Reduced signal for polygenic adaptation of height in UK Biobank

Jeremy J. Berg[*,†,1], Arbel Harpak[*,1,2], Nasa Sinnott-Armstrong[*,3],
Anja Moltke Jørgensen[4], Hakhamanesh Mostafavi[1], Yair Field[3], Evan A Boyle[3], Xinjun Zhang[5],
Fernando Racimo[4], Jonathan K Pritchard[†,2,3,6], Graham Coop[†,7]

[*] Joint first author.

[1] Department of Biological Sciences, Columbia University, New York, NY

[2] Department of Biology, Stanford University, Stanford CA

[3] Department of Genetics, Stanford University, Stanford CA

[4] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark.

[5] Department of Anthropology, University of California, Davis.

[6] Howard Hughes Medical Institute, Stanford University, Stanford CA

[7] Center for Population Biology and Department of Evolution and Ecology, University of California, Davis.

[†] To whom correspondence should be addressed:

`jeremy.jackson.berg@gmail.com`, `pritch@stanford.edu`, and `gmcoop@ucdavis.edu`

Abstract

There is considerable variation in average height across European populations, with individuals in the northwest being taller, on average, than those in the southeast. During the past six years, a series of papers reported that polygenic scores for height also show a north to south gradient, and that this cline results from natural selection. These polygenic analyses relied on external estimates of SNP effects on height, taken from the GIANT consortium and from smaller replication studies. Here, we describe a new analysis based on SNP effect estimates from a large independent data set, the UK Biobank (UKB). We find that the signals of selection using UKB effect-size estimates for height are strongly attenuated, though not entirely absent. Because multiple prior lines of evidence provided independent support for directional selection on height, there is no single simple explanation for all the discrepancies. Nonetheless, our current view is that previous analyses were likely confounded by population stratification and so the conclusion of strong polygenic adaptation in Europe now lacks clear support. Moreover, these discrepancies highlight (1) that current methods for correcting for population structure in GWAS may not always be sufficient for polygenic trait analyses, and (2) that claims of polygenic differences between populations should be treated with caution until these issues are better understood.

Introduction

In recent years, there has been great progress in understanding the polygenic basis of a wide variety of complex traits. One significant development has been advent of "polygenic scores", which aim to predict individual phenotypes using a linear combination of allelic contributions to a given trait across many sites.

One important application of polygenic scores has been the study of polygenic adaptation—the adaptive change of a phenotype through small allele frequency shifts at many sites that affect the phenotype. Thus far, the clearest example of polygenic adaptation has seemed to come from adaptation of height in Europe. However, as we will show here, this signal is strongly attenuated or absent using new data from the UK Biobank [1], now calling this example into question.

While our focus here is on the adaptation of height in Europe, we see this as a case study for understanding the challenges in comparing polygenic scores across populations. Compared to other complex traits, height is particularly well-characterized, and the evidence for adaptation of height in Europeans seemed clear. Thus our work highlights a need for caution in this area until these issues are more fully understood [2].

Starting in 2012, a series of papers identified multiple lines of evidence suggesting that average polygenic scores for height increase from south-to-north across Europe (Table 1). Analyses from multiple groups have concluded that the sharpness of this cline is inconsistent with a neutral model of evolution, suggesting that natural selection drove these differences in allele frequencies and polygenic scores [3, 4, 5, 6, 7, 8, 9]. Significant differences in polygenic scores for height have also been reported among ancient populations, and are also believed to have been driven by selection [10, 11, 7]. In parallel, analysis of recent changes of allele frequencies in a large British sample suggested that natural selection drove a concerted increase in the frequency of height-increasing alleles in the ancestors of modern Britons during the last 2,000 years [12].

All such studies rely on estimates of individual allelic effects on height, as calculated from genome-wide association studies (GWAS). These GWAS estimates are then combined with population-genetic analysis to test for selection. Under a null model of no directional change, we would not expect "tall" alleles to increase (or decrease) in frequency in concert; thus, loosely speaking, a systematic shift in frequency of "tall" alleles in the same direction has been interpreted as evidence for selection.

GWAS data used to study adaptation of height. Until recently, the largest height GWAS data set came from the GIANT consortium (253,288 individuals as of 2014). This is the primary GWAS underlying most studies of adaptation of height [13, 14]. Additionally, several groups have used other, smaller, data sets to replicate signals found using GIANT. In particular, because it is known that population structure may be a confounder in GWAS studies, and that this would affect inferences of polygenic adaptation [3, 4, 5], several of the groups sought to replicate signals using family-based analyses, which protect against structure confounding [15, 16, 17]

The first replication, by Turchin et al. (2012) [3], showed that the effect sizes of the top 1,400 GIANT associations (based on the 2010 GWAS) [13] were statistically consistent with effect sizes re-estimated in a smaller sibling-based regression approach using data from the Framingham Heart Study (4,819 individuals across 1,761 nuclear sibships). Sibling-based regression is presumed to be immune to confounding by population structure, and so the agreement of effect sizes between studies was taken as validation of the north-south gradient observed when using the GIANT effect sizes.

The second, partially independent, replication came from Zoledziewska et al. (2015), who selected 691 height-associated SNPs on the basis of the GIANT association study, and then computed polygenic scores using effect sizes re-estimated in a cohort of 6,307 individuals of Sardinian ancestry. Zoledziewska et al. determined that the average polygenic score of Sardinian individuals was significantly lower than observed for other European populations, consistent with the previously reported north-south gradient of polygenic scores [6].

A third replication was performed by Robinson et al. (2015), who used a different sibling-based GWAS to identify associations [5, 18] (17,500 sibling pairs, we denote this sibling-based dataset by "R15-sibs"). These authors showed that the north-south frequency gradient replicates using SNPs ascertained from the sibling-based GWAS. This replication is stronger than that performed by either Turchin et al., or that by Zoledziewska et al., as the cohort is larger and the SNP ascertainment did not rely on GIANT. As pointed

GWAS	Approach	Signal	Reference
GIANT 2010 validation: Framingham sibs	European frequency cline of top SNPs	strong	Turchin 2012
GIANT 2010	Polygenic measures of pop. frequency differences	strong	Berg 2014
GIANT	Polygenic measures of pop. frequency differences Polygenic diffs between ancient and modern pops	strong strong strong strong	Berg 2017 Racimo 2018 Guo 2018 Mathieson 2015
GIANT validation: R15-sibs	Heterogeneity of polygenic scores among pops	strong	Robinson 2015
Sardinia cohort	Low polygenic height scores in Sardinians. Effect estimates from Sardinian cohort at GIANT hit SNPs	strong	Zoledziewska 2015
GIANT and R15-sibs	Singleton density (SDS) in UK sample vs GWAS Also: LD-score regression (SDS vs GWAS)	strong strong	Field 2016
UK Biobank	Population frequency differences Singleton density (SDS) in UK sample LD score (SDS vs GWAS)	weak or absent weak or absent weak	This paper* This paper* This paper*

Table 1: Studies reporting signals of height adaptation in Europeans. *Prior to the UK Biobank data set, studies consistently found evidence for polygenic adaptation of height. Notes: Most of the papers marked as having "strong" signals report p-values* $< 10^{-5}$*, and sometimes* $<< 10^{-5}$*. In the present paper, the UK Biobank analyses generally yield p-values* $> 10^{-3}$*. *See also results from [21].*

out by Robinson et al., this two-step procedure has the potential to introduce an ascertainment bias, even if the effects are correctly estimated in the replication study (we note that a small fraction of the GIANT samples are contained within the R15-sibs analysis, so the effect sizes are not strictly independent; however, because of the sibling design, any bias due to stratification in GIANT should be absent in R15-sibs). The R15-sibs study was also used by Field et al. to verify a signal of recent selection in ancestors of the current British population. Field et al. found that the signal of selection was even stronger when using R15-sibs than when using GIANT.

Lastly, Field et al. also used LD-score regression to test for height adaptation in the British while controlling for population structure [19, 20]. While LD-score regression is typically used to estimate genetic covariance between two phenotypes, Field et al. used it to test for a relationship between height effects and recent increase in frequency (measured by SDS)—and found a strong covariance of the two consistent with selection driving allele frequency change at height loci.

Weak replication in UK Biobank. In 2017, the UK Biobank (UKB) released genotype and phenotype data for nearly 500,000 residents of the United Kingdom [1]. This dataset has been an important new resource for GWAS of height and many other traits, due to the large sample size, the relatively unstructured population (compared to international studies such as GIANT), and the opportunity for researchers to work directly with the genotype data rather than with processed summary statistics.

Here, we report that previously found signals of selection on height either do not replicate, or are considerably weaker, when using effect-size estimates from the UKB. Similar findings have been obtained independently by other groups [21, 22]. The data do not yet point to a single, simple explanation that covers all of the different observations. As discussed below, our current interpretation is that the result based on UKB is likely correct: evidence for polygenic selection on height is weak or absent; and that the earlier height studies were impacted by confounding due to population structure stratification.

More generally, our work shows that population structure corrections in GWAS may not always work exactly as expected, and that further study is needed to understand the limitations of these methods. We anticipate that current methods are likely sufficient for most applications–in particular for identifying genotype-phenotype associations. However applications such as detecting differences in average polygenic

scores among populations, and detecting polygenic adaptation, are exquisitely sensitive to small amounts of uncorrected structure, and such analyses should be undertaken with great care.

Results

GWAS data sets. We downloaded or generated six different height GWAS data sets, each relying on different subsets of individuals or using different analysis methods as follows. The bold-faced text give the identifiers by which we will refer to each data set throughout this paper. These include two previous studies that show strong evidence for polygenic adaptation:

GIANT: (n=253k) 2014 GIANT consortium meta-analysis of 79 separate GWAS for height in individuals of European ancestry, with each study independently controlling for population structure via the inclusion of principal components as covariates [14].
R15-sibs: (n=35k) Family-based sib-pair analysis of data from European cohorts [18, 5].

We also considered four different versions of analysis of the UK Biobank data, using different subsets of individuals and different processing pipelines:

UKB-GB: (n=337k) Linear regression controlling for 10 principle components of ancestry (unrelated British ancestry individuals only) [23].
UKB-Eur: (n=459k) Structure correction using linear mixed models; all individuals of European ancestry; relatives included [24].
UKB–GB-NoPCs: (n=337k) Linear regression without any structure correction—with only genotype, age, sex and sequencing array as covariates (unrelated British ancestry individuals only) [newly calculated, see Section S1].
UKB-sibs: (n=35k) Family-based sib-pair analysis [newly calculated, see Section S1].

Importantly we find that all of these studies show high pairwise genetic correlations (estimated by LD-score regression), consistent with the view that all of them capture a largely-shared signal of the genetic basis of height (Table 2).

	GIANT	R15-sibs	UKB-Eur	UKB–GB-NoPCs	UKB-GB	UKB-sibs
GIANT		(0.04)	(0.01)	(0.01)	(0.01)	(0.05)
R15-sibs	0.98		(0.04)	(0.04)	(0.05)	(0.08)
UKB-Eur	1.03	0.87		(0.004)	(0.004)	(0.05)
UKB–GB-NoPCs	1.01	0.82	1.00		(0.002)	(0.05)
UKB-GB	1.03	0.89	1.02	1.00		(0.05)
UKB-sibs	1.02	0.93	1.06	1.02	1.06	

Table 2: Pairwise genetic correlations between GWAS data sets. *The table shows genetic correlation estimates (lower triangle) and standard errors (upper triangle) between each of the height data sets, estimated using LD-score regression [20]. All trait pairs show a strong genetic correlation, as expected for different studies of the same trait.*

Signal of selection across Eurasia. One well-studied signal of adaptation of height in Europe has been the observation that, among height-associated SNPs, the "tall" alleles tend to be at higher frequencies in northern populations. Equivalently, the average polygenic scores of individuals in northern populations tend

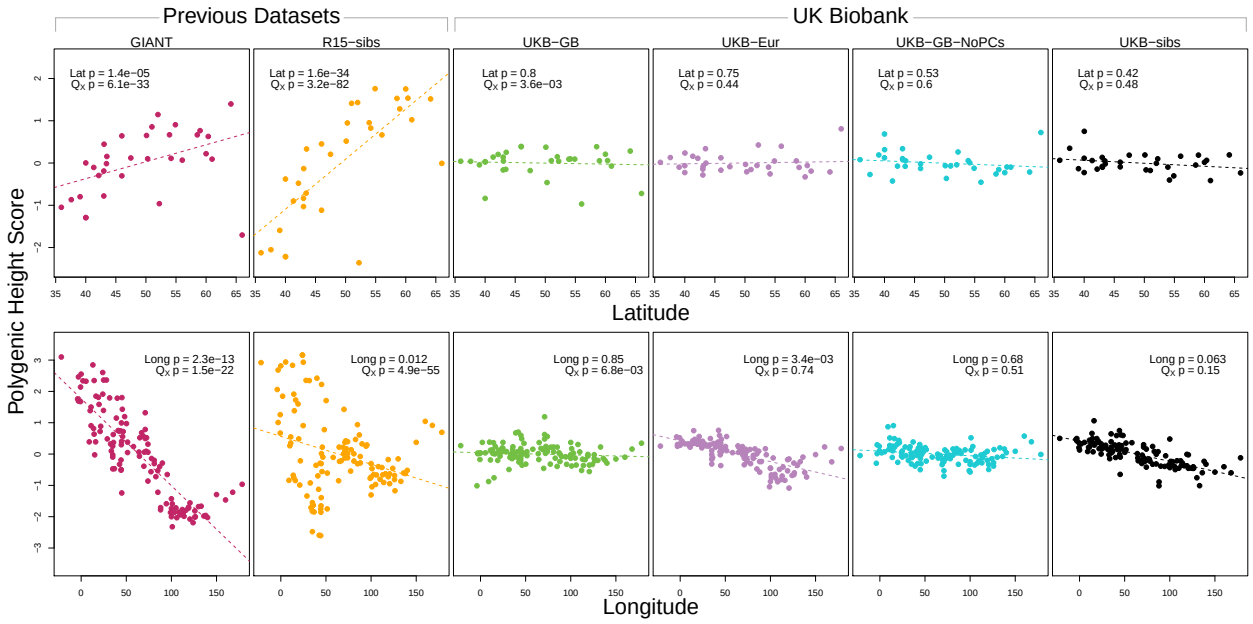

Figure 1: Polygenic scores across Eurasian populations, for different GWAS data sets. *The top row shows European populations from the combined 1000 Genomes plus Human Origins panel, plotted against latitude, while the bottom row shows all Eurasian populations from the same combined dataset, plotted against longitude.*

to be higher than in southern populations.

To evaluate this signal for each dataset, we independently ascertained the SNP with the smallest p-value within each of 1700 approximately independent LD blocks [25, 7] (subject to the constraint that MAF $>$ 0.05 within the GBR 1000 Genomes population). We used these loci to calculate average polygenic scores for each of a set of European population samples taken from the 1000 Genomes and Human Origins panels [26, 27] (see Section S2 for statistical details).

As expected, we find highly significant latitudinal gradients in both the GIANT and R15-sibs data (Figure 1, top row). However, this signal does not replicate in any of the four UK Biobank datasets.

We also tested whether the polygenic scores are over-dispersed compared to a neutral model, without requiring any relationship with latitude (the Q_X test from [4]). Here we find a similar pattern: we strongly reject neutrality using both the GIANT and R15-sibs datasets, but see little evidence against neutrality among the UK Biobank datasets. The sole exception is for the UKB-GB dataset, though the rejection of neutrality in this dataset is marginal compared to that observed with GIANT and R15-sibs, and it does not align with latitude.

While most studies have focused on a latitudinal cline in Europe, a preprint by Berg et al. (2017) [7] also recently reported a cline of polygenic scores decreasing from west to east across all of Eurasia. Extending this analysis across all six datasets, we observe similarly inconsistent signals (Figure 1, bottom row). Only the GIANT dataset shows the clear longitudinal signal reported by Berg et al, though the R15-sibs dataset is again strongly over-dispersed in general, and retains some of the longitudinal signal. Interestingly, we find a weakly significant relationship between longitude and polygenic score in the UKB-Eur dataset (though not in the other UKB data sets), suggesting systematic differences between the results based on British-only and pan-European samples.

We also experimented with using a larger number of significant SNPs. We found that this led to significant values of Q_X, when using UKB-GB effect sizes to ascertain SNPs using a similar procedure to Robinson et al. (2015) [5]. However this signal was sensitive to the particular method of ascertainment, and seems to be diffuse, with part of the signal coming from closely linked SNPs. Thus we conclude that this signal is not robust and may, at least partially, arise from a violation of the assumption of independence among SNPs

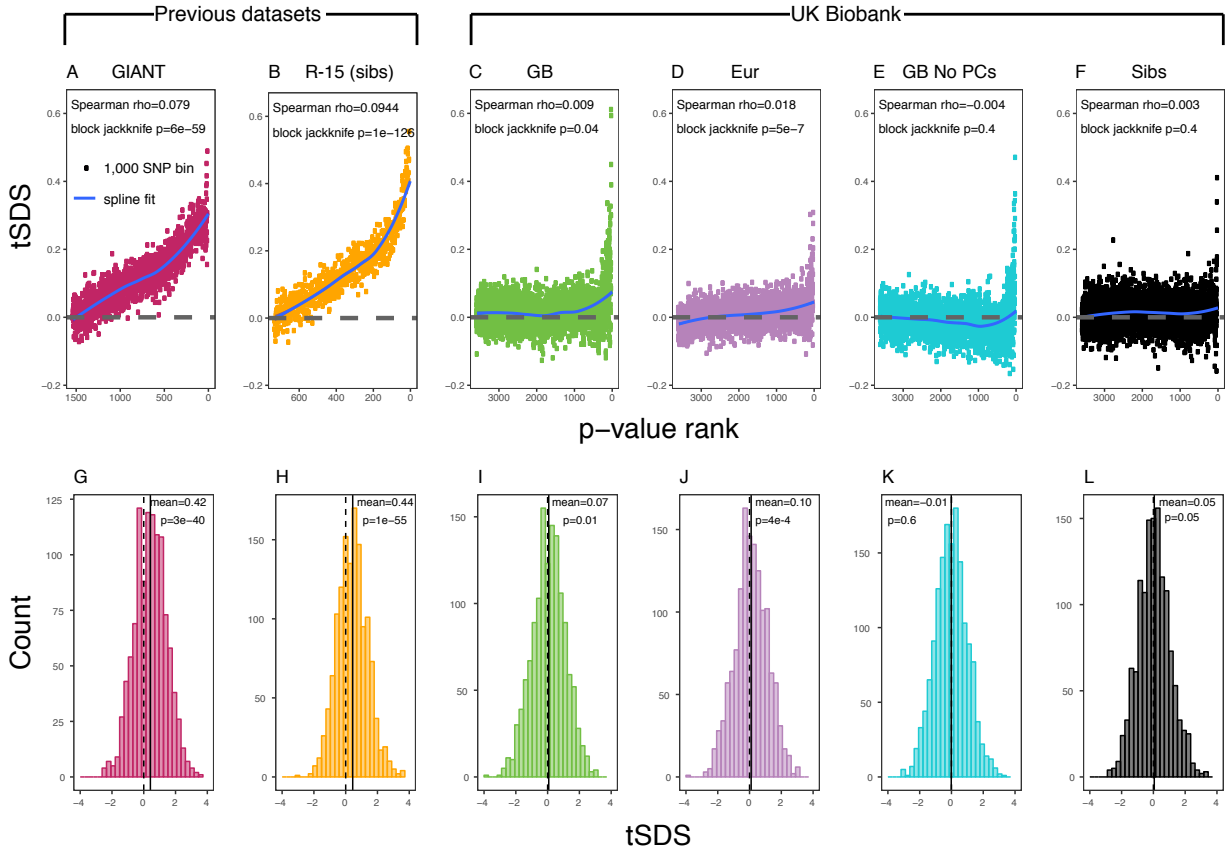

Figure 2: SDS signals for recent selection, assessed using different height GWAS. (A-F) *Each point shows the average tSDS (SDS polarized to height-increasing allele) of 1000 consecutive SNPs in the ordered list of GWAS p-values. Positive values of tSDS are taken as evidence for selection for increased height, and a global monotonic increase—as seen in panels A and B—suggests highly polygenic selection.* **(G-L)** *tSDS distribution for the most significant SNPs in each GWAS, thinned according to LD to represent approximately independent signals. Dashed vertical lines show tSDS=0, as expected under the neutral null; solid vertical lines show mean tSDS. A significantly positive mean value of tSDS indicates selection for increased height.*

that underlies our neutral model (Section S3). Additionally, we tested different frequency, effect size and probability-of-association cutoffs to determine which SNPs we included in the computation of the scores, but found none of these cutoffs affected the discrepancy observed between the GIANT and UKB-GB datasets (Section S4).

SDS signal of selection in Britain. We next evaluated the Singleton Density Score (SDS) signal of selection for increased height in the British population, previously reported by Field et al. [12]. SDS estimates recent changes in allele frequencies at each SNP within a population by comparing the distances to the nearest singletons linked to each of the focal SNP's allele. Field et al. applied SDS using the UK10K sample [28] to investigate allele frequency changes in the ancestors of the British. SDS can be polarized according to the sign of a GWAS effect at each SNP–this is denoted trait-SDS, or tSDS. Here, tSDS>0 indicates that a height-increasing allele has risen in frequency in the recent past; tSDS<0 correspondingly indicates a decrease in frequency of the height-increasing allele. A systematic pattern of tSDS>0 is consistent with directional selection for increased height.

Using both GIANT and R15-sibs, Field et al. found a genome-wide pattern of positive tSDS, indicating

that on average, height-increasing alleles have increased in frequency in the last ∼75 generations. tSDS also showed a steady increase with the significance of a SNP's association with height. We replicate these trends in Figure 2A,B.

This tSDS trend is greatly attenuated in all four GWAS versions performed on the UK Biobank sample (Figures 2C-F). The correlation between UKB-GB GWAS p-value and tSDS is weak (Spearman $\rho = 0.009$, block-jackknife $p = 0.04$). This correlation is stronger for the UKB-Eur GWAS ($\rho = 0.018$, $p = 5e - 07$). Since the UKB-Eur GWAS is not limited to British individuals—but instead includes all European ancestry individuals—this might suggest that residual European population structure continues to confound UKB-Eur effect estimates, despite the use of LMM correction for structure.

We wondered whether the main reason for the weakened trend in UKB-GB is the conservative PC-correction. This could occur if the genetic contribution to height is highly correlated with population structure axes. If this were the case, we would expect the correlation between GWAS p-value and tSDS to still be observed in a UKB GWAS without population structure correction (namely, in UKB–GB-NoPCs). However, we see no evidence for this correlation (block jackknife p=0.6). Taken together with the UKB–GB-NoPCs polygenic score analysis (Figure 1), the lack of signal in UKB–GB-NoPCs suggests that the main reason that UKB is less confounded by population structure than GIANT is the relatively-homogeneous ancestry of the UKB British sample—rather than differences in GWAS correction procedures.

Lastly, we examined tSDS at the most significant height-associated SNPs of each UKB GWAS (as before, ascertained in approximately-independent LD blocks). Significant SNPs show a positive average tSDS (Figure 2I,K,L; t-test p<0.05)—with the exception of the UKB–GB-NoPCs GWAS (Figure 2J) in which the average tSDS is not significantly different from zero (t-test p=0.6).

Relationship between GWAS estimates and European population structure. We have now shown that signals of polygenic adaptation of height are greatly reduced in the UKB data relative to the GIANT and R15-sibs data sets. To better understand the differences among the data sets, we ascertained $1,652$ approximately-independent lead SNPs based on the GIANT p-values to form the basis of comparison between the GIANT and UKB-GB data sets.

Figure 3A shows the effect sizes of ancestral alleles, as estimated using GIANT (x-axis) and UKB-GB (y-axis). The two data sets are highly correlated ($r^2 = 0.78$, $p < 2.2 \times 10^{-16}$), consistent with the strong genetic correlation estimated in Table 2. The fact that the slope is <1 probably reflects, at least in part, the standard winner's curse effect for SNPs ascertained in one study and replicated in another.

Importantly however, we also see clear evidence that the *differences* between the GIANT and UKB-GB effect sizes are partly related to European population structure. Specifically, for each SNP we plotted the difference in allele frequency between northern and southern European samples (specifically, between the British (GBR) and Tuscan (TSI) subsets of 1000 Genomes) versus the difference in effect size between GIANT and UKB-GB. These differences have a significant correlation ($r^2 = 0.06$, $p < 2.2 \times 10^{-16}$), indicating that alleles that are more frequent in GBR, compared to TSI, tend to have more positive effect sizes in GIANT than in UKB-GB, and vice versa. We also observed a similar signal for frequency differences between TSI and the Han Chinese in Beijing (Figure S12).

Similar patterns are present in a comparison of the R15-sibs and UKB-GB datasets when ascertaining from R15-sibs p-values (Figure 3, panels C and D; 1,642 SNPs). Here, the correlation between effect size estimates is much lower ($r^2 = 0.14$, $p < 2.2 \times 10^{-16}$), likely due to the much smaller sample size of R15-sibs. However, the correlation between the effect-size difference and the GBR-TSI allele frequency difference remains ($r^2 = 0.07$, $p < 2.2 \times 10^{-16}$). In contrast, when SNPs are ascertained on the basis of their UKB-GB p-value, these patterns are considerably weaker in both the GIANT and R15-sibs datasets (Figure S13).

Finally, an unexpected feature of the R15-sibs dataset can be seen in Figure 3D: there is a strong skew for the ancestral allele to be associated with increased height ($1,201$ out of the $1,642$ SNPs ascertained with R15-sibs p-values have positive effect sizes in R15-sibs). This pattern is not present in any of the other datasets. We do not currently have any explanation for this observation.

Together, these observations suggest the hypothesis that while all of the datasets primarily capture real signals of association with height, both the GIANT and R15-sibs effect sizes are also partly confounded by

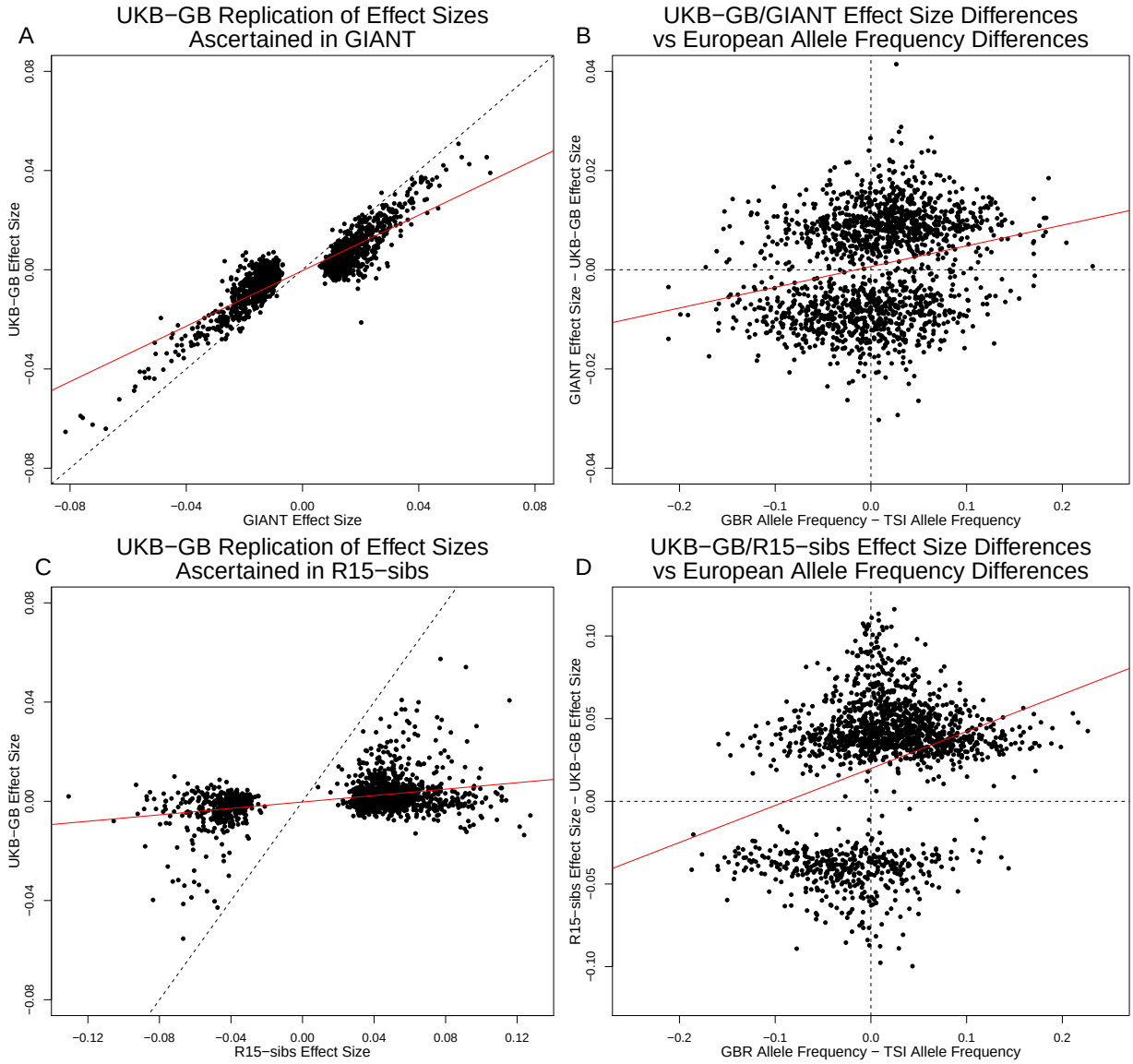

Figure 3: Effect size estimates and population structure. Top Row: *SNPs ascertained using GIANT compared with UKB-GB.* **(A)** *The x- and y-axes show the estimated effect sizes of SNPs in GIANT and in UKB-GB. Note that the signals are highly correlated overall, indicating that these partially capture a shared signal (presumably true effects of these SNPs on height).* **(B)** *The x-axis shows the difference in ancestral allele frequency for each SNP between 1000 Genomes GBR and TSI; the y-axis shows the difference in effect size as estimated by GIANT and UKB-GB. These two variables are significantly correlated, indicating that a component of the difference between GIANT and UKB-GB is related to the major axis of population structure across Europe.* **Bottom Row:** *SNPs ascertained using R15-sibs compared with UKB-GB.* **(C)** *The same plot as panel (A), but ascertaining with and plotting R15-sibs effect sizes rather than GIANT* **(D).** *Similarly, the same as (B), but with the R15-sibs ascertainment and effect sizes.*

population structure corresponding to major axes of variation in Europe and Eurasia. This could drive false positive signals in geographic-based analyses of polygenic adaptation. Furthermore, since SDS measured in Britain correlates with north-south frequency differences [12], this could also drive false positives for SDS.

To explore this further, we next turn to an analysis of the data sets based on LD-score regression.

LD score regression signal. Another line of evidence in Field et al. [12] came from LD-score regression (LDSR, [19, 20]). LD-score regression applies the principle that, under a polygenic model, SNPs in regions of stronger LD (quantified by "LD score") should tag more causal variants and therefore have larger squared estimated effect sizes. Similarly, if two traits share a genetic basis, then the correlation between estimated effect sizes of these traits should increase with LD score. Meanwhile, confounders such as population structure are expected to affect SNPs of different LD score equally, and therefore affect the intercept but not the slope of a linear regression to LD score [19] (we return to this point below and in Supplementary Note S6).

While LD-score regression is commonly used to estimate the genetic covariance between pairs of phenotypes [20], Field et al. used it to test for a relationship between height and SDS. SDS is similar to GWAS effect estimates in that the expected change in frequency of an allele depends on both its own fitness effect and on the effects of those in linkage disequilibrium with it. Field et al. predicted that the covariance between estimated marginal height effect and SDS should increase with LD score—and found this to be the case using both GIANT and R15-sibs. This provided further evidence for polygenic adaptation for increased stature in Britain.

Here, we revisit Field et al.'s observations (Figure 4A,B). Both GIANT and R15-sibs exhibit a highly significant LD-score regression slope (scaled GIANT slope= 0.17, $p = 5 \times 10^{-9}$; scaled R15-sibs slope= 0.46, $p = 7 \times 10^{-17}$), as well as a highly significant intercept (GIANT intercept= 0.093, $p = 4 \times 10^{-71}$; R15-sibs intercept= 0.119, $p = 2 \times 10^{-87}$). These large intercepts suggest that both GWAS suffer from stratification along an axis of population structure that is correlated with SDS in the British population. In contrast, in LD-score regression with the UKB-GB GWAS, the intercept is not significant ($p = 0.10$), suggesting that UKB-GB is not strongly stratified (or at least, not along an axis that correlates with SDS). The slope is $\sim$1/3 as large as in GIANT, though still modestly significant ($p = 1.2 \times 10^{-3}$, Figure 4C).

LDSR of population frequency differences. We next extended Field et al.'s LD score rationale from SDS to test whether SNP effects on height affected allele frequency differentiation between northern and southern Europe. We used 1000 Genomes British (GBR) and Tuscan (TSI) samples as proxies for northern and southern ancestry respectively. To control for the correlation between allele frequencies and LD score, we normalized the frequency differences to have variance 1 within 1% average minor allele frequency bins. For shorthand, we refer to this measure as [GBR-TSI]. Under a model of selection driving allele frequency differences, we would expect the covariance of [GBR-TSI] and effect sizes to increase with LD score. We regressed the product [GBR-TSI] $\times$ effect size (estimated in previous and UKB GWAS) against LD score.

In contrast to SDS, we find that none of the GWAS data sets show a strongly positive slope (Figure 4D-F): the slope is approximately zero in GIANT, weakly positive in R15-sibs (p=0.002), and weakly negative in UKB-GB (p=0.09).

We see extremely strong evidence for positive intercepts in GIANT (p=4×10^{-80}) and R15-sibs (p=9×10^{-161}), but not in UKB-GB (p=0.05). The large intercepts in GIANT and R15-sibs are consistent with population confounding affecting both of these GWAS, as the North-South allele frequency difference is systematically correlated with the effect sizes in these GWAS independently of LD score (see Section S6.4 for a more technical discussion). However, the relative lack of slope in these analyses suggests that the LD score signal for SDS must be driven by a component of frequency change that is largely uncorrelated with the [GBR-TSI] axis of variation.

Population structure confounds LD-score regression slope. The original LD-score regression paper noted that in the presence of linked selection, allele frequency differentiation could plausibly increase with LD score. However, they concluded that this effect was negligible in the examples they considered [19].

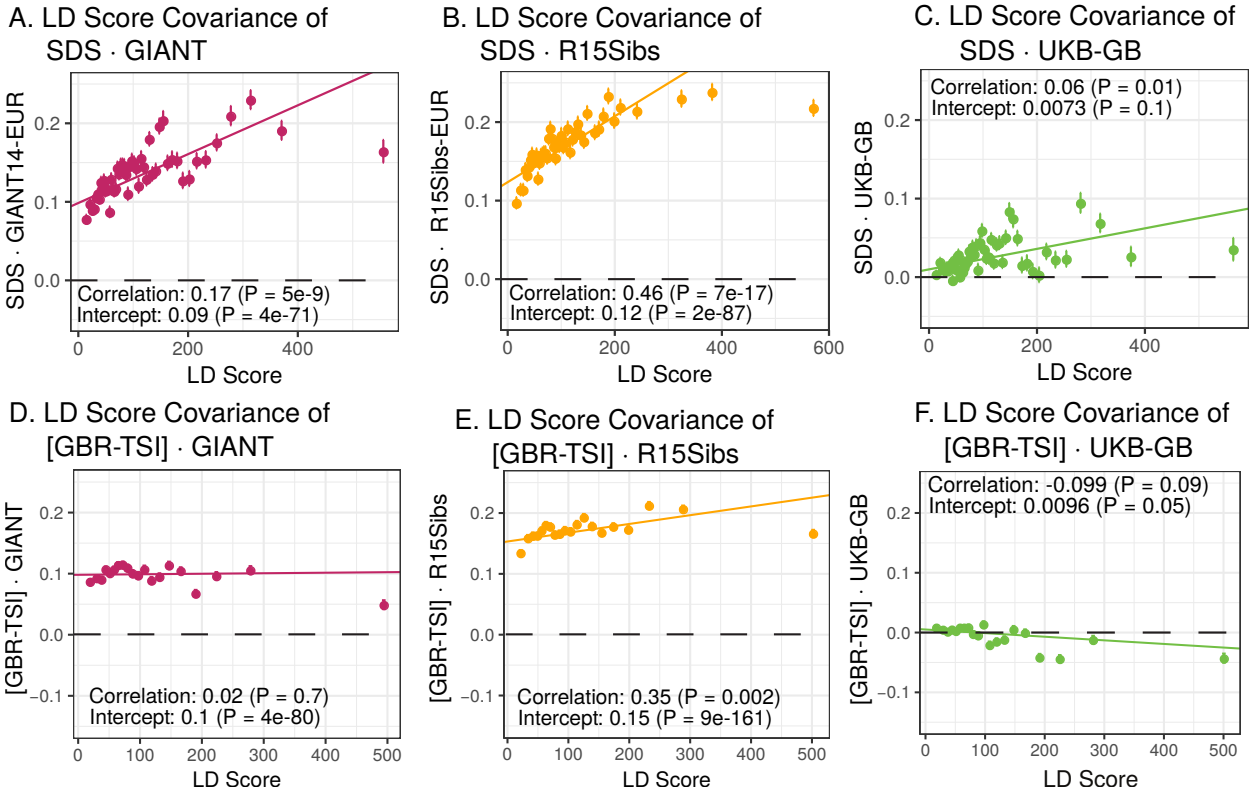

Figure 4: SDS LD-score regression analyses. **(A)**, **(B)**, *and* **(C)** *LD-score covariance analysis of SDS with GIANT, R15-sibs, and UKB-GB, respectively. The x-axis of each plot shows LD score, and the y-axis shows the average value of the product of effect size on height and SDS, for all SNPs in a bin. Genetic correlation estimates are a function of slope, reference LD scores, and the sample size [20]. Both the slope and intercept are substantially attenuated in UKB-GB.* **(D)**, **(E)** *and* **(F)** *Genetic covariance between GBR-TSI frequency differences vs. GIANT, R15-sibs, and UKB-GB. GIANT and R15-sibs show highly significant nonzero intercepts, consistent with a signal of population structure in both data sets, while UKB-GB does not. In addition, R15-sibs shows a significant slope with LD score.*

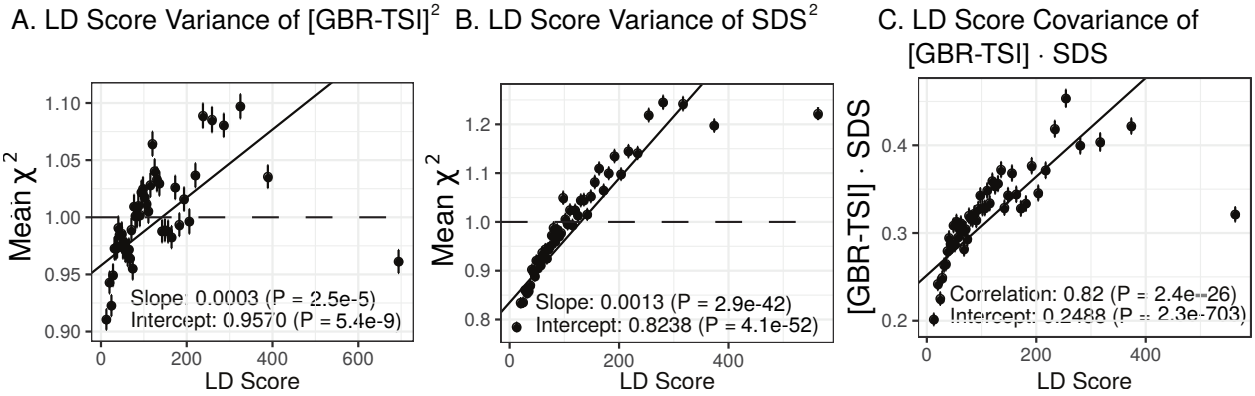

Figure 5: Population allele frequencies show genetic correlation with European height GWAS. (A), **(B** *and* **(C)** *Magnitude of squared GBR-TSI allele frequency differences, squared SDS effect sizes, and the product of allele frequency and SDS increase with LD score. Both SDS and GBR-TSI frequency difference are standardized and normalized within 1% minor allele frequency bins.*

Given our results in Figure 4, where we see strong signals for SDS and height in GIANT and R15-sibs but not in UKB-GB, we decided to revisit this assumption. We find a strong correlation between [GBR-TSI]^2 and LD scores ($p = 2.5 \times 10^{-5}$, Figure 5A). We find a similarly strong correlation of allele frequency differentiation and LD score for lower levels of population differentiation—between self-identified "Irish" and "White British" individuals in the UK Biobank ($p = 2.5 \times 10^{-7}$, Figure S14). These correlations indicate that allele frequency differentiation increase on average with local LD levels. Finally, we find a similar pattern regressing SDS^2 and [GBR-TSI]×SDS against LD scores (Figure 5B,C). These results show that that the signals of greater allele frequency change in regions of stronger LD are shared between [GBR-TSI] and SDS.

Background selection and LD score. The correlation we observe between LD scores and allele frequency differentiation could be generated by the genome-wide effects of linked selection. While a range of different modes of linked selection likely act in humans, one of the simplest is background selection [29, 30, 31]. Background selection (BGS) on neutral polymorphisms results from the purging of linked, strongly deleterious alleles. Because any neutral allele that is in strong LD with a deleterious mutation will also be purged from the population, the primary effect of BGS is a reduction in the number of chromosomes that contribute descendants in the next generation. The impact of BGS can therefore be thought of approximately as increasing the rate of genetic drift in genomic regions of strong LD relative to regions of weak LD. Therefore, SNPs with larger LD scores will experience stronger BGS and a higher rate of genetic drift, and this effect could generate a positive relationship between LD scores and allele frequency differentiation.

In Supplementary Section S6, we derive a simple model for the effect of BGS on the relationship between allele frequency divergence and LD scores. Empirically, we find that LD scores are positively correlated with the strength of background selection [31], and that our simple model of background selection is capable of explaining much of the relationship between LD scores and allele frequency divergence that we observe in Figure 5 (see Sections S6.2 and S6.3, Figures S17-S19). Further, in the presence of BGS, bivariate regression of a stratified GWAS together with a measure of allele frequency differentiation can result in a positive slope, provided that the axis of stratification is correlated with the chosen measure of allele frequency divergence.

Summary of LD Score Regression Results. What conclusions should we draw from our LD score regression analyses? The significant positive intercepts observed for LDSR of both GIANT and R15-sibs with [GBR-TSI] suggest that both data sets suffer from population confounding along a north-to-south axis within Europe. These observations are consistent with the evidence presented in Figure 3. A positive slope in such analyses was previously interpreted as evidence of positive selection either on height or a close genetic correlate (and presumed to be robust to stratification). However, BGS can, and empirically does, violate the assumptions of LD score regression in a way that may generate a positive slope. We therefore interpret the positive slopes observed for the LDSR signals for GIANT, R15-sibs with SDS as likely resulting from a combination of stratification and BGS. A similar conclusion applies to the positive slope observed for R15-sibs × [GBR-TSI]. It is unclear why stratification plus BGS should have elevated the slope for GIANT ×SDS, but not for GIANT ×[GBR-TSI]. This may suggest that the apparent SDS selection signal found in

GIANT may be driven by an axis of variation that is not strongly correlated with [GBR-TSI]. We view this as an area worthy of further theoretical exploration going forward.

Discussion

To summarize the key observations, we have reported the following:

- Multiple analyses based on GIANT and R15-sibs indicate strong signals of selection on height.
- However, the same signals of selection are absent or greatly attenuated in UK Biobank data. In some, but not all, analyses of frequency differentiation and SDS we still detect weakly significant signals of polygenic adaptation (Figures 1 and 2).
- The GIANT height GWAS is overall highly correlated with UKB-GB, but differs specifically by having an additional correlation with the main gradient of allele frequency variation across Europe, as modeled by frequency differences between GBR and TSI (Figures 3). LD score analysis of [GBR-TSI]$\times$ GIANT effect-size also suggests that GIANT is stratified along this axis (positive intercept in Figure 4D).
- Selection signals in the R15-sibs data are consistent with, and in some cases even stronger, than the corresponding signals in GIANT. They are inconsistent with analyses using UK Biobank data. While sib-based studies are designed to be impervious to population structure [GBR-TSI]$\times$ R15-sibs effect-size also shows a highly positive intercept in the LD-score analysis presented in Figure 4E.
- LD Score analyses show a much stronger relationship between SDS and GIANT or R15-sibs than between SDS and UKB-GB. LD score regression is generally considered to be robust to population structure (but see the discussion in [19]). However, the intensity of background selection increases with LD score (Figure 5, Section S6), and this has likely inflated the LD score-based signal of selection in GIANT and R15-sibs.

In principle, it is possible that height in the UKB is confounded in a way that suppresses the signal of height adaptation; it is also possible that current methods for detection of polygenic adaptation are underpowered or flawed. However, we found no evidence for either of these hypotheses. Instead, multiple lines of evidence suggest that population-structure confounding in GIANT and R15-sibs may be the driver of the discrepancy with UKB-based analyses.

GIANT was conducted across a large number of study sites that provided summary statistics to the overall consortium. While the overall value of this pioneering data set is not in question, it would not be surprising in retrospect if residual population structure leaked into this GWAS. The sib design used by Robinson et al offered a strong independent replication of the polygenic adaptation signal, that should have been impervious to population structure concerns. However, the study does show potential signs of stratification; we currently have no insight into what could have driven this issue.

Lastly, we must conclude that the strong signal of LD-score genetic covariance between SDS and both GIANT and R15-sibs is largely spurious. This would imply that the LDSR slope is not robust to population structure confounding. Specifically, We demonstrated that background selection—through its correlation with LD score—can potentially generate a spurious LDSR slope.

Together, these discrepancies highlight the extreme sensitivity of analyses of polygenic adaptation–and reports of polygenic score differences between populations–to the input GWAS [2]. On the other hand, there is still strong evidence that typical GWAS studies do indeed capture genuine signals of genotype-phenotype associations. This is exemplified by the strong correlations in effect sizes between top hits from GIANT with those from UK Biobank (Figure 3A), and by the high genetic correlations among the different height studies as measured from all SNPs (Table 2). More generally, GWAS studies show other features indicating that much of the signal is genuine: for example, strong functional enrichments within active chromatin from trait-relevant tissues [32].

Nonetheless, we have shown that GIANT effect-size estimates also contain a component related to European population structure (Figure 3B). While this is a small effect at individual SNPs, it is strong enough

to overwhelm polygenic analyses of population differences. We suspect that meta-analyses such as GIANT, that collate summary statistics from many sources, may be particularly sensitive to structure confounding. This is likely to be a problem for other GWAS meta-analyses that have used this design, especially for studies of traits with phenotypic variation across populations.

Furthermore, it is striking that even within the UK Biobank data, we see marked differences between results based on UKB-GB and UKB-Eur (Figure 2C, I vs. D, J and S15). The study subjects in the two data sets were largely overlapping, and both were computed using widely-accepted structure-correction methods. These discrepancies hint that, in the more-demanding setting of European ancestry variation and related individuals, the linear mixed model approach did not provide full protection against structure confounding. This highlights a need for renewed exploration of the robustness of these methods, especially in the context of polygenic prediction.

Looking forward, we recommend that studies of polygenic adaptation should focus on data sets that minimize population structure (such as subsets of UKB), and where the investigators have access to full genotype data, including family data, so that they can explore sensitivity to different datasets and analysis pipelines.

Acknowledgements

We thank Manuel Rivas for assisting with the UK Biobank (Application #24983). We thank the Coop lab, Pritchard lab, Przeworski lab, Sella lab, Doc Edge, John Novembre, Guy Sella, Molly Przeworski, Joshua Schraiber, Loic Yengo and the authors of Robinson et al. (2015) and Sohail, Maier et al. (2018) for helpful conversations and feedback on earlier drafts. J.J.B. was supported in part by NIH R01 GM115889 to Guy Sella and in part by NIH Grant F32 GM126787 to J.J.B. A.H. was supported, in part, by a fellowship from the Stanford Center for Computational, Evolutionary and Human Genomics. N.S.-A. was supported by a Stanford Graduate Fellowship and by the Department of Defense through a National Defense Science and Engineering Grant. H.M was supported by NIH R01 GM121372 to Molly Przeworski. G.C. was partially supported by NIH R01 GM108779. Work in the Pritchard lab was supported in part by NIH R01 HG008140.

References

[1] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

[2] John Novembre and Nicholas H Barton. Tread lightly interpreting polygenic tests of selection. *Genetics*, 208(4):1351–1355, 2018.

[3] Michael C Turchin, Charleston WK Chiang, Cameron D Palmer, Sriram Sankararaman, David Reich, Joel N Hirschhorn, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, 44(9):1015–1019, 2012.

[4] Jeremy J Berg and Graham Coop. A population genetic signal of polygenic adaptation. *PLoS Genetics*, 10(8):e1004412, 2014.

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

[6] Magdalena Zoledziewska, Carlo Sidore, Charleston WK Chiang, Serena Sanna, Antonella Mulas, Maristella Steri, Fabio Busonero, Joseph H Marcus, Michele Marongiu, Andrea Maschio, et al. Height-reducing variants and selection for short stature in Sardinia. *Nature Genetics*, 47(11):1352, 2015.

[7] Jeremy J Berg, Xinjun Zhang, and Graham Coop. Polygenic adaptation has impacted multiple anthropometric traits. *bioRxiv*, 2017.

[8] Fernando Racimo, Jeremy J Berg, and Joseph K Pickrell. Detecting polygenic adaptation in admixture graphs. *Genetics*, 208(4):1565–1584, 2018.

[9] Jing Guo, Yang Wu, Zhihong Zhu, Zhili Zheng, Maciej Trzaskowski, Jian Zeng, Matthew R Robinson, Peter M Visscher, and Jian Yang. Global genetic differentiation of complex traits shaped by natural selection in humans. *Nature communications*, 9(1):1865, 2018.

[10] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

[11] Rui Martiniano, Lara M Cassidy, Ros Ó'Maoldúin, Russell McLaughlin, Nuno M Silva, Licinio Manco, Daniel Fidalgo, Tania Pereira, Maria J Coelho, Miguel Serra, et al. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genetics*, 13(7):e1006852, 2017.

[12] Yair Field, Evan A Boyle, Natalie Telis, Ziyue Gao, Kyle J Gaulton, David Golan, Loic Yengo, Ghislain Rocheleau, Philippe Froguel, Mark I McCarthy, et al. Detection of human adaptation during the past 2000 years. *Science*, 354(6313):760–764, 2016.

[13] Hana Lango Allen, Karol Estrada, Guillame Lettre, Sonja I. Berndt, Michael N. Weedon, Fernando Rivadeneira, Cristen J. Willer, Anne U. Jackson, Sailaja Vedantam, Soumya Raychaudhuri, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[14] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186, 2014.

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

[16] Richard S Spielman and Warren J Ewens. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *The American Journal of Human Genetics*, 62(2):450–458, 1998.

[17] Gonçalo R Abecasis, Lon R Cardon, and WOC Cookson. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics*, 66(1):279–292, 2000.

[18] Gibran Hemani, Jian Yang, Anna Vinkhuyzen, Joseph E Powell, Gonneke Willemsen, Jouke Jan Hottenga, Abdel Abdellaoui, Massimo Mangino, Ana M Valdes, Sarah E Medland, Pamela A Madden, Andrew C Heath, Anjali K Henders, Dale R Nyholt, Eco J C de Geus, Patrik K E Magnusson, Erik Ingelsson, Grant W Montgomery, Timothy D Spector, Dorret I Boomsma, Nancy L Pedersen, Nicholas G Martin, and Peter M Visscher. Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs. *American Journal of Human Genetics*, 93(5):865–875, 2013.

[19] Brendan Bulik-Sullivan, Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

[20] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11):1236–1241, 2015.

[21] Mashaal Sohail, Robert Maier, Andrea Ganna, Alexander Bloemendal, Alicia Martin, Mark Daly, Nicholas Patterson, Benjamin M Neale, Iain Mathieson, David Reich, and Shamil Sunyaev. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*, 2018.

[22] Lawrence H Uricchio, Hugo C Kitano, Alexander Gusev, and Noah Zaitlen. An evolutionary compass for elucidating selection mechanisms shaping complex traits. *bioRxiv*, 2018.

[23] Claire Churchhouse, Benjamin M Neale, et al. Rapid gwas of thousands of phenotypes for 337,000 samples in the uk biobank. https://sites.google.com/broadinstitute.org/ukbbgwasresults/home?authuser=0 Accessed: 2018-04-27.

[24] Po-Ru Loh, Gleb Kichaev, Steven Gazal, Armin P Schoech, and Alkes L Price. Mixed model association for biobank-scale data sets. *bioRxiv*, 2017.

[25] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283–285, 2016.

[26] The 1000 Genomes Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[27] Lazaridis, Iosif and Patterson, Nick and Mittnick, Alissa and Renaud, Gabriel and Mallick, Swapan and Kirsanow, Karola and Sudmant, Peter H. and Schraiber, Joshua G. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513:409–413, 2014.

[28] UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, 2015.

[29] Brian Charlesworth, M T Morgan, and Deborah Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

[30] Brian Charlesworth. Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15(5):538–543, 1998.

[31] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, 5(5):e1000471, 2009.

[32] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, 2015.

[33] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[34] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[35] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[36] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014.

[37] Gabriel Renaud. glactools: a command-line toolset for the management of genotype likelihoods and allele counts. *Bioinformatics*, 1:3, 2017.

[38] Hans R Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

[39] Frank MTA Busing, Erik Meijer, and Rien Van Der Leeden. Delete-m jackknife for unequal m. *Statistics and Computing*, 9(1):3–8, 1999.

[40] Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099, 2010.

[41] Loic Yengo, Jian Yang, and Peter M Visscher. Expectation of the intercept from bivariate ld score regression in the presence of population stratification. *bioRxiv*, 2018.

S1 Newly calculated GWAS

Figure 1 and 2 display analyses based on six different GWAS. Two of these GWAS were newly calculated by us using UK Biobank data. Below, we describe the specifics of these two GWAS.

UKB-GB-NoPCs. To preform this GWAS, we used the following *plink v. 2.0* [33] with command line as follows:

plink2 --memory 64G --threads 16 --linear
--bpfile ⟨UKB imputed SNPs bp file⟩
--keep ⟨id list of individuals self-identified as "White British"⟩
--out ⟨output file⟩
--pheno ⟨standing height phenotype file (UKB phenotype 50.0.0)⟩
--covar ⟨covariates file⟩

The covariates file included only the sex, age and sequencing array for each individual id. We filtered all A↔G or C↔T SNPs–to prevent the possibility of strand errors. Finally, we excluded SNPs for which SDS was not calculated in Field et al. 2016 [12].

UKB-sibs. We used the estimated kinship coefficient (ϕ) and the proportion of SNPs for which the individuals share no allele (IBS0) provided by the UK Biobank, to call siblings as pairs with

$$\frac{1}{2^{5/2}} < \phi < \frac{1}{2^{3/2}}$$

and IBS0 > 0.0012—following the conditions used by [1]. We further filtered sibling pairs such that both individuals were White British, their reported sex matched their inferred sex, were not identified by the UK Biobank as "outliers" based on heterozygosity and missing rate nor had an excessive number relatives in the data, and had height measurements. We standardized height values for each sex based on its mean and standard deviation (SD) values in the 336,810 unrelated White British samples: mean 175.9cm and SD 6.7cm for males, and mean 162.7cm and SD 6.2cm for females. We also removed pairs if one of the siblings was more than 5 SD away from the mean. After applying all filters, 19,268 sibling pairs remained, equaling 35,524 individuals in 17,275 families. We performed an association analysis on 10,879,183 biallelic SNPs included in UKB-GB (converting dosages from imputation to genotype calls using no hard calling threshold), using *plink* v. 1.9 [33] QFAM procedure with the following command:

plink2 --bfile ⟨UKB hard-called SNPs file⟩
--out ⟨output file⟩
--qfam mperm=100000

The family relationships, as well as the phenotypic values, were encoded in *plink* FAM files.

S2 Neutrality tests for polygenic scores

In Figure 1, we employ two separate tests to assess the evidence that the distribution of polygenic scores among populations is driven in part by adaptive divergence. Both are based on a simple null model introduced by Berg and Coop (2014)[4], which states that the distribution of polygenic scores under neutrality should be approximately multivariate normal. Here, we give a brief overview of the assumptions and calculations underlying the null model, before describing the two tests used in Figure 1. For a more complete treatment, see Berg and Coop (2014).

Let $\vec{p}_\ell$ be the vector of population allele frequencies at SNP ℓ, while α_ℓ is the effect size for SNP $\ell \in \{1, ..., L\}$. Then, population level polygenic scores are given by

$$\vec{Z} = 2\sum_\ell \alpha_\ell \vec{p}_\ell. \tag{1}$$

Under neutrality, the distribution of polygenic scores among populations should be approximately

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

where

$$\mu = 2\sum_\ell \alpha_\ell \epsilon_\ell \tag{3}$$

$$V_A = 2\sum_\ell \alpha_\ell^2 \epsilon_\ell \left(1 - \epsilon_\ell\right) \tag{4}$$

where ϵ_ℓ is the mean of $\vec{p}_\ell$ across populations. The matrix $\mathbf{F}$ gives the population level co-ancestry among populations. Here, we calculate the matrix $\mathbf{F}$ directly from the same set of SNPs used to calculate polygenic scores, which is a conservative procedure. Concretely, let

$$\vec{x}_\ell = \frac{\vec{p}_\ell - \epsilon_\ell}{\sqrt{\epsilon_\ell\left(1 - \epsilon_\ell\right)}}. \tag{5}$$

Then, if $\mathbf{X}$ is a matrix with the $\vec{x}_\ell$ as columns, we have

$$\mathbf{F} = \frac{1}{L-1}\mathbf{X}\mathbf{X}^T. \tag{6}$$

Now, based on this null model, we perform two separate neutrality tests. One is a general over-dispersion test (i.e. the "Q_X test" from Berg and Coop (2014)), for which the test statistic is

$$Q_X = \frac{\left(\vec{Z} - \mu\right)^T \mathbf{F}^{-1}\left(\vec{Z} - \mu\right)}{2V_A}. \tag{7}$$

For M populations, this statistic is expected to have a χ^2_{M-1} distribution under the multivariate normal null model (Eq 2). An unusually large value of Q_X indicates that the neutral null model is a poor fit, and is therefore taken as evidence in favor of selection.

We also apply a second, more specific test, to test for evidence of a correlation with a specific geographic axis that is unusually strong compared to the neutral expectation. For any vector $\vec{Y}$, if $\vec{Z}$ has a multivariate normal distribution given by Eq 2, then

$$\vec{Y}^T\vec{Z} \sim N\left(\mu\vec{Y}^T\vec{1}, 2V_A\vec{Y}^T\mathbf{F}\vec{Y}\right) \tag{8}$$

and therefore

$$\left(\frac{\vec{Y}^T\vec{Z} - \mu\vec{Y}^T\vec{1}}{2V_A\vec{Y}^T\mathbf{F}\vec{Y}}\right)^2 \sim \chi^2_1 \tag{9}$$

under the multivariate normal null. This fact can be used to test for an unexpectedly strong association between polygenic scores and a geographic axis by choosing $\vec{Y}$ to be the vector of latitudes or longitude across populations.

S3 Expanded SNP Sets

Some analyses of polygenic score variation among populations have used many more than the ~ 1700 SNPs we use in our main text analyses, in the hope of increasing power to detect adaptive divergence (e.g. [5]). Here, we use three alternative ascertainment schemes that increase the number of SNPs used, and apply them to the UKB-GB GWAS to determine the resulting effect on the signal of selection:

20k: 19,848 genotyped SNPs ascertained from the UKB-GB dataset by running plink's clumping procedure with $r^2 < 0.1$, a maximum clump size of 1Mb, $p < 0.01$, and using 10,000 randomly selected unrelated White British samples as the reference for LD structure.
5k: 4,880 SNPs with the smallest p values subsampled from the 20k ascertainment.
HapMap5k: 5,675 SNPs ascertained from UKB-GB GWAS SNPs after first restricting to HapMap3 SNPs [34], using the same plink clumping procedure as the 20k ascertainment. This HapMap3 ascertainment was performed in order to mimic the ascertainment in [5].

For each expanded SNP set, we applied both the general Q_X test for overdispersion, as well as the specific test for a correlation with latitude (both tests are outlined in Section S2). In all three datasets, the relationship between polygenic scores and latitude was consistent with neutrality. However, in both the 20k and HapMap5k datasets, we can reject the neutral model, as the Q_X p value is 1.68×10^{-3} for 20k and 9.88×10^{-9} for HapMap5k. On the other hand, 5k is not significant, with a p value of 0.75.

We were concerned that the rejection of the neutral null with 20k and HapMap5k ascertainments may be partly due to the higher proximity of SNPs included—leading to deviations from the independent evolution assumption of the neutral model underlying the Q_X hypothesis test (Section S2). To investigate this, we leveraged a decomposition of the Q_X statistic in terms of the underlying loci used to calculate the polygenic scores. Specifically, Berg and Coop (2014) showed that Q_X can be expressed in terms of and "F_{ST}-like" component, which describes the extent to which loci underlying the polygenic scores are marginally overdispersed, and an "LD-like" component, which describes the extent to which pairs of loci which affect the trait covary in their allele frequencies across populations. This decomposition can be written as

$$Q_X = \underbrace{(M-1)\frac{2\sum_\ell \alpha_\ell^2 Var\left(\vec{p}_\ell\right)}{V_A}}_{F_{ST}-\text{like term}} + \underbrace{(M-1)\frac{2\sum_{\ell\neq\ell'} \alpha_\ell\alpha_{\ell'} Cov\left(\vec{p}_\ell,\vec{p}_{\ell'}\right)}{V_A}}_{\text{LD-like term}}. \tag{10}$$

Here, we have assumed that the allele frequencies, p_ℓ, have been transformed so as to remove the influence of population structure. See the discussion surrounding equations 12-14 in Berg and Coop (2014) for a more complete explanation of this transformation.

Here, we extend this decomposition further, breaking the LD-like term into components as a function of the degree of physical separation of SNPs along the chromsome. Specifically, we define a set of partial Q_X statistics ($pQ_X(k)$), such that $pQ_X(k)$ gives the contribution to Q_X from sites which are k SNPs apart on the chromosome (where only SNPs included in the a given ascertianment are included for the purposes of counting how many SNPs apart any two SNPs are).

$$pQ_X(k) = (M-1)\frac{2\sum_{\ell,\ell'\in\mathbb{S}_k} \alpha_\ell\alpha_{\ell'} Cov\left(\vec{p}_\ell,\vec{p}_{\ell'}\right)}{V_A} \tag{11}$$

where $\mathbb{S}_k$ denotes the set of SNP pairs which are k SNPs apart on the same chromosome. So $pQ_X(0)$ would give the "F_{ST} term", while $pQ_X(1)$ gives the component of the "LD term" that comes from covariance between pairs of SNPs which do not have another SNP (that is included in the polygenic scores) physically located between them. $pQ_X(2)$ gives the component that comes from covariance between pairs of SNPs separated by exactly one other SNP included in the polygenic scores, $pQ_X(3)$ the component from SNPs separated by exactly two intervening SNPs, etc. We let $\mathbb{S}_\infty$ be the set of pairs which are on separate chromosomes, so that $pQ_X(\infty)$ gives the contribution to Q_X coming from pairs of SNPs on different chromosomes.

This decomposition retains the property that

$$Q_X = \sum_{k=0}^{K_{\max}} pQ_X(k) + pQ_X(\infty), \tag{12}$$

where $K_{\max}$ is the maximum separation of two SNPs on any chromosome. We note that this in no longer strictly a decomposition of Q_X as the $pQ_X(k)$ terms are not independent of one another, but they are uncorrelated under the neutral null.

In figures and S1, S2, and S3, we show the pQx statistics for various k values in these three different ascertainments. In both the 20k and the HapMap5k ascertainments, pQ_X is higher for low k values–i.e. there is more signal coming from covariance among SNP pairss which are physically close to one another on the chromosome than from distant pairs. This indicates a role for linkage in generating the signals detected in these two ascertainments. In contrast, we see no linkage-associated signal in the 5k ascertainment (in fact, we see no signal whatsoever). The major difference between the signal we observe in the 20k ascertainment and that in the HapMap5k ascertainment is that $pQ_X(\infty)$ is strongly positive for the HapMap5k ascertainment, whereas it is weakly negative for the 20k ascertainment. This difference in the strength of between-population LD between loci on separate chromosomes is largely responsible for the fact that the neutral null hypothesis is strongly rejected for the 20k ascertainment, but only weakly so for the HapMap5k ascertainment.

This heterogeneity of signals across different ascertainments suggests that the signals we do observe are unlikely to be the result of selection—but rather result from some other process or phenomenon which we do not fully understand. Perhaps the most unusual observation is the fact that the among chromosome component of Q_X (i.e. $pQ_X(\infty)$) is so strong from HapMap5k, when it is absent under all other ascertainments. This suggests a role for some ascertainment bias impacting SNPs included in the HapMap3 SNP set. This seems plausible, as SNPs included in the HapMap3 SNP set have an elevated minor allele frequency as compared to a genome-wide sample.While it seems plausible that patterns of among population LD would differ for SNPs included on genotyping platforms, it is not clear why among-population LD should be systematically positive with respect to the SNPs' effect on height.

To better understand the signal observed in the HapMap5k ascertainment, we make use of an alternate decomposition of the Q_X statistic. First, write the eigenvector decomposition of $\mathbf{F}$ as $\mathbf{U\Lambda U^T}$. The mth column of $\mathbf{U}$ ($\vec{U}_m$) gives the mth eigenvector of $\mathbf{F}$, and the mth diagonal entry of $\mathbf{\Lambda}$ (λ_m) gives the mth eigenvalue of $\mathbf{F}$. Note that because this eigen-decomposition is performed on the population level covariance matrix, they capture only the major axes of variation among our pre-specified population labels, in contrast to how PCA is usually done at the individual level in demographic inference applications. Now, we can define a statistic

$$Q_U(m) = \frac{\left(\left(\vec{Z} - \mu\right)^T \vec{U}_m\right)^2}{2\lambda_m V_A} \tag{13}$$

which has a χ^2_1 distribution under the neutral null hypothesis. These statistics, like the pQ_X statistics, have the property that Q_X is given simply by their sum:

$$Q_X = \frac{\left(\vec{Z} - \mu\right)^T \mathbf{F}^{-1} \left(\vec{Z} - \mu\right)}{2V_A} \tag{14}$$

$$= \frac{\left(\vec{Z} - \mu\right)^T \mathbf{U\Lambda}^{-1}\mathbf{U}^T \left(\vec{Z} - \mu\right)}{2V_A} \tag{15}$$

$$= \sum_m \frac{\left(\left(\vec{Z} - \mu\right)^T \vec{U}_m\right)^2}{2\lambda_m V_A} \tag{16}$$

$$= \sum_m Q_U(m). \tag{17}$$

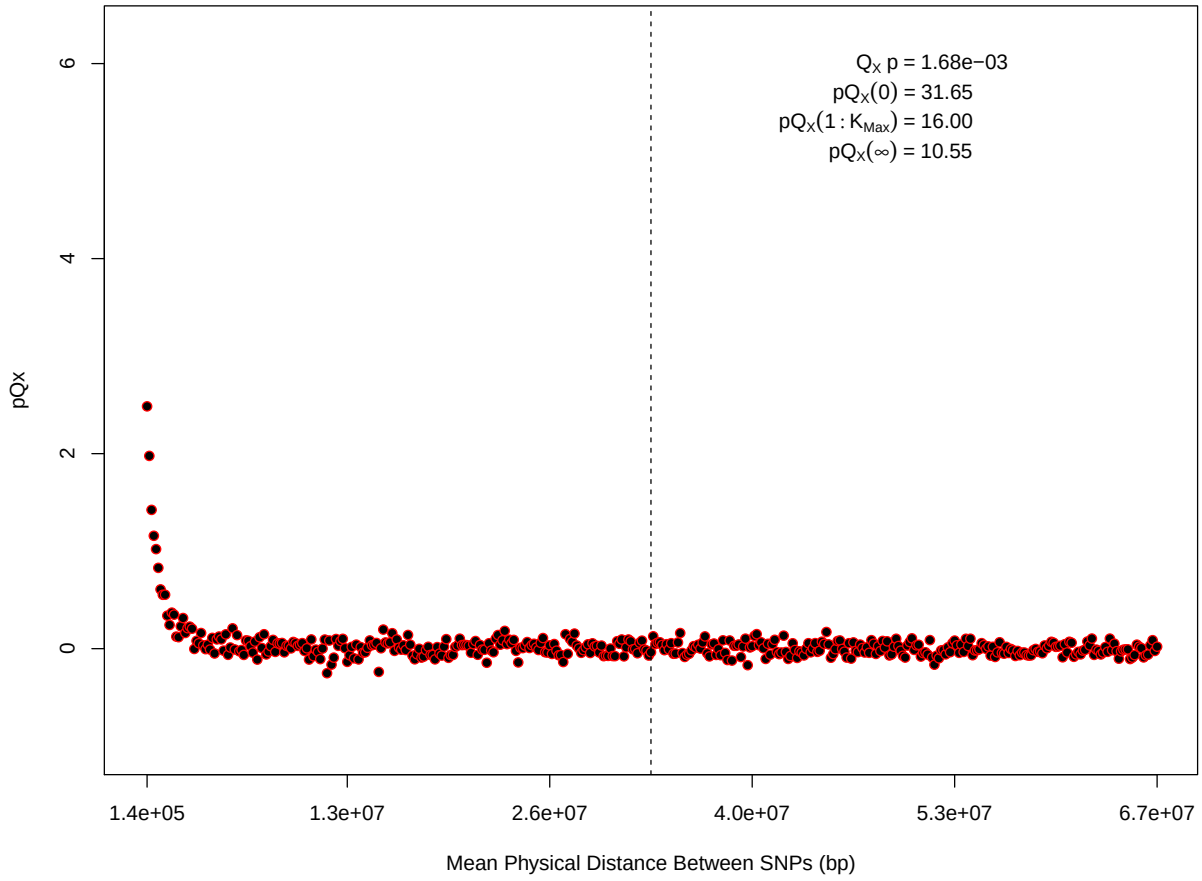

Figure S1: *$pQ_X(k)$ statistics for $k = 1 : 450$ for the 20k dataset. The x axis gives the average physical distance between all pairs of SNPs contributing to a given $pQ_X(k)$ statistic. The uptick in $pQ_X(k)$ on the left side of the plot (i.e. small values of k) indicates that SNPs which are physically close to one another and have the same sign in their effect on height covary across population disproportionately as compared to more distant pairs of SNPs. Note that the number of pairs of SNPs ($|\mathbb{S}_k|$) contributing to a given $pQ_X(k)$ decreases as k increases, as smaller chromosomes have fewer pairs at larger distances than they do at shorter distances. This leads to a decrease in the variance of $pQ_X(k)$ under the null as k increases. However, this decline in variance is not responsible for the decay in signal as k increases, as $|\mathbb{S}_k|$ remains approximately constant until well past the dashed vertical line, which indicates the distance between between the ends of chromosome 21 (the shortest chromosome, and therefore the first to drop out of the $pQ_X(k)$ calculation).*

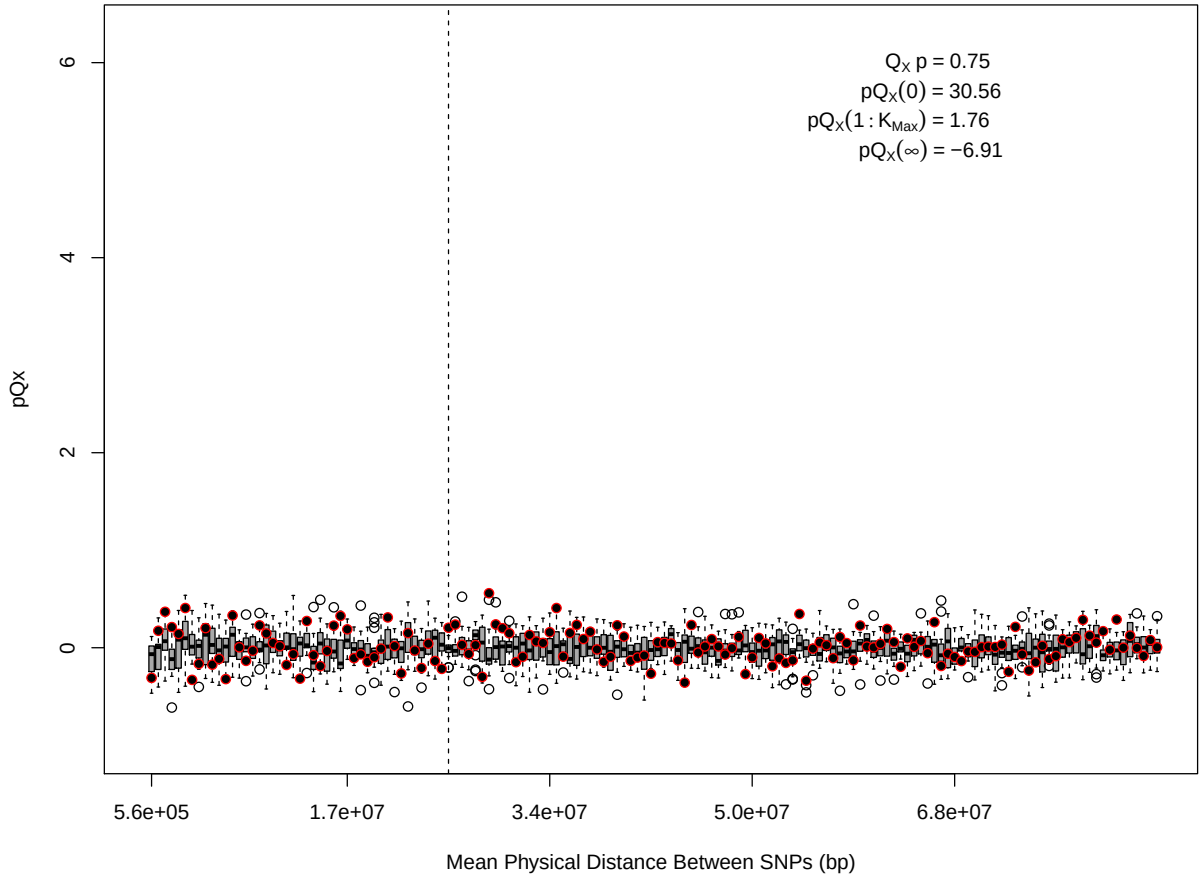

Figure S2: *$pQ_X(k)$ statistics for $k = 1 : 150$ for the 5k dataset. The x axis gives the average physical distance between all pairs of SNPs contributing to a given $pQ_X(k)$ statistic. The boxplots give an empirical null distribution of $pQ_X(k)$ statistics derived from permuting the signs of all effect sizes independently (this empirical null was omitted from Figure S1 due to computational expense). In this case, SNPs that are physically close to one another do not contribute disproportionately to the signal.*

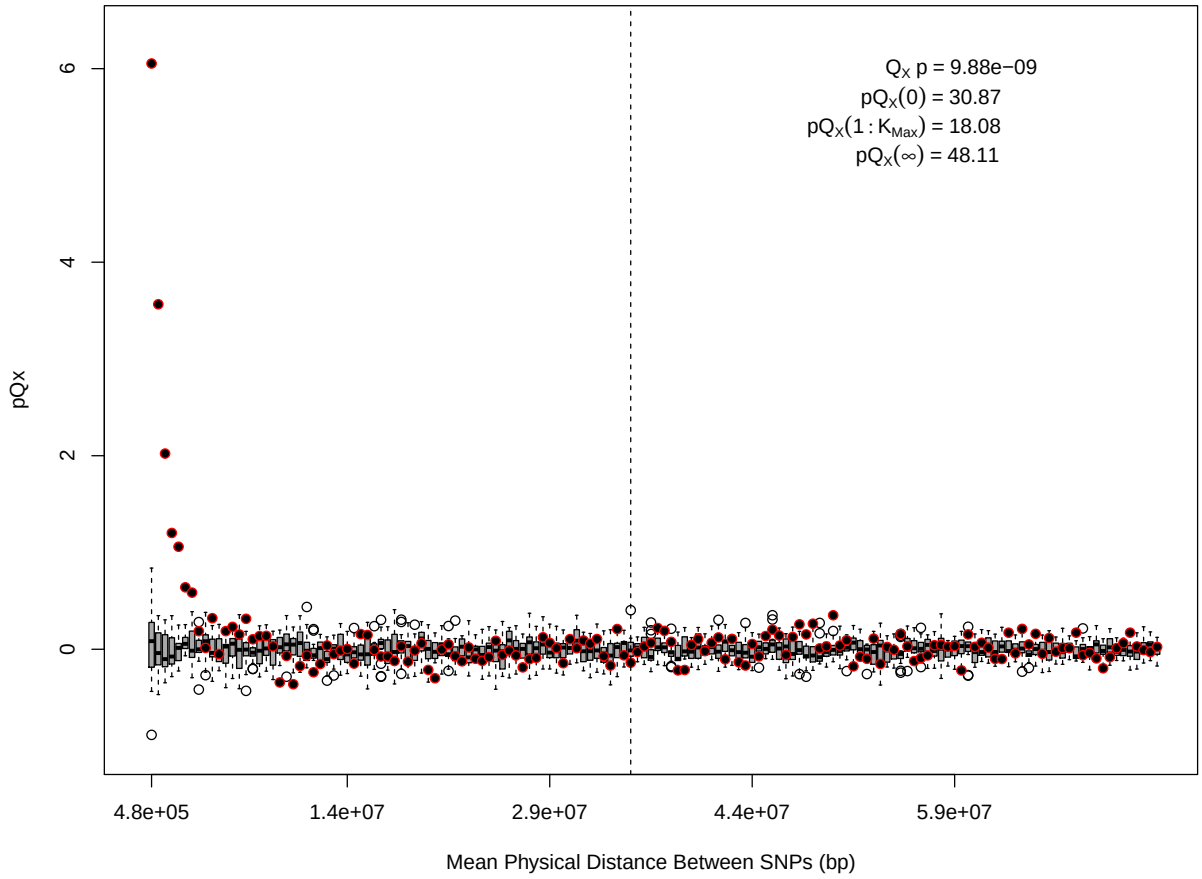

Figure S3: *$pQ_X(k)$ statistics for $k = 1 : 150$ for the HapMap5k dataset. The x axis gives the average physical distance between all pairs of SNPs contributing to a given $pQ_X(k)$ statistic. The boxplots give an empirical null distribution of $pQ_X(k)$ statistics derived from permuting the signs of all effect sizes independently (this empirical null was omitted from Figure S1 due to computational expense). The uptick in signal from pairs of SNPs physically nearby to one another is present in this dataset, again suggesting a role for physical linkage in contributing to the signal. However, note that in contrast to the 20k and 5k ascertainments, the HapMap5k ascertainment also has a large amount of signal from $pQ_X(\infty)$, which cannot be explained by linkage.*

An unusually large value of $Q_U(m)$ for a given choice of m is an indication that the polygenic scores are more strongly correlated with the mth axis of population structure than expected under the neutral null model. Therefore, once a signal is detected with Q_X, the Q_U statistics can be used to understand which specific axes of divergence among populations are responsible for generating the signal in Q_X.

In Figure S4, we show a QQ plot of the $-log_{10}$ p values for the HapMap5k ascertainment, derived from comparing these Q_U statistics from the European set of populations to the χ^2_1 distribution. It is particularly noteworthy that the signal in this ascertainment is diffuse, resulting from inflation of nearly all of the Q_U statistics, rather than just a few. This is a statement that the signal detected in the HapMap5k ascertainment results from the polygenic scores simply being more variable along all axes, rather than one particular axis of population structure. In general, we are skeptical that this represents a real signal of selection, particularly given how sensitive it is to ascertainment.

One hypothesis is that the HapMap5k ascertainment could suggest ancient assortative mating on the basis of height. Specifically, our neutral null model assumes that all loci drift independently. However, assortative mating on the basis of a phenotype will lead to a build-up of within population LD that is positive with respect to the direction of allelic effects on the trait—even among distant or unlinked alleles. As populations drifted apart, within-population LD due to assortative mating would get converted into among-population LD—causing a deviation from our null-model assumption of independent evolution across all loci. This phenomenon would result in populations drifting apart in height-associated loci faster than expected by the rest of the genome. This hypothesis is consistent with the diffusion of Q_X across all Q_U terms in HapMap5k. This hypothesis is also consistent with higher pQ_X for physically proximate SNPs, as assortative mating would leads to a stronger buildup of trait LD among pairs of loci which are tightly linked than for those that are not, which would lead to stronger among population LD among these loci as populations diverge

However, under this hypothesis, it is not clear why we would expect the uptick in $pQx(k)$ for small k to be present in the HapMap5k and 20k datasets but not the 5k dataset, or why the $pQx(\infty)$ signal should be present in only the HapMap5k dataset. At this point, we leave the assortative mating hypothesis outlined above as speculative, and leave further investigations for future work.

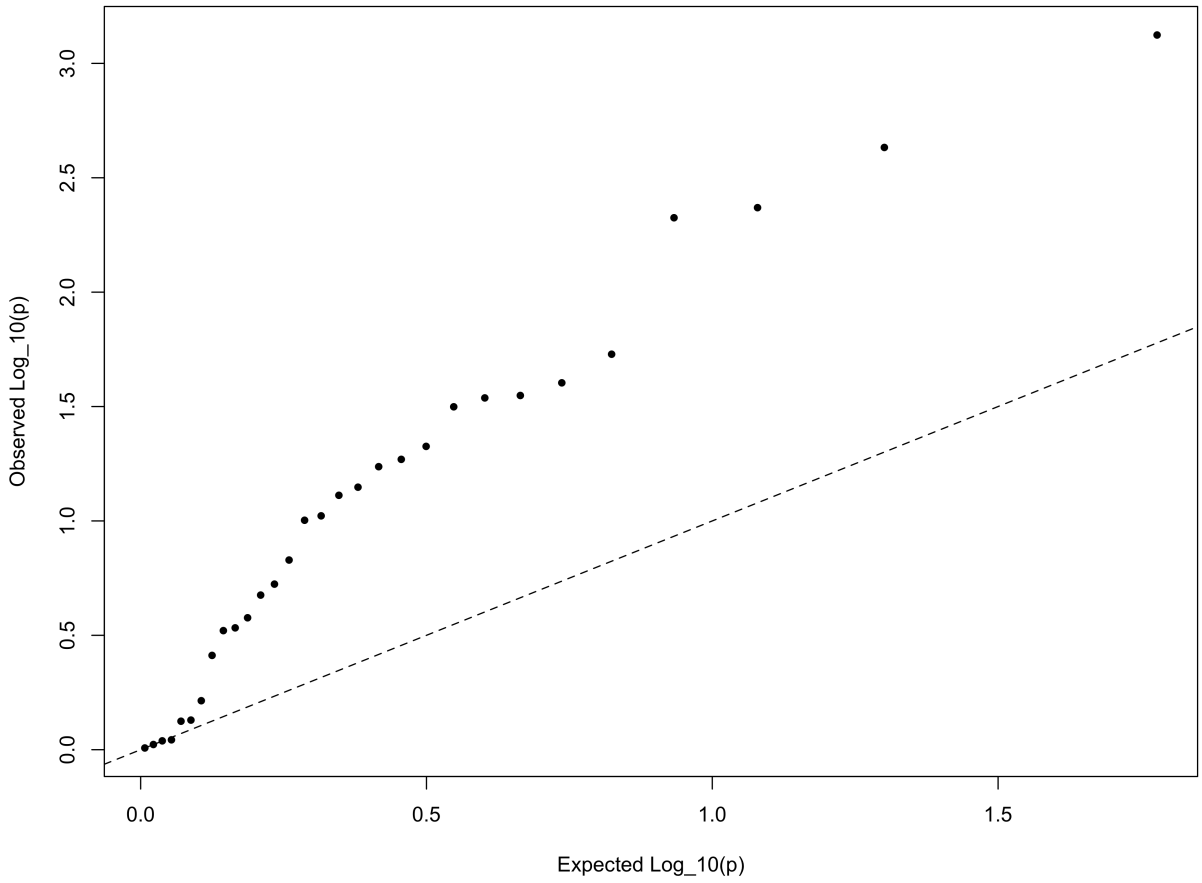

Figure S4: *The QQ plot of $-log_{10}$ P values for the Q_U statistics calculated a within-Europe sample using the HapMap5k ascertainment. The systematic inflation of Q_U indicates a non-specific rejection of neutrality: polygenic scores are more variable in all directions. This pattern is not expected under adaptive divergence of labeled populations.*

S4 Robustness of differences in differentiation signal to filtering schemes

We attempted to replicate the signal for polygenic adaptation on height previously found using GWAS data [14] from the GIANT consortium [3, 4, 7, 5], but this time using the GWAS on standing height performed by the Neale lab with the UK Biobank (UKB) dataset. In particular, we were interested in checking if the failure to replicate the signal was due to filtering on particular sets of SNPs in one or the other dataset.

We focused on present-day populations from phase 3 of the 1000 Genomes Project [35] (Figure S5). We divided the genome into 1,700 approximately independent LD blocks, using fgwas [25, 36], and extracted, for each of the two GWAS for height, the SNP with the highest posterior probability of association (PPA) from each block, using. This resulted in a total of 1,700 SNPs (1 per block). Unless otherwise stated, we computed scores using the subset of these SNPs that were located in blocks with high per-block posterior probability of association ($PPA > 95\%$), after retrieving the allele frequencies of these SNPs in the 1000 Genomes population panels, using glactools [37]. We tested different types of filters to assess how they influenced the results.

Figure S6 (upper row) shows that genetic scores computed for each of the 1000 Genomes phase 3 populations. In each plot below in which we report a P-value, this P-value comes from calculating the Q_X statistic, and assuming this statistic is chi-squared distributed [4, 7]. The candidate SNPs used for calculating the genetic scores were filtered so that the average minor allele frequency across populations was more than or equal to 5%.

To investigate the effect of the per-block posterior probability of association (Block PPA) on the genetic scores, we also used two alternative PPA thresholds for including a block in the computation of the PPA score: 0 (i.e. including all blocks, lower row of Figure S6) and 0.5 (middle row of Figure S6) shows that this filtering has little effect in the difference in results between the two GWASs.

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

Figure S5: Present-day populations from 1000 Genomes Project Phase 3 used to build population-level polygenic scores, colored by their respective super-population code

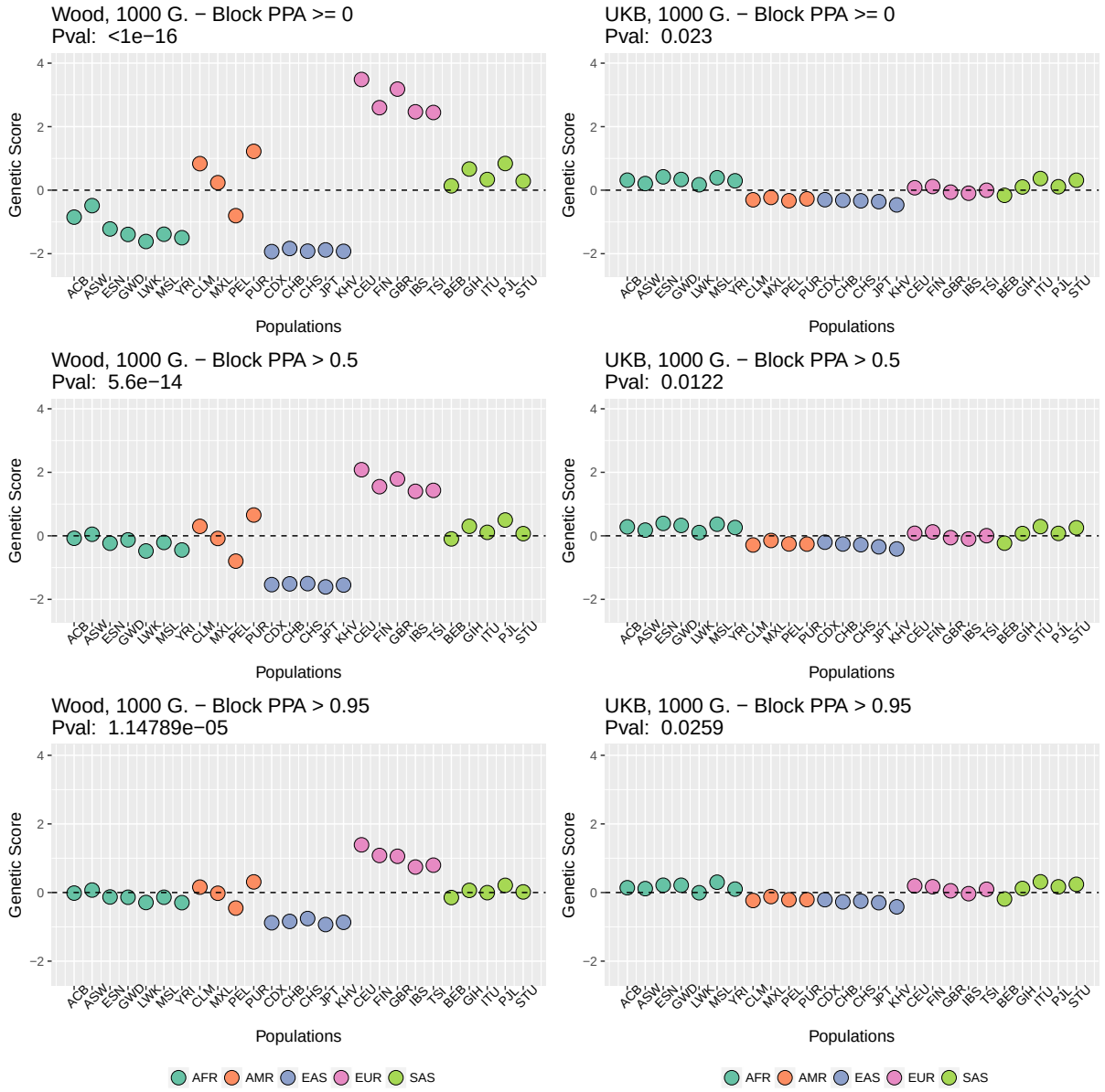

Figure S6: Genetic scores in present-day populations, colored by their super-population code, and created using different block-PPA thresholds. Left column: Wood et al. GWAS. Right column: Neale lab UK Biobank GWAS

To visualize the contribution of each SNP to the difference in scores between two populations with high differentiation in the Wood et al. GWAS (CHB and CEU), we produced a contour plot in which we display the absolute effect size of each SNP contributing in the computation of the genetics scores, plotted as a function of the difference in the frequency of the trait-increasing allele for that SNP in the two populations (Figure S7).

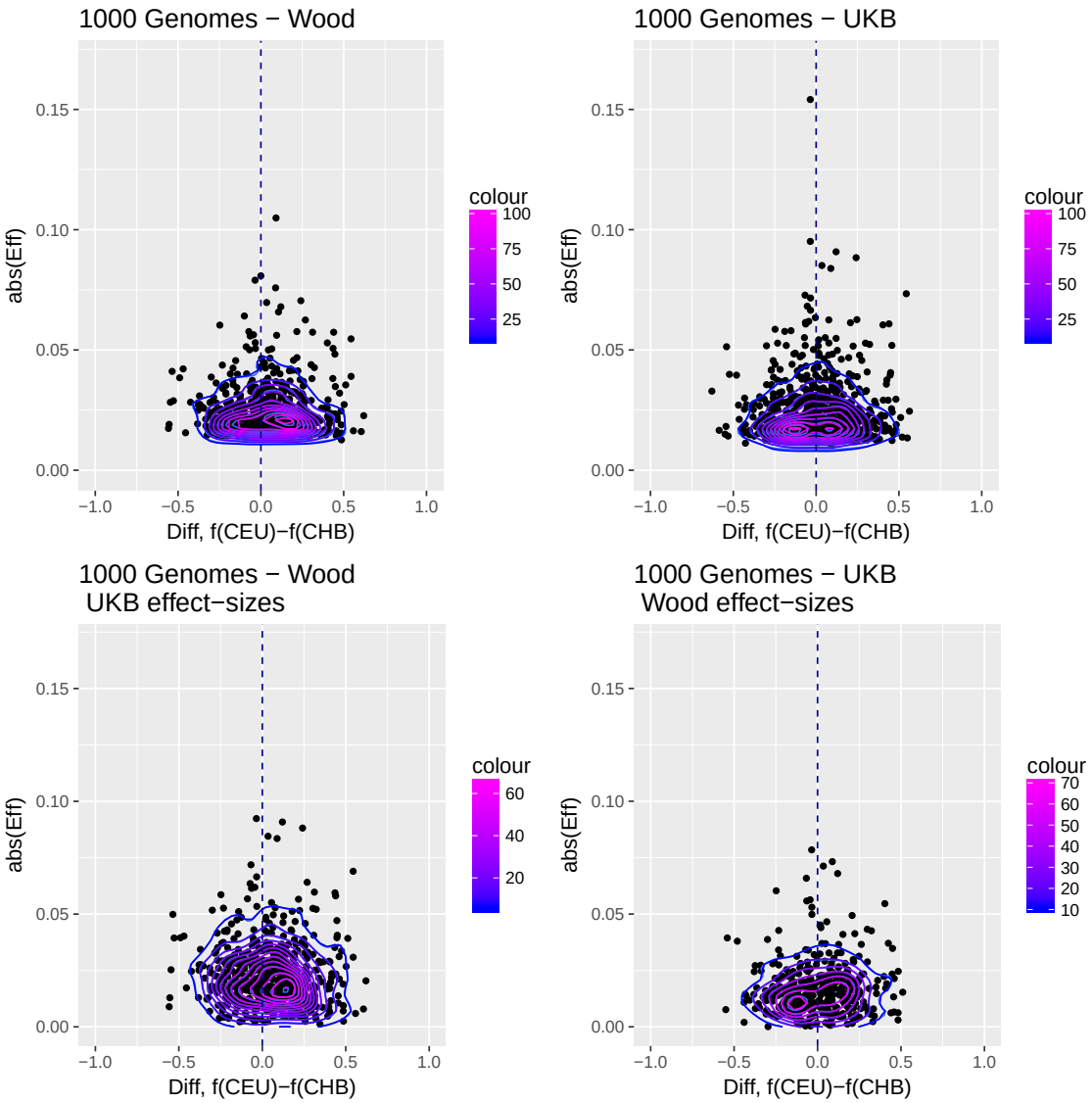

Figure S7: Distribution of the absolute value of effect sizes (y-axis) plotted as a function of the difference in frequency of the trait-increasing allele between CEU and CHB (x-axis), for candidate SNPs used to build genetic scores. Top left: trait-associated SNPs from Wood et al., with effect sizes from the same GWAS. Top right: trait-associated SNPs from the Neale lab GWAS, with effect sizes from the same GWAS. Bottom left: trait-associated SNPs from Wood et al., but with their corresponding effect sizes from the Neale lab GWAS. Bottom right: trait-associated SNPs from the Neale lab GWAS, but with their corresponding effect sizes from Wood et al. Contour colors denote the density of SNPs in different regions of each plot.

Figure S7 shows that the distribution of the difference in scores between the two populations is shifted in favor of CEU when using the Wood et al. dataset, but not when using the UKB dataset. When selecting SNPs via PPAs from the Wood et al. dataset but using their UKB effect sizes, the distribution of differences is also shifted in favor of CEU, but this does not occur when performing the converse: using PPAs from UKB to select SNPs, but plotting their effect sizes from Wood et al.

This figure also reveals that there are a number of SNPs in the UKB dataset with high effect sizes and very small differences in allele frequency between the two populations. These SNPs tend to have allele frequencies near the boundaries of extinction or fixation in both populations, suggesting they could possibly be under

the influence of negative selection. To investigate the contribution of these high-effect SNPs on the overall genetic scores with the UKB dataset, we removed their corresponding blocks from the score computation, and re-calculated the genetic scores for all populations. We chose a minimum absolute effect size equal to 0.08 for removal of SNPs, and the 6 SNPs in the UKB dataset which are above this threshold were therefore excluded from the analysis. This filtering, however, does not seem to serve to recover the Wood et al. signal (Figure S8).

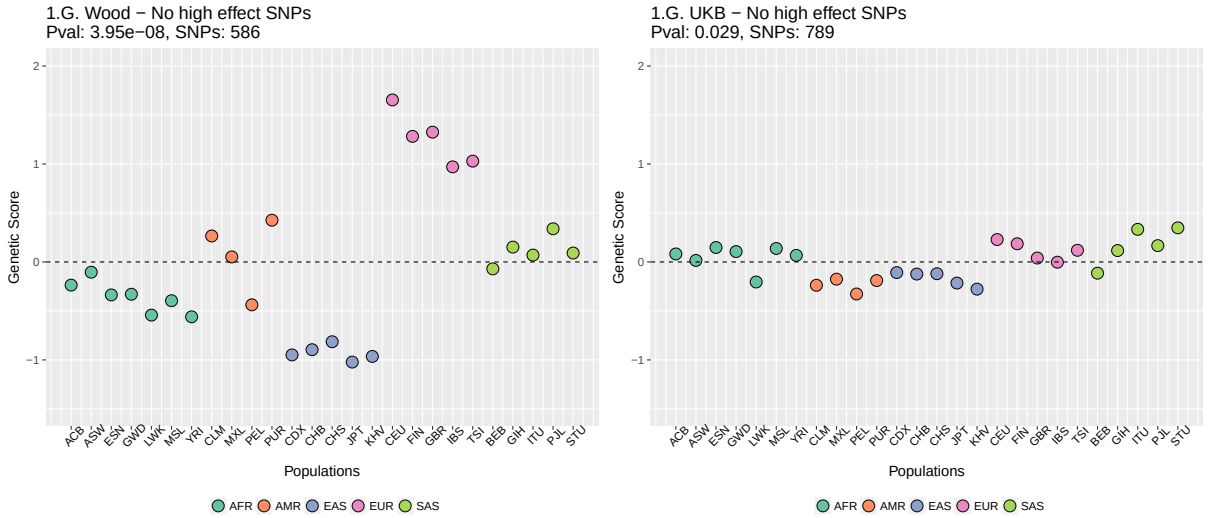

Figure S8: Genetic scores for present-day populations, after excluding 6 high-effect SNPs from UKB, colored by super-population code. Left: Wood et al. GWAS. Right: Neale lab UK Biobank GWAS.

In Figure S9 we restrict the candidate SNPs used, by only allowing SNPs that have minor allele frequencies larger than 0.05 in all populations. This is different from our previous default allele frequency filtering, in which we only required the average of the minor allele frequency across populations to be larger than 0.05. Nevertheless, this filtering does not recover the Wood et al. signal either.

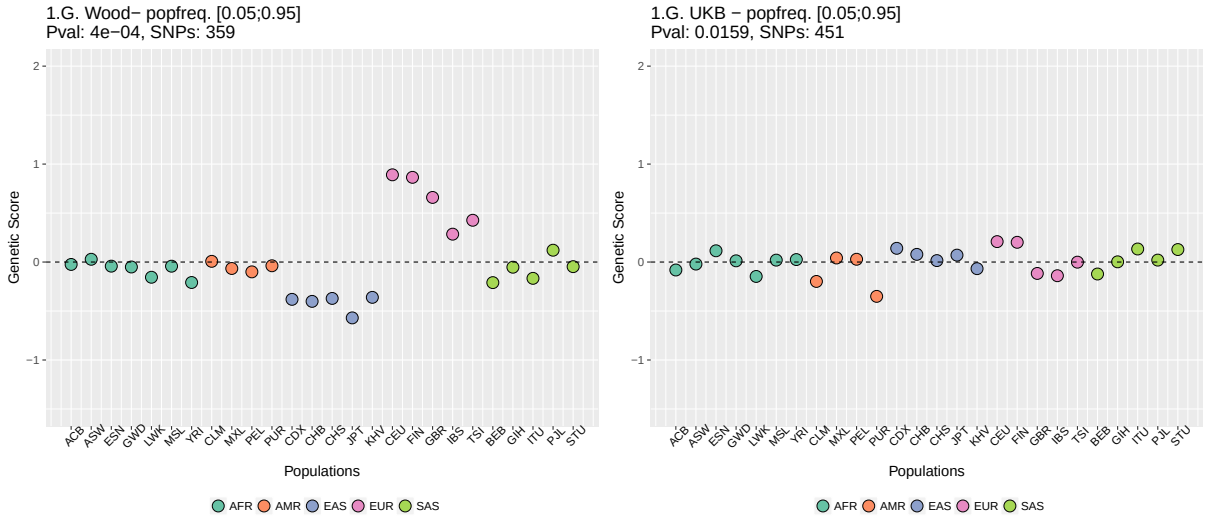

Figure S9: Genetic scores computed only with SNPs that have minor allele frequencies larger than 0.05 in all populations. Left: Wood et al. GWAS. Right: Neale lab UK Biobank GWAS.

We also looked into whether the candidate SNPs found using the UK Biobank dataset were also present in the Wood et al. GWAS, but perhaps with much smaller effect sizes, and this was somehow affecting the genetic scores made using the UKB data. In figure S10 all UK Biobank candidate SNPs that were also found in Wood et al. were evaluated and if a SNP's absolute effect size in Wood et al. was smaller than or equal to 0.05, the SNP was excluded from the UK Biobank candidate set.

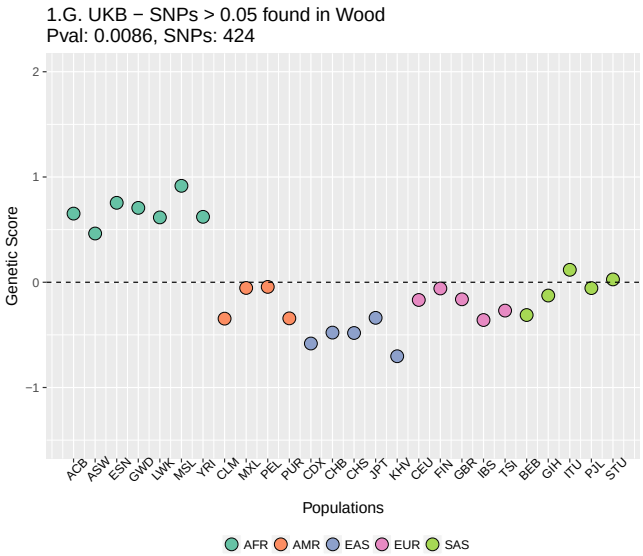

Figure S10: Genetic scores computed using the UK Biobank data, after removing SNPs with absolute effect sizes smaller than or equal to 0.05 in Wood et al.

We also excluded all UKB-candidate SNPs found in Wood et al. with absolute effect sizes smaller than or equal to 0.01, and recomputed the scores using the UK Biobank effect sizes (Figure S11).

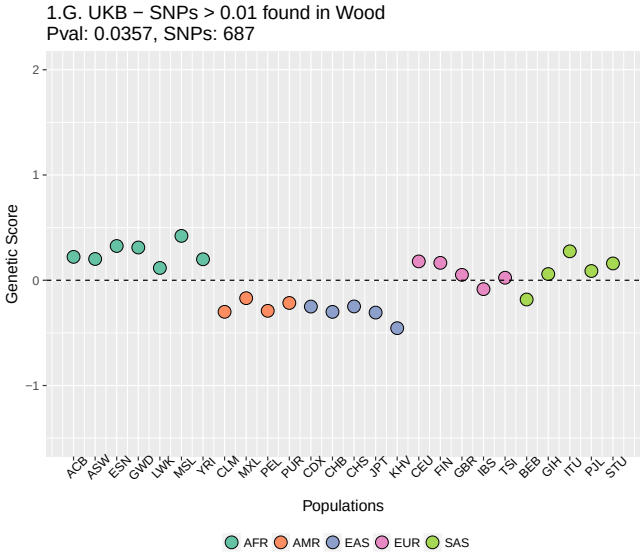

Figure S11: Genetic scores computed using the UK Biobank data, after removing SNPs with absolute effect sizes smaller than or equal to 0.01 in Wood et al.

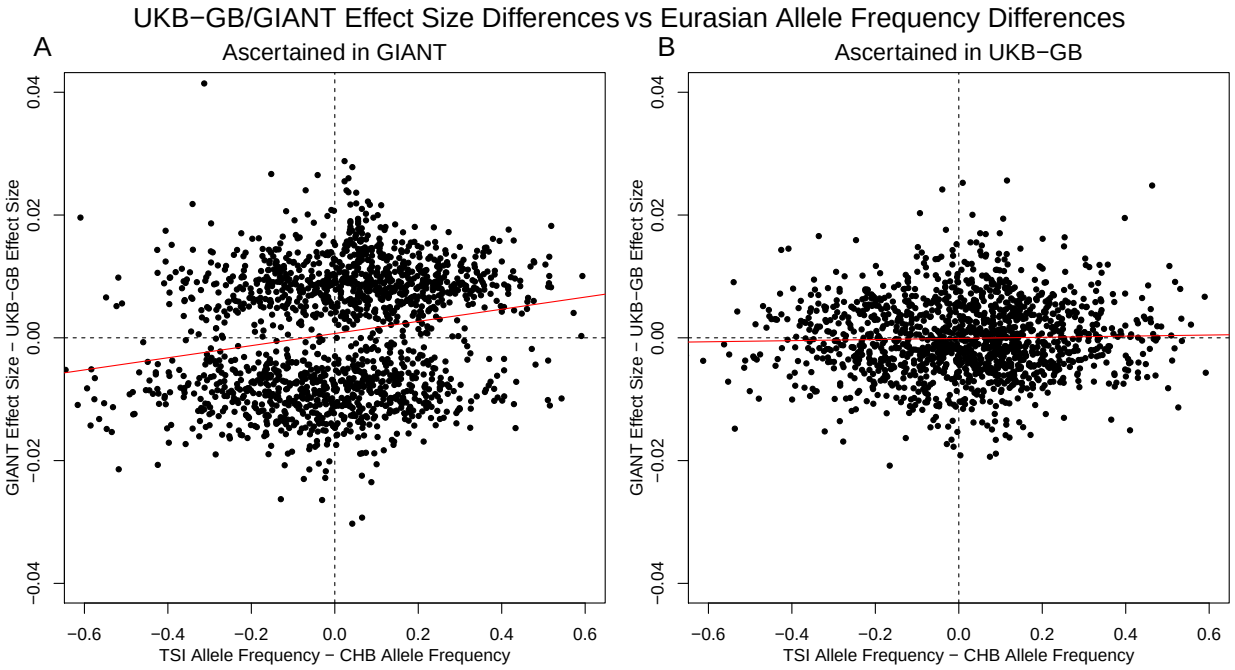

Figure S12: (A) *SNPs are ascertained using GIANT, and the difference in effect sizes between GIANT and UKB-GB is plotted against the difference in frequency between the TSI and CHB 1000 genomes population samples. These variables exhibit a highly significant correlation* ($r^2 = 0.04$, $p = 1.43 \times 10^{-15}$). **(B)** *SNPs are ascertained using UKB-GB, and the difference in effect sizes between GIANT and UKB-GB is again plotted against the difference in frequency between the TSI and CHB 1000 genomes population samples. In this case, there is no significant correlation* ($r^2 = 1.8 \times 10^{-4}$, $p = 0.26$)

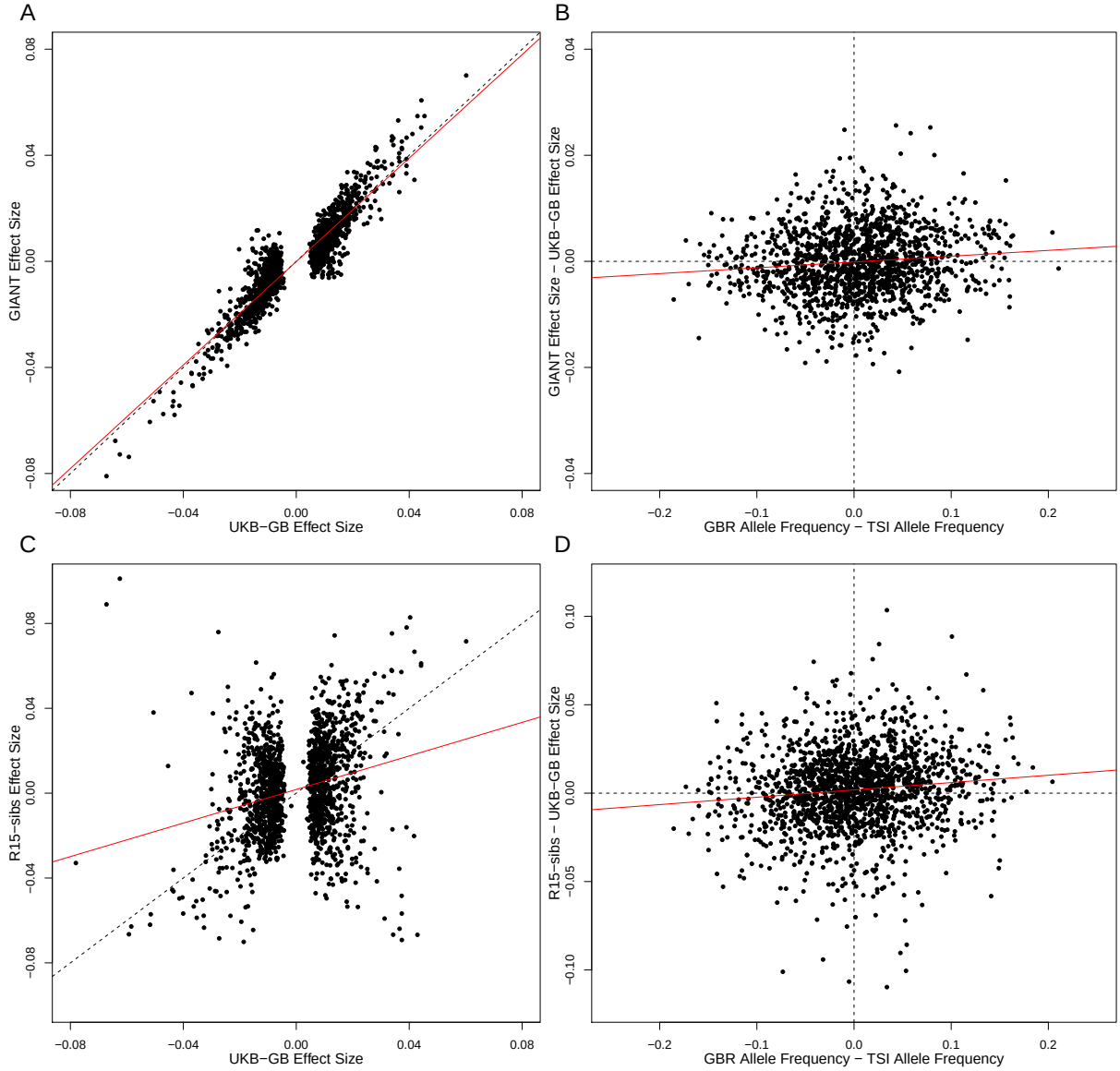

Figure S13: *Each panel in this figure shows the same comparisons as Figure 3, except that SNPs have been ascertained with UKB-GB p values, rather than those of GIANT or R15-sibs. Note that in both cases, the correlation between the difference in effect sizes and the GBR-TSI allele frequency difference is substantially reduced in both cases.*

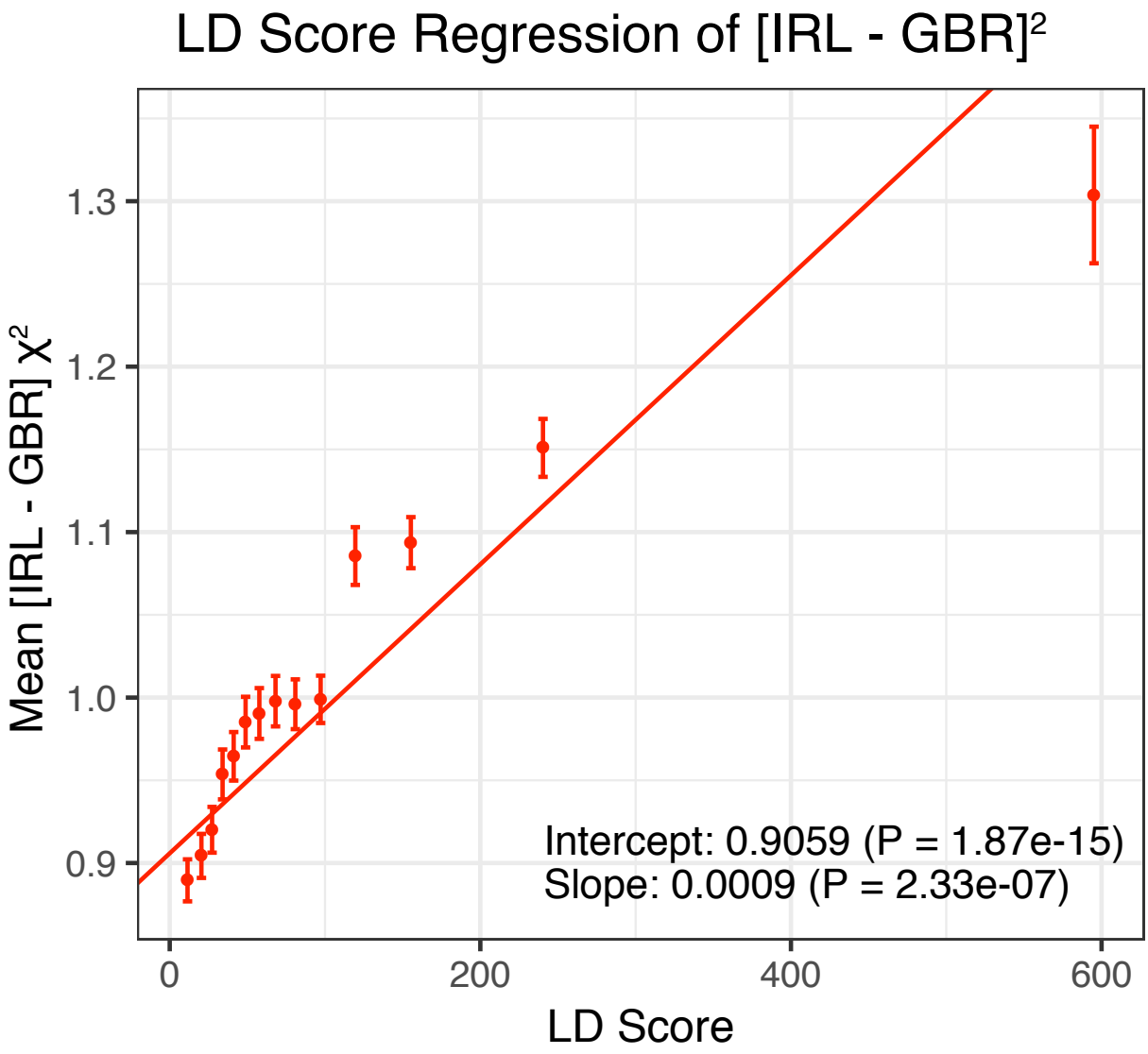

Figure S14: The "heritability" estimate of the difference in allele frequency between individuals who identified as Irish and those who identified as White British in the UK Biobank (related individuals removed).

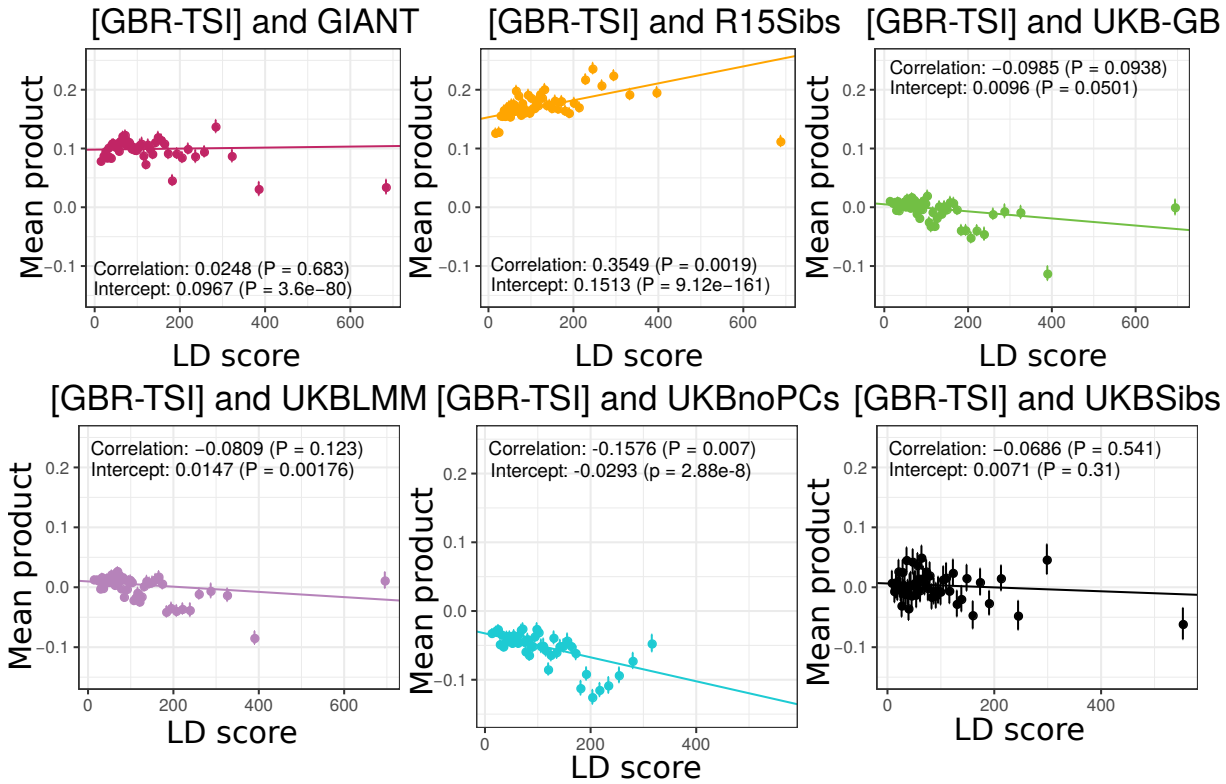

Figure S15: The "genetic correlation" estimates of the difference in allele frequency between GBR and TSI from 1000 Genomes versus each of the height traits under study.

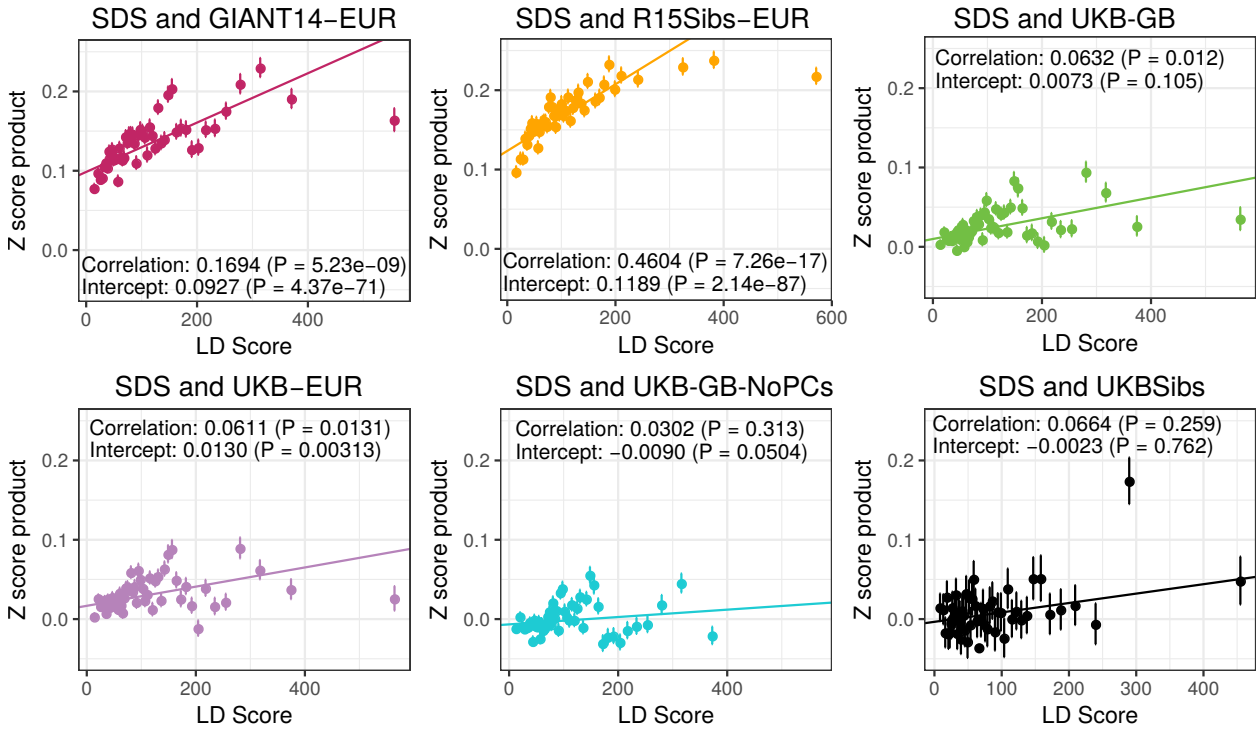

Figure S16: The "genetic correlation" estimates of SDS versus each of the height traits under study.

S5 Assessing significance of the correlation between GWAS p-value and tSDS

Figure 2 illustrates the correlation between tSDS and GWAS p-value (p-value for the strength of association with height). These p-values are computed while accounting for LD between SNPs that might inflate the significance of the correlation between the two. To do this, we used a blocked-jackknife approach [38, 39] to estimate the standard error of our Spearman correlation point estimate, $\hat{\rho}$. For each GWAS, SNPs were assigned to one of $b = 200$ contiguous blocks based on concatenated genomic coordinates. tSDS values should not be correlated across such large blocks. For each block i, we computed the Spearman correlation in the i'th jackknife sample, $\hat{\rho}^b_{(-i)}$—i.e. the Spearman correlation across all SNPs but the SNPs in block i. We then estimated the standard error of the point Spearman estimate by $\hat{\sigma}$, where

$$\hat{\sigma}^2 = \frac{b-1}{b} \sum_{i=1}^{b} (\hat{\rho}^b_{(-i)} - \bar{\rho^b}),$$

and

$$\bar{\rho^b} = \frac{1}{b} \sum_{i=1}^{b} \hat{\rho}^b_{(-i)}$$

is the average of jackknife samples. Finally, we compute a p-value for the null hypothesis

$$H_0 : \rho = 0,$$

by approximating $\hat{\rho}$ as Normally distributed under the null with standard deviation $\hat{\sigma}$, namely

$$\hat{\rho} \sim N(0, \hat{\sigma}).$$

S6 LD score regression and linked selection

In this section we discuss how linked selection, specifically BGS, may be a potential confounder of LD score regression. In the first section (S6.1) we discuss the intuition behind univariate LD score regression and how BGS can cause a correlation between LD score and allele frequency differentiation. In the second section (S6.2) we show empirically how LD score and BGS covary across the genome, and how this can account for the empirical patterns of LD score correlating with allele frequency differentiation. In the third section (S6.3) we show the BGS confounding of the slope and intercept of the univariate LD score regression. In the final section (S6.4) we work through bivariate LD score regression and show that it can be used to highlight the confounding of GWAS by specific axes of population structure.

Through this supplement we discuss the potential issue with linked selection in terms of BGS. However, it is likely that basic intuition of theses results, i.e. that linked selection is confounder of LD score regression, apply more generally to other models of linked selection (e.g. selective sweeps).

S6.1 Introduction

Bulik-Sulivan et al. [19, 20] introduced LD score regression as a robust way to assess the impact of population structure confounding on GWAS, and to robustly assess heritabilities and genetic correlations in GWAS even in the presence of such confounding. The LD score of a SNP (i) is found by summing up LD (R^2) in a genomic window of W surrounding SNPs:

$$\ell_i = \sum_{j=0}^{W} R_{i,j}^2. \tag{18}$$

Following the logic laid out in the appendix of Bulik-Sulivan et al. [19], consider a GWAS done using a sample drawn from two populations, with a sample of $N/2$ draws from each population. The trait is controlled by a very large number of loci (M), and the total narrow-sense heritability of the trait is h_g^2. The GWAS is partially confounded by population structure, as the squared difference in mean phenotype between the populations is a, and the allele frequency differentiation between the populations is F_{ST}. The expected χ_i^2 statistic of the trait association of the ith SNP is

$$\mathbf{E}[\chi_i^2] = \frac{Nh_g^2}{M}\ell_i + 1 + aNF_{ST}, \tag{19}$$

following Equation 2.7 of Bulik-Sulivan et al [19].

The basic idea of LD score regression is that we regress χ_i^2 on ℓ_i, the deviation of the estimated intercept away from 1 gives aNF_{ST}, the confounding by population structure, while the slope of the regression gives $\frac{Nh_g^2}{M}$. Underlying this separation of the confounding effects of population structure (aF_{ST}) and the heritability (h_g^2) is the assumption that F_{ST} is not correlated with LD score. However, as noted by Bulik-Sulivan et al. [19] this assumption maybe violated by background selection (BGS). In short, regions of low recombination (and thus higher LD score) experience more BGS—which in turn drives higher F_{ST} [30].

To a first approximation, the effects of strong BGS in a well-mixed, constant-sized population can be modeled by a reduction in the effective population size, as the rate of drift increases in regions subject to BGS. We can express this mathematically by saying that SNP i experiences an effective population size B_iN_e, where N_e is the effective population size in the absence of BGS and B_i is the reduction due to BGS. The expected LD between SNP i and another SNP L bp apart is

$$\mathbf{E}(R^2) \approx 1/(1 + 4N_eB_ir_{BP,i}L)$$

where $r_{BP,i}$ is the recombination rate surrounding SNP i.

F_{ST}, in turn, is a decreasing function of N_eB_i. For example, if the two populations at hand split T generations ago, without subsequent gene-flow or population size changes,

$$\mathbf{E}(F_{ST}) \approx T/(4N_eB_i) \tag{20}$$

(this approximation holds for small values of T/N_e). Similar inverse dependences of F_{ST} on B_i can be derived in other models of weak population structure [30].

S6.2 Empirical results on LD score and BGS

To explore the empirical relationship between LD score, recombination rate and BGS we make use of the B values estimated along the human genome by McVicker et al [31]. We use the 1000 genomes CEU LD scores [19], and the Kong et al recombination rates [40] (the latter are standardized by the genome-wide average recombination rate).

In Figure S17 we plot the LD score, averaged in 100kb windows, as a function of recombination rates and McVicker's B values. As expected, LD scores are higher in regions of low recombination and regions of stronger background selection (lower B).

Based on a simple model of BGS (Equation 20), $F_{ST} \propto 1/B$, therefore in Figure S18 we plot the relationship between LD scores and $1/B$ values each averaged in 30 quantiles of LD score.

In the main text (Figure 5A and S14) we plotted the relationship between LD score and the χ^2 statistic for allele frequency differentiation. To make our χ^2 statistic comparable to F_{ST} we standardized it. To do this we note that because population membership is not a genetic trait, setting $h^2 = 0$ in Equation 19 we obtain

$$\mathbf{E}[\chi_i^2] = 1 + aNF_{ST}, \tag{21}$$

Therefore, to make our χ_i^2 statistic comparable to F_{ST} we standardize our χ_i^2 as:

$$(\chi_i^2 - 1)/\overline{\chi_i^2}, \tag{22}$$

where the overbar in the denominator signifies a genome-wide average. In Figure S18 we plot the expected relationship between LD score and standardized χ_i^2 predicted under our simple BGS model (plugging in McVicker B values for the intensity of background selection). We compare it to the the empirical relationship between LD score and the standardized χ_i^2 statistics for the Irish-British and GBR-TSI allele frequency differences. The agreement between the empirical results and the BGS-theoretical predictions is reasonable, suggesting that a model of BGS, as parameterized by McVicker's B, could explain the confounding in LD score regression by linked selection.

S6.3 Predicted effect on linked selection on the slope and intercept of LD score regression.

The expectations of the slope and intercept of univariate LD score were derived in the absence of linked selection. In this section we show how these expectations can be distorted by BGS.

In the regression of $\chi_i^2 \sim \ell_i$ the slope is:

$$\beta_{\chi^2;\ell} = \frac{\mathrm{Cov}(\chi_i^2, \ell_i)}{\mathrm{Var}(\ell_i))} \tag{23}$$

$$= \frac{\frac{Nh_g^2}{M}\mathrm{Var}(\ell_i) + aN\mathrm{Cov}(\ell_i, F_{ST,i})}{\mathrm{Var}(\ell_i)} \tag{24}$$

$$= \frac{Nh_g^2}{M} + aN\beta_{F_{ST};\ell} \tag{25}$$

where $\beta_{F_{ST};\ell}$ is the slope of F_{ST} regressed on LD score. Therefore the slope of the univariate LD score regression is biased upwards by linked selection. The intercept is

$$\alpha_{\chi^2;\ell} = \overline{\chi^2} - \beta_{\chi^2;\ell}\overline{\ell} = 1 + aN\big(\overline{F_{ST}} - \beta_{F_{ST};\ell}\overline{\ell}\big), \tag{26}$$

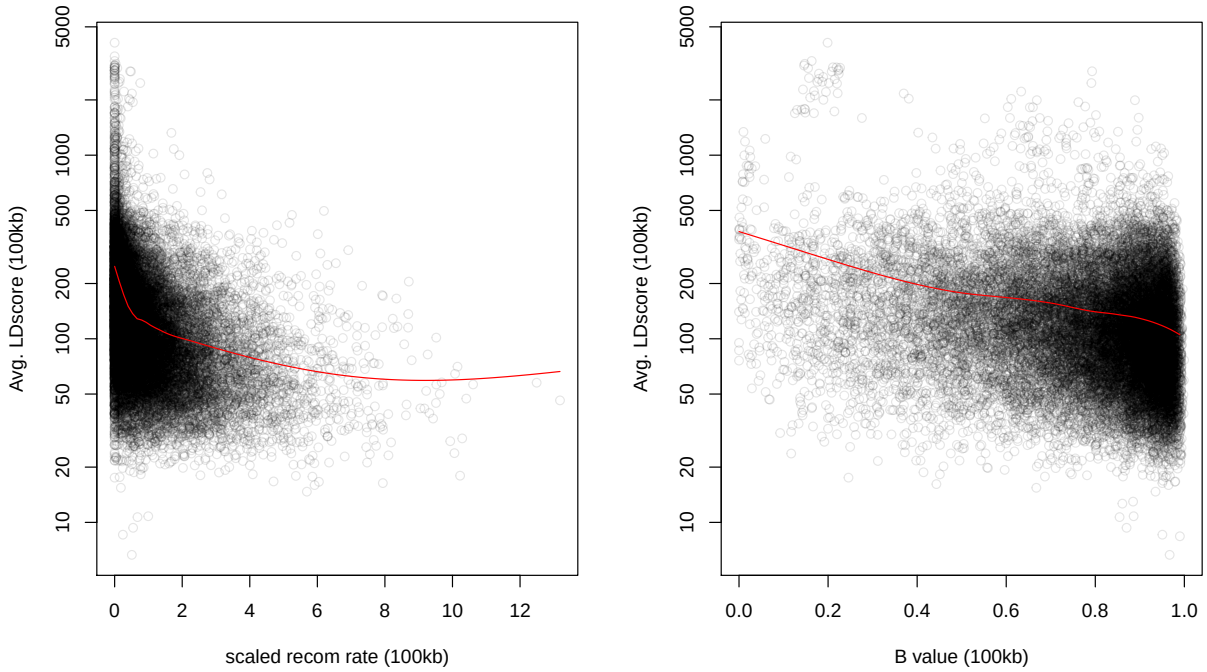

Figure S17: Windows with lower recombination rates and B values have higher LD scores. The autosome is divided into 100 kb windows and the average LD score, B-value, and standardized recombination rate is calculated in each bin. The red lines are a lowess fit as a guide to the eye.

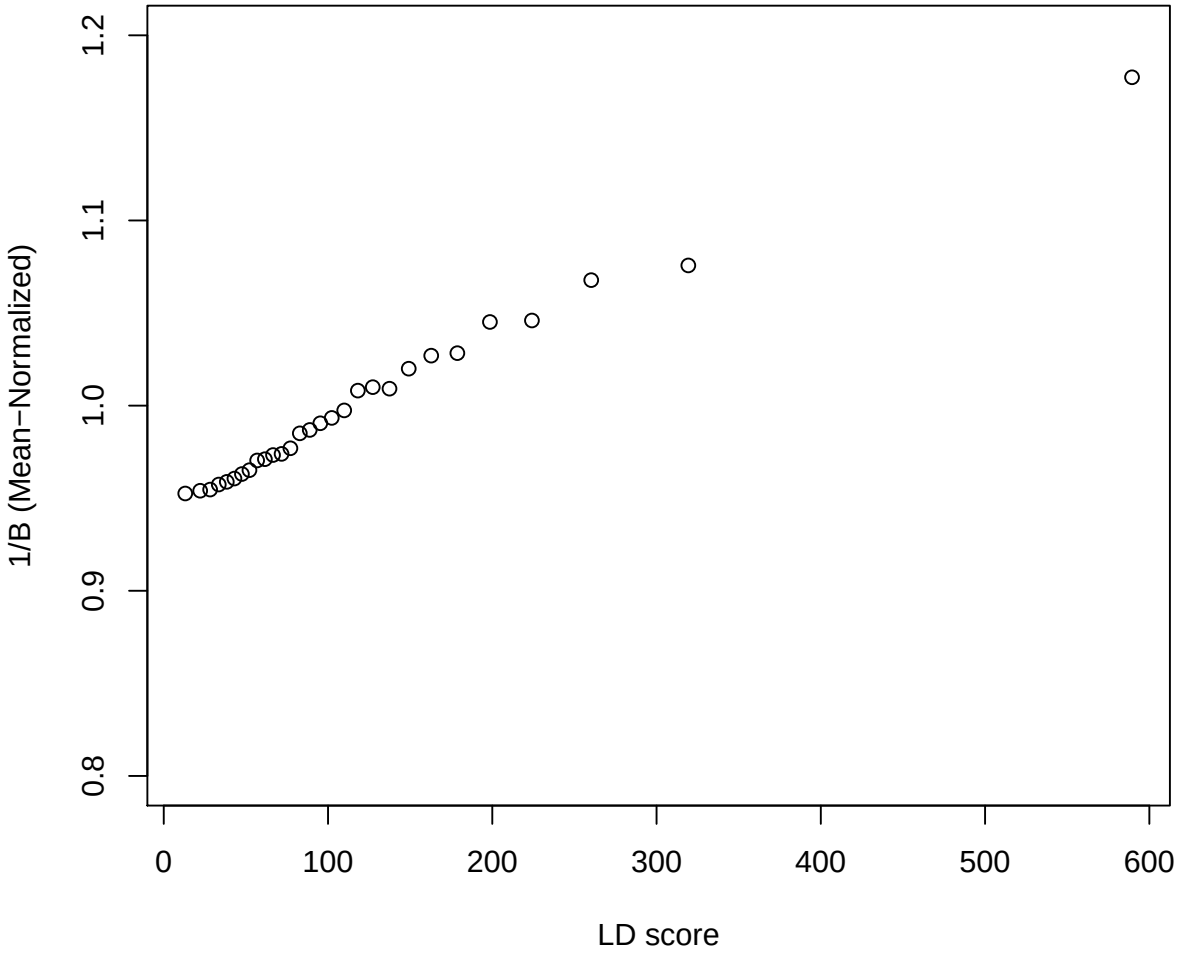

Figure S18: A plot across 30 quantiles of genome-wide LD score our simple BGS model of differentiation, parameterized by McVicker's B (eq 20).

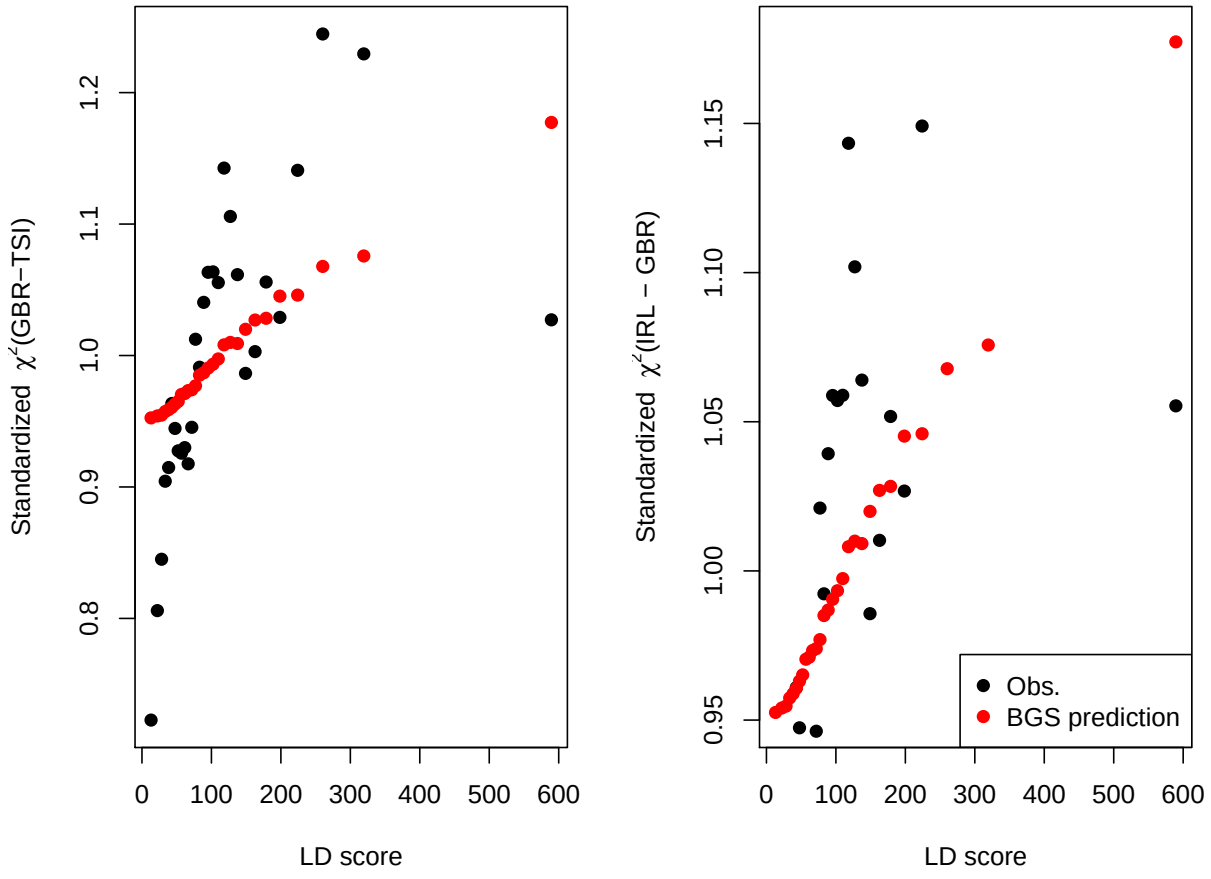

Figure S19: A plot across 30 quantiles of LD score a standardized χ^2 (eq 22) of allele frequency differentiation (black dots) and that expected under our simple BGS model parameterized by McVicker's B (red dots, eq 20, standardized by its genome-wide mean). Note that the red dots are the same values in both panels, and match those given in Figure S18.

where the bars denote genome-wide averages. In other words, the intercept is suppressed by $aN\beta_{F_{ST};\ell}\overline{\ell}$. Another useful way to write the intercept is

$$\alpha_{\chi^2;\ell} = aN\overline{F_{ST}}\big(1 - \frac{\beta_{F_{ST};\ell}}{\overline{F_{ST}}}\overline{\ell}\big), \tag{27}$$

as $\frac{\beta_{F_{ST};\ell}}{\overline{F_{ST}}}$ is the slope of the $\frac{F_{ST_i}}{\overline{F_{ST}}} \sim \overline{\ell_i}$ regression—i.e. the effect of LD score on the relative reduction in F_{ST} from its mean.

S6.4 Using LD score regression to assess 'genetic correlations' with allele frequency differentiation.

In the main text we plot the (Height GWAS effect size) × (Allele frequency difference) LD score regression (Figures 4D-F & S15). In a number of cases we see a strong intercept for this regression, and in some cases a significant slope. Here we show how a non-zero intercept may be a signal of stratification in the original GWAS along the axis represented by the allele frequency difference, while a non-zero slope may demonstrate that this stratification has interacted with BGS.

The logic of assessing genetic correlations via LD score [20] is that at each SNP (i) we have a pair $(Z_{i,1}, Z_{i,2})$: scores for phenotypes 1 and 2 and the genetic correlation (ρ_g) between the phenotypes is captured by the slope of the regression $(Z_{i,1} \cdot Z_{i,2}) \sim \ell_i$. Imagine that these Z's were estimated by conducting a GWAS of the two traits in a sample of size N_1 and N_2 respectively, with a sample overlap of N_s individuals. The intercept of this regression, under the assumptions of [20], is determined by the phenotypic correlation (ρ) in the N_S overlapping samples. Bulik-Sullivan et al [20] show that under their assumptions of no stratification and no linked selection,

$$\mathbf{E}[Z_{i,1}Z_{i,2}] = \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_i + \frac{N_s\rho}{\sqrt{N_1N_2}} \tag{28}$$

Yengo et al [41] extended this to the case of a phenotype from a stratified population. Consider as before a population that consists of two equally sized samples from two populations with allele frequency differentiation F_{ST}. The difference in mean phenotype 1 and 2 between the two populations are σ_1 and σ_2 respectively. Yengo et al [41] show

$$\mathbf{E}[Z_{i,1}Z_{i,2}] = \frac{\sqrt{N_1N_2}\rho_g}{M}\ell_i + \frac{N_s\rho}{\sqrt{N_1N_2}} + \rho_g F_{ST}^2\sqrt{N_1N_2} + \frac{N_s^2 F_{ST}\sigma_1\sigma_2}{\sqrt{N_1N_2}} \tag{29}$$

this is equation (17) of [41], up to slight differences in notation.

Lets return to our case of the LDscore regression of (Height GWAS effect size) × (Allele frequency difference). Assume for the moment that our 'Allele frequency difference' (e.g. GBR-TSI) measures the difference in allele frequency between the two populations stratifying our GWAS. In our case, let phenotype 1 be a phenotype (e.g. height) and let 2 be an individual's population membership (e.g. 1 if in population 1 and 0 if in population 2) $Z_{i,H}$ and the $Z_{i,P}$ score-proxy of the allele frequency difference. The two phenotypes are measured in the same cohort (such that $N_1 = N_2 = N_S$. The difference in mean phenotype (height) between the two populations is σ_1. The mean difference in population membership is 1. As we can assume that population membership is not a genetic trait it follows that $\rho_g = 0$. However, there is a 'phenotypic' correlation between population membership and height, as height differs between our two populations stratifying our GWAS ($\rho = \sigma_1 \times 1$). Following the logic of eqn (29) then

$$\mathbf{E}[Z_{i,H}Z_{i,P}] \approx A\sigma_1 + CF_{ST}\sigma_1 \tag{30}$$

where A and C are constants. Note the strong similarity of eq 30 to the univariate LD score regression for allele frequency χ^2 (eq 21). In reality the population samples (GBR and TSI) used to assess European N-S allele frequencies differences, in Figure 4D-F, and related figures, are not the population samples used in the GWAS. However, the spirit of eq 30 holds if the confounding in a GWAS falls along this N-S axis. A

significant intercept of this regression potentially indicates that some portion of the phenotypic variance (e.g. height) in the GWAS samples was confounded by residual N-S population structure and this problem has been transmitted through into the GWAS effect sizes. This LD score regression is not necessarily expected to have any slope as eq 30 does not include the LD score (ℓ_i). However, if the population structure confounding (F_{ST}) in the GWAS samples is correlated with LD score (ℓ_i), e.g. due to BGS, then a slope will be induced (in a manner similar to to eq 25).