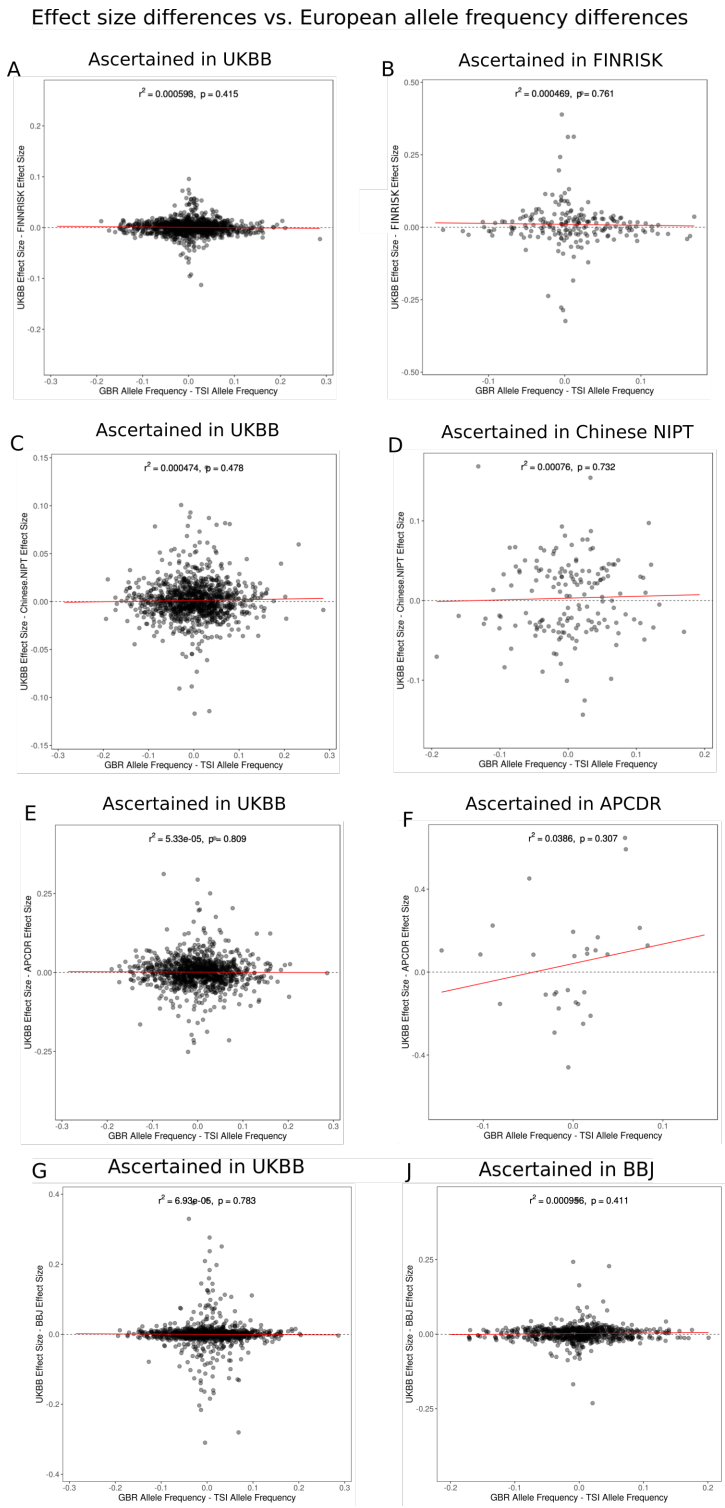

Figure S22: Regression of GWAS effect size differences against allele frequency differences between the northern and southern European populations (GBR and TSI). **A,B**. We selected SNPs for this analysis that had P < 1e−5.

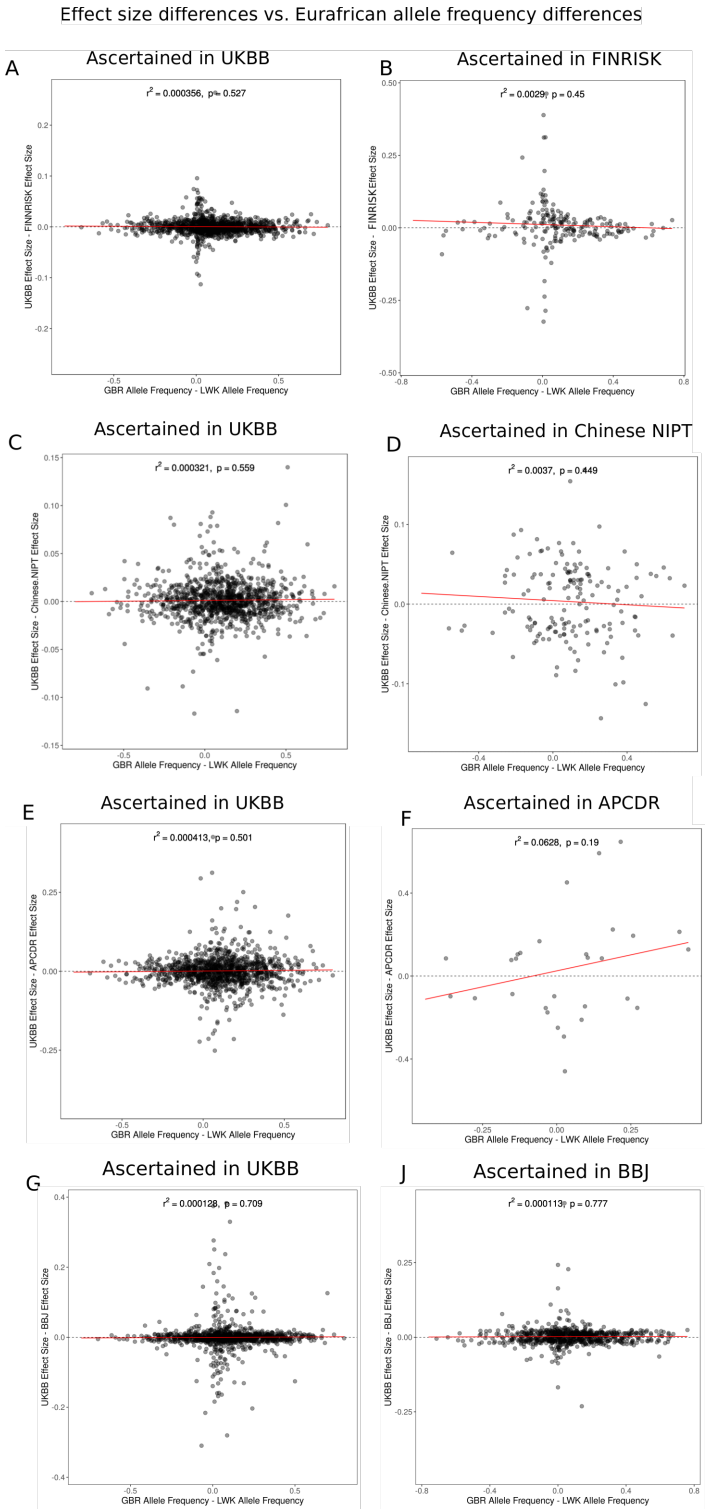

Figure S24: Each panel in this figure shows the same regressions as Figure S22, except that the difference in allele frequency is between the European and African population panels (GBR and LWK).

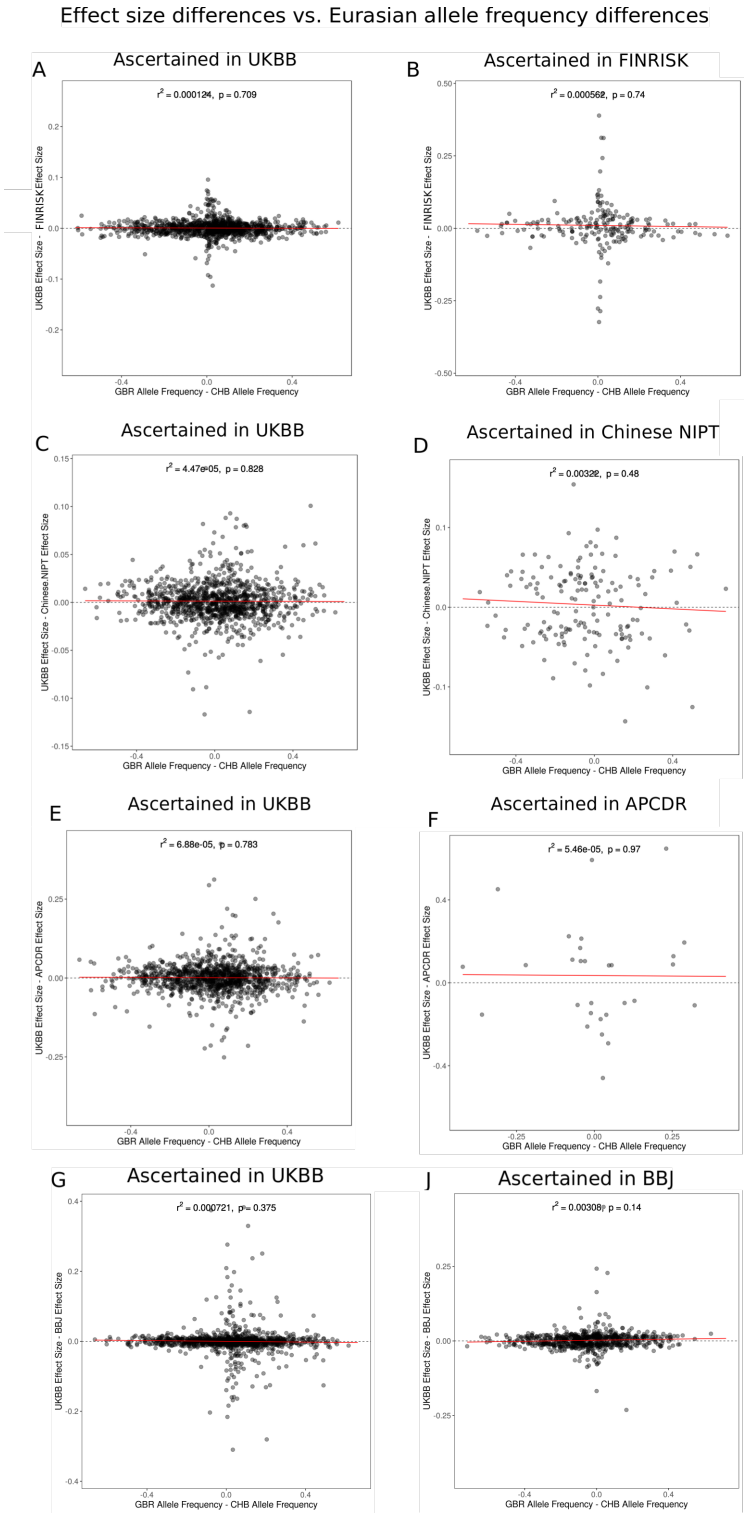

Figure S23: Each panel in this figure shows the same regressions as Figure S22, except that the difference in allele frequency is between the European and Asian population panels (GBR and CHB).

Supplementary Tables

Table S1: Full population descriptions of 1000 Genomes Project panels used in our analysis

Population code	Description	Super population
CHB	Han Chinese in,Beijing, China	EAS
JPT	Japanese in Tokyo,Japan	EAS
CHB	Southern Han,Chinese	EAS
CDX	Chinese Dai in,Xishuangbanna, China	EAS
KHV	Kinh in Ho Chi,Minh City, Vietnam	EAS
CEU	Utah Residents with Northern and Western European Ancestries	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England,and Scotland	EUR
IBS	Iberian Population,in Spain	EUR
YRI	Yoruba in Ibadan,,Nigeria	AFR
LWK	Luhya in Webuye,,Kenya	AFR
GWD	Gambian in Western,Divisions in the Gambia	AFR
MSL	Mende in Sierra,Leone	AFR
ESN	Esan in Nigeria	AFR
ASW	Americans of,African Ancestry in SW USA	AFR
ACB	African Caribbeans,in Barbados	AFR
MXL	Mexican Ancestry,from Los Angeles USA	AMR
PUR	Puerto Ricans from,Puerto Rico	AMR
CLM	Colombians from,Medellin, Colombia	AMR
PEL	Peruvians from,Lima, Peru	AMR
GIB	Gujarati Indian,from Houston, Texas	SAS
PJL	Punjabi from,Lahore, Pakistan	SAS
BEB	Bengali from,Bangladesh	SAS
STU	Sri Lankan Tamil,from the UK	SAS
ITU	Indian Telugu from the UK	SAS

Table S2: Full list of all traits tested and the number of individuals is shown for the GWAS in which this data was available.

		GWAS						
	Traits	UKBB	BBJ	FINRISK	PAGE	APCDR	Chinese NIPT	GIANT
Anthropometric	BMI	359,983	158,284	24,722	49,335	-	60,652	231,410
	Height	360,388	159,095	24,725	49,781	4,778	61,321	253,288
	Waist circumference	360,564	-	-	33,942	-	-	-
	Waist hip ratio (WHR)	-	-	24,671	33,904	-	-	-
Blood pressure	diastolic blood pressure (DBP)	340,162	136,615	21,588	35,433	-	-	-
	Systolic blood pressure (SBP)	340,159	136,597	21,591	35,433	-	-	-
Inflammatory	Platelet	361,141	108,208	6,404	29,328	-	-	-
	Monocyte	349,856	62,076	-	-	-	-	-
	C-reactive protein (CRP)	-	75,391	16,529	28,537	-	-	-
	Hemoglobin (HbA1c)	350,474	42,790	-	11,178	-	-	-
	Mean corpuscular hemoglobin concentration (MCHC)	350,468	108,728	-	19,803	-	-	-
	Mean corpuscular hemoglobin (MCH)	350,472	108,054	-	-	-	-	-
	Mean corpuscular volume (MCV)	350,473	108,256	-	-	-	-	-
	Basophil count	349,856	62,076	-	-	-	-	-
	Hematocrit count	350,475	108,757	-	-	-	-	-
	White blood cell count (WBC)	350,470	107,964	-	28,534	-	-	-
	Red blood cell count (RBC)	350,475	108,794	-	-	-	-	-
	Lymphocyte count	349,856	62,076	-	-	-	-	-
	Monocyte count	349,856	62,076	-	-	-	-	-
	Neutrophil count	349,856	62,076	-	-	-	-	-
Kidney-related	Creatinine	350,812	142,097	6,376	-	-	-	-
Liver-related	Billirubin	-	110,207	-	-	4,778	-	-
Metabolic	Cholesterol	361,141	128,305	-		4,778	-	-
	High density lipoprotein (HDL)	-	70,657	21,620	33,063	4,778	-	-
	Low density lipoprotein (LDL)	-	72,866	21,250	32,221	4,778	-	-
	Triglycerides	-	105,597	21,619	33,096	4,778	-	-
	Glucose	-	93,146	4,418	23,923	-	-	-
Electrolits	Potassium	350,053	132,938	-	-	-	-	-
	Sodium	350,061	127,304	-	-	-	-	-
Lifestyle	Cigarettes per day	25,348	-	-	15,862	-	-	-

Table S3: SNP-based heritability and LD Score regression ratio and intercept estimates (with standard errors in parentheses) for height measured in different cohorts. LD scores were computed using the closest population in the 1000 Genomes Project to each GWAS cohort (meta-analyses of multiple populations were not included here, but see Table 4). The APCDR heritability estimate is not shown because it was estimated to be negative, due to the small sample size of the cohort. For Chinese NIPT GWAS we filtered out all the sites with INFO scores less than 0.4.

GWAS cohort	Genome-wide significant SNPs (P 5e-8)	Observed scale heritability (SE)	LD regression ratio (SE)	LD regression intercept (SE)	Population
UKBB	30891	0.3911 (0.0211)	0.1887 (0.0099)	1.7056 (0.0371)	GBR
BBJ	9976	0.4168 (0.019)	0.1142 (0.0128)	1.1829 (0.0204)	JPT
Chinese NIPT	1573	0.2594 (0.0298)	0.2878 (0.0452)	1.1641 (0.0257)	CHB
FINRISK	415	0.4042 (0.0401)	0.3609 (0.0389)	1.1305 (0.0141)	FIN
APCDR	0	NA	14.8323 (9.7253)	1.0156 (0.0102)	LWK

Table S4: Pairwise Pearson correlation coefficient between effect size estimates from the UK Biobank GWAS and from another GWAS. The SNPs used were determine based on their P-value in the UKBB. n = number of SNPs used to compute the correlation.

	GIANT	FINRISK	PAGE	Chinese NIPT	BBJ	APCDR	SNP filtering scheme
Filtering based on UKBB P-values	0.867 (n = 1700)	0.531 (n = 1700)	0.339 (n = 1700)	0.510 (n = 1700)	0.219 (n = 1700)	0.061 (n = 1700)	SNP with lowest P-value in block
	0.948 (n = 1059)	0.738 (n = 1124)	0.624 (n = 1139)	0.625 (n = 1064)	0.360 (n = 1095)	0.107 (n = 1100)	SNP with lowest P-value in block, if P <1e-5
	0.958 (n = 809)	0.790 (n = 870)	0.615 (n = 881)	0.654 (n = 814)	0.X (n = X)	0.087 (n = 856)	SNP with lowest P-value in block, if P <5e-8

Table S5: Pairwise Pearson correlation coefficient between effect size estimates from the UK Biobank GWAS and from another GWAS. The SNPs used were determine based on their P-value in the non-UKBB study. n = number of SNPs used to compute the correlation.

	GIANT	FINRISK	PAGE	Chinese NIPT	BBJ	APCDR	SNP filtering scheme
Filtering based on non-UKBB P-values	0.698 (n = 1700)	0.167 (n = 1700)	0.128 (n = 1700)	0.335 (n = 1700)	0.018 (n = 1700)	0.121 (n = 1700)	SNP with lowest P-value in block
	0.931 (n = 738)	0.322 (n = 199)	0.258 (n = 366)	0.687 (n = 157)	0.561 (n = 709)	0.138 (n = 29)	SNP with lowest P-value in block if P <1e-5
	0.938 (n = 570)	0.334 (n = 136)	0.260 (n = 257)	0.654 (n = 112)	0.X (n = X)	0.087 (n = 15)	SNP with lowest P-value in block if P <5e-8