Quantification of genetic components of population differentiation in UK Biobank traits reveals signals of polygenic selection

Xuanyao Liu[1,2*], Po-Ru Loh[3,4], Luke J. O'Connor[1], Steven Gazal[1,4], Armin Schoech[1], Robert M. Maier[4], Nick Patterson[4], Alkes L. Price[1,4,5*]

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
2. Department of Human Genetics, The University of Chicago, Chicago, Illinois, USA
3. Division of Genetics, Harvard Medical School, Boston, Massachusetts, USA
4. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
5. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.

*Correspondence should be addressed to X.L. (xuanyao@uchicago.edu) or A.L.P. (aprice@hsph.harvard.edu).

Abstract

The genetic architecture of most human complex traits is highly polygenic, motivating efforts to detect polygenic selection involving a large number of loci. In contrast to previous work relying on top GWAS loci, we developed a method that uses genome-wide association statistics and linkage disequilibrium patterns to estimate the genome-wide genetic component of population differentiation of a complex trait along a continuous gradient, enabling powerful inference of polygenic selection. We analyzed 43 UK Biobank traits and focused on PC1 and North-South and East-West birth coordinates across 337K unrelated British-ancestry samples, for which our method produced close to unbiased estimates of genetic components of population differentiation and high power to detect polygenic selection in simulations across different trait architectures. For PC1, we identified signals of polygenic selection for height (74.5±16.7% of 9.3% total correlation with PC1 attributable to genome-wide genetic effects; $P = 8.4\times10^{-6}$) and red hair pigmentation (95.9±24.7% of total correlation with PC1 attributable to genome-wide genetic effects; $P = 1.1\times10^{-4}$); the bulk of the signal remained when removing genome-wide significant loci, even though red hair pigmentation includes loci of large effect. We also detected polygenic selection for height, systolic blood pressure, BMI and basal metabolic rate along North-South birth coordinate, and height and systolic blood pressure along East-West birth coordinate. Our method detects polygenic selection in modern human populations with very subtle population structure and elucidates the relative contributions of genetic and non-genetic components of trait population differences.

Introduction

The genetic architecture of human complex traits is highly polygenic[1-7]. Natural selection on polygenic traits could occur in polygenic fashion, via small shifts in allele frequencies across a large number of loci[8,9]. Signals of polygenic selection have been detected for several complex traits, including height and body mass index (BMI), by correlating SNP effect sizes from genome-wide association studies (GWAS) with population differences in allele frequencies or by computing singleton density scores[10-17]. However, methods for detecting polygenic selection generally restrict their analyses to genome-wide significant associated SNPs or relatively small sets of top associated SNPs, which generally capture only a small proportion of trait heritability[18]. This may limit power to detect polygenic selection for highly polygenic traits and precludes estimation of the genetic component of population differences in phenotype, a fundamental population genetic quantity.

In this study, we developed a method, PopDiff, that quantifies the genetic component of population differences in phenotype, using association statistics from genome-wide SNPs and linkage disequilibrium (LD) between SNPs; the method can be applied to continuous gradients of genetic ancestry, which are often an effective way to model subtle population structure[19-21]. A significantly non-zero value of the genetic component after accounting for effects of genetic drift indicates a signal of polygenic selection. We applied PopDiff to 43 UK Biobank traits (N=337K unrelated British-ancestry samples[22]), analyzing three continuous gradients: the top principal component (PC1) and North-South and East-West birth coordinates. We detected signals of polygenic selection for several traits, including traits not previously reported to be under polygenic selection.

Results

Overview of methods

Population differences in phenotype may have genetic and/or environmental components. Our method, PopDiff, quantifies the genetic component of population differences in phenotype to detect signals of polygenic selection. We first consider the special case of two discrete subpopulations and no LD. In this case, given a phenotype whose mean differs between the two subpopulations, an unbiased estimate of the genetic component of the phenotypic difference can be obtained by summing the estimated contribution of each SNP to the phenotypic difference, based on the product of the difference in allele frequency and the estimated effect size. (Throughout this paper, "genetic component" is defined as the component explained by a specified set of SNPs and may exclude other genetic effects.) Although genetic differences between subpopulations can arise due to either genetic drift or polygenic selection, the effects of genetic drift vary stochastically across the genome, such that standard errors computed using a block-jackknife include the effects of genetic drift and enable a statistical test for polygenic selection. It is straightforward to generalize from two discrete subpopulations to a pre-specified continuous gradient of genetic ancestry (e.g. based on principal components (PCs) or birth coordinates) by replacing the difference between subpopulations with the correlation to the continuous gradient. In the presence of LD, which can cause effects of linked SNPs to be double-counted, we multiply estimated SNP effects by the inverse of the LD matrix to correct for LD[23]. To reduce noise, we regularize the LD matrix as described previously[24]. (This regularization can introduce a conservative bias in estimates but increases power to detect polygenic selection; we thoroughly investigate this bias-variance tradeoff in our simulations). We note that analyses of polygenic selection can potentially be confounded by uncorrected population stratification[28,29]. We correct for 10 PCs when estimating marginal SNP effects, but careful consideration of possible uncorrected population stratification is warranted (see Population stratification section).

In detail, we estimate the genetic component ΔG of population differentiation ΔY (defined as the correlation between continuous ancestry gradient and phenotype Y) via

$$\widehat{\Delta G} = \frac{\eta^T X}{N}(\hat{D} + \gamma I)^{-1}\hat{\alpha}, \quad (1)$$

where η is a pre-specified N x 1 vector quantifying a continuous gradient of genetic ancestry across samples (normalized to mean 0 and variance 1), X is the N x M matrix of normalized genotypes, $\hat{D}$ is an M x M banded LD matrix computed using all samples, γ is a scalar regularization parameter[24], I is the M x M identity matrix, $\hat{\alpha}$ is the M x 1 vector of normalized estimated marginal (inclusive of LD) effect sizes for each SNP, M is the number of SNPs, and N is the number of samples. We note that $(\hat{D} + \gamma I)^{-1}\hat{\alpha}$ is an estimate of normalized *causal* SNP effect sizes β, so that $\widehat{\Delta G}$ is an estimate of $\Delta G = \frac{\eta^T X}{N}\beta$. The genetic proportion of population differentiation, denoted $\%G$, is estimated as $\widehat{\Delta G}/\Delta Y$. As noted above, we estimate the standard error of $\widehat{\Delta G}$ using a block-jackknife that includes the effects of genetic drift, so that a significantly nonzero $\widehat{\Delta G}$ indicates a signal of polygenic selection. Details of our PopDiff method are described in the Methods section; we have released open-source software implementing the method (see URLs).

In this study, we analyzed data from 337,536 unrelated British-ancestry samples from the full UK Biobank release[22] (see URLs). We considered 43 highly heritable traits and three continuous ancestry gradients η: PC1 of the 337,536 unrelated British-ancestry samples, North-South birth coordinates and East-West birth coordinates. (We did not consider lower PCs, which correspond to exceedingly subtle genetic effects: F_{ST} <0.0001; Table S2 of ref. 21). We applied our PopDiff method to 67 (η, trait) pairs for which the population differentiation ΔY (correlation between η and phenotype Y) had absolute value greater than 0.01 (Table S1). In all analyses, we computed the LD matrix $\hat{D}$ using the full set of 337,536 samples. In analyses of North-South and East-West birth coordinates, we computed marginal effect size estimates $\hat{\alpha}$ using the full set of 337,536 samples, correcting for 10 PCs. In analyses of PC1, this choice of $\hat{\alpha}$ would be mathematically guaranteed to produce a $\widehat{\Delta G}$ estimate of 0 (since correlations to PC1 are subtracted out, see Methods), and for this reason we instead computed $\hat{\alpha}$ using 10 random (non-

overlapping) subsets of 33,754 samples (correcting for 10 PCs within each subset) and meta-analyzed the results.

Simulations

We performed simulations using real UK Biobank genotypes to assess the bias, type I error and power of our PopDiff method. We performed simulations using both PC1 and North-South birth coordinates. We used unrelated British-ancestry samples (N=337,536) and genome-wide genotyped SNPs (M=516,086 SNPs after QC; see Methods); our simulations used exactly the same sample set and SNP set as our analyses of real traits. Phenotypes were simulated using an additive model with SNP-heritability set to 0.2, similar to most UK Biobank traits[25,26] (see Methods). The proportion of causal SNPs was set to 100% (the default value), 10% or 1%, and the regularization parameter γ was set to 0.1 (as in ref. 24), 0.2, 0.5 or 1.0. In PC1 simulations, we estimated $\hat{\alpha}$ using a random subset of unrelated British-ancestry samples (N=33,754, correcting for 10 PCs within the subset), analogous to our analyses of real traits.

We performed null simulations (heritable phenotype with $\Delta Y = 0.11$, $\Delta G = 0$, $\%G = 0$) to assess bias and type I error. We first performed simulations using PC1. Estimates of ΔG and $\%G = \Delta G/\Delta Y$ were unbiased at all values of γ, although estimates were extremely noisy at γ=0.1 (Figure 1A and Table S2A). Type I error was properly controlled at all values of γ (Figure 1B and Table S2B; conservative at γ=0.1 due to noisy estimates), and jackknife standard errors (s.e.) were similar to empirical standard deviations (s.d.) of ΔG estimates (Table S2A). We also performed simulations using North-South birth coordinates. Once again, estimates of ΔG and $\%G = \Delta G/\Delta Y$ were unbiased at all values of γ, although estimates were extremely noisy at γ=0.1 (Figure S1A and Table S3A). Type I error was properly controlled for γ=0.1, 0.2, 0.5 but not for γ=1.0 (Figure S1B and Table S3B; conservative at γ=0.1 due to noisy estimates), and jackknife s.e. were similar to empirical s.d. for all values of γ (Table S3A); we chose γ=0.5 as the default setting (see below). Notably, we obtained similar results at different values of the proportion of

causal SNPs (100%, 10% or 1%), both for simulations using PC1 (Table S4) and for simulations using North-South birth coordinates (Table S5).

We performed causal simulations (ΔY = 0.11, ΔG = 0.086, $\%G$ = 78%) to assess bias and power. We first performed simulations using PC1. Estimates of ΔG and $\%G = \Delta G/\Delta Y$ were unbiased but extremely noisy at γ=0.1, slightly upward biased at γ=0.2, and slightly conservative at γ=0.5 and γ=1.0 (Figure 1C and Table S6A). Power increased as a function of γ, with high power at γ=0.5 and very high power at γ=1.0 (Figure 1D and Table S6B). We also performed simulations using North-South birth coordinates. Estimates of ΔG and $\%G = \Delta G/\Delta Y$ were upward biased and extremely noisy at γ=0.1, slightly upward biased at γ=0.2, close to unbiased at γ=0.5, and conservative at γ=1.0 (Figure S1C and Table S7A). Power increased as a function of γ, with very high power at γ=0.5 and γ=1.0 (Figure S1D and Table S7B). As in null simulations, we obtained similar results at different values of the proportion of causal SNPs (100%, 10% or 1%), both for simulations using PC1 (Table S8) and for simulations using North-South birth coordinates (Table S9). Based on the results of both null and causal simulations, we chose γ=0.5 as the default regularization parameter value in all of our analyses of real traits, as this parameter value consistently controls false positives, produces close to unbiased or slightly conservative estimates, and achieves high power. (For completeness, we also report results of secondary analyses at γ=1 in our analyses of real traits.)

Polygenic selection along PC1 in UK Biobank

We considered 43 UK Biobank traits, restricting to 337,536 unrelated British-ancestry samples (average N= 321,389 phenotyped samples; Table S10). We defined a continuous ancestry gradient η based on PC1 of the full set of 337,536 samples, representing a north-south axis separating southern England from Northern Ireland[21]. We applied our PopDiff method to 22 (PC1, trait) pairs for which the population differentiation ΔY (correlation between PC1 and phenotype Y) had absolute value greater than 0.01 (Table S1). All analyses were corrected for 67 hypotheses tested, which include other choices of η (see below).

Results are displayed in Figure 2 and Table S11. We identified two traits with statistically significant %G for PC1 ($p<0.05/67=7.5\times10^{-4}$), implicating polygenic selection: height and red hair pigmentation. For height ($\Delta Y=0.093$; individuals with ancestry from southern England are taller on average than individuals with ancestry from Northern Ireland), our estimate of %G was 74.5% (s.e.=16.7%; $p=8.4\times10^{-6}$), implying that differences in height along PC1 are primarily due to selection and cannot be explained by genetic drift. We note that height has previously been reported to be under polygenic selection[10-16]. For red hair pigmentation ($\Delta Y=-0.039$; red hair is more common in individuals with ancestry from Northern Ireland than in individuals with ancestry from southern England), our estimate of %G was 95.9% (s.e.=24.7%; $p=1.1\times10^{-4}$), implying that differences in red hair pigmentation along PC1 are primarily due to selection and cannot be explained by genetic drift. We note that the genetic architecture of red hair pigmentation includes large-effect loci, with 12 genome-wide significant loci explaining 7.2% of trait variance (Table S12, comparable to 6.9% in ref. 27). We repeated our analysis after removing these 12 loci and surrounding regions (±1Mb), and confirmed that the signal of polygenic selection remained (%G=73.2%, s.e.=21.3%; $p=6.1\times10^{-4}$). This demonstrates that polygenic selection can affect traits whose genetic architectures include large-effect loci. We are not currently aware of previous evidence of polygenic selection on red hair pigmentation, although a previous study reported polygenic selection on skin pigmentation[11]. In secondary analyses at $\gamma=1$, %G estimates for height and red hair pigmentation were lower (consistent with Figure 1C), but remained statistically significant (Table S13A).

Population stratification

Recent work has suggested that previous studies of polygenic selection may be confounded by uncorrected population stratification, compromising their results[28,29]. Correcting for population stratification is clearly very important in analyses of polygenic selection, as we determined that repeating our analyses with no correction for population stratification produced unstable results (Table S14; e.g. height %G > 2000%, $p<10^{-200}$). We note that uncorrected population stratification may be either environmentally driven (driven by environmental components of population differences in phenotype) or

genetically driven (driven by genetic components of population differences in phenotype, caused by polygenic selection). We are primarily concerned about the former case, as the latter case represents true-positive (not false-positive) signals of polygenic selection—although estimates of %G could still be inflated in the latter case.

We performed a series of analyses to assess whether our results are robust to uncorrected population stratification. We first considered simulations. Our null PC1 simulations described above (heritable phenotype with $\Delta Y = 0.11$, $\Delta G = 0$, $\%G = 0$; Figure 1A,B), which correspond to the case of environmentally driven population stratification, achieved correct calibration. However, it is also of interest to check whether environmentally driven population stratification along other genetic gradients could lead to false-positive signals of polygenic selection along PC1. To assess this, we performed additional simulations in which we simulated environmentally driven population stratification along North-South birth coordinate (analogous to Figure S1A,B) but evaluated evidence of polygenic selection along PC1. Results are reported in Figure S2 and Table S15. We confirmed that type I error was properly controlled.

We next considered analyses of UK Biobank traits. We repeated our PC1 analyses by estimating $\hat{\alpha}$ using 10 (non-overlapping) subsets of 33,754 samples ordered by PC1 values (similar values of PC1 within each subset, so as to minimize stratification), correcting for PCs within each subset and meta-analyzing the results. We confirmed that %G remained statistically significant for both height (%G=123.0%, s.e.=14.7%; p=4.9×10^{-17}) and red hair pigmentation (%G=93.54%, s.e.=35.26%; p=7.9×10^{-3}); results for all 22 traits are reported in Table S16. We also repeated our PC1 analysis of height using family-based effect size estimates from ref. 28, and determined that the %G estimate along British PC1 was 95.5% (s.e.=40.8%, p=0.019); in secondary analyses at γ=1, the %G estimate remained positive, but was smaller and non-significant (Table S13B). We note that our results involving British PC1 are orthogonal to previous findings involving European PC1[28]: we repeated our analysis using a European PC1 computed using all N=460K European-ancestry samples[22,26], and determined that the correlation

between European PC1 and British PC1 loadings was only −0.017, and that the %G for height along European PC1 (relative to ΔY = −2.5%) was −20.7% (s.e.=10.2%, p=0.043), which is consistent with the ref. 28 finding that the genetic component of the population difference in height along European PC1 has the opposite sign of the total population difference. Overall, these secondary analyses support our findings (as well as those of ref. 28).

Polygenic selection along North-South birth coordinate in UK Biobank

We next defined a continuous ancestry gradient η based on North-South birth coordinate. We applied our PopDiff method to 24 (North-South birth coordinate, trait) pairs for which the population differentiation ΔY (correlation between North-South birth coordinate and phenotype Y) had absolute value greater than 0.01 (Table S1). All analyses were corrected for 67 hypotheses tested.

Results are displayed in Figure 3 and Table S17. We identified four traits with statistically significant %G for North-South birth coordinate (p<0.05/67=7.5×10^{-4}): height, BMI, basal metabolic rate and systolic blood pressure. For height (ΔY=-0.091; individuals born in the southern UK are taller on average than individuals born in the northern UK), our estimate of %G was 124.6% (s.e.=15.9%; p=4.6×10^{-15}), implying that differences in height along North-South birth coordinate are predominantly genetic and cannot be explained by genetic drift. For BMI (ΔY=0.040; individuals born in the northern UK have larger BMI on average than individuals born in the southern UK), our estimate of %G was 128.2% (s.e. = 25.7%; p=5.8×10^{-7}), implying that differences in BMI along North-South birth coordinate are predominantly genetic and cannot be explained by genetic drift. Both height and BMI have previously been reported to be under polygenic selection[10-16]. We note that our estimates of %G are not significantly larger than 100%; however, values of %G larger than 100% are possible if genetic and environmental geographic effects have opposite signs, as previously reported for BMI (ref. 12).

We are not currently aware of previous evidence of polygenic selection on systolic blood pressure and basal metabolic rate. For systolic blood pressure (ΔY=0.032; individuals

born in the northern UK have higher systolic blood pressure on average than individuals born in the southern UK), our estimate of %G was 83.3% (s.e.=19.1%; p=1.3×10^{-5}), implying that differences in systolic blood pressure along North-South birth coordinate are primarily due to selection and cannot be explained by genetic drift. For basal metabolic rate (ΔY=-0.032; individuals born in the southern UK have higher basal metabolic rate on average than individuals born in the northern UK), our estimate of %G was 125.6% (s.e.=20.0%; p=3.3×10^{-10}), implying that differences in basal metabolic rate along North-South birth coordinate are primarily due to selection and cannot be explained by genetic drift. We note that both systolic blood pressure and basal metabolic rate have significant genetic correlation with height and BMI in UK Biobank data (Table S18; estimated using cross-trait LD score regression[30]). For both systolic blood pressure and basal metabolic rate, %G became only nominally significant ($0.05/67=7.5\times10^{-4}<p<0.05$) when computing association statistics using height as a covariate (Table S19), suggesting that polygenic selection on these traits may be impacted by polygenic selection on height. On the other hand, for both of these traits, %G remained highly significant when computing association statistics using BMI as a covariate (Table S19), although we caution that adjusting association statistics for heritable covariates can introduce collider bias[31]. In secondary analyses at γ=1, %G estimates were lower for most traits, but remained statistically significant (Table S20).

Polygenic selection along East-West birth coordinate in UK Biobank

Finally, we defined a continuous ancestry gradient based on East-West birth coordinate. We applied our PopDiff method to 21 (East-West birth coordinate, trait) pairs for which the population differentiation ΔY (correlation between East-West birth coordinate and phenotype Y) had absolute value greater than 0.01 (Table S1). All analyses were corrected for 67 hypotheses tested.

Results are displayed in Figure 4 and Table S21. We identified two traits with statistically significant %G for East-West birth coordinate ($p<0.05/67=7.5\times10^{-4}$): height and systolic blood pressure, both of which were also statistically significant in our analysis of North-South birth coordinate. For height (ΔY=0.087; individuals born in the eastern UK are

taller on average than individuals born in the western UK), our estimate of %G was 78.0% (s.e.=21.2%; p=2.4×10^{-4}), implying that differences in height along East-West birth coordinate are primarily due to selection and cannot be be explained by genetic drift. For systolic blood pressure (ΔY=-0.031; individuals born in the western UK have higher systolic blood pressure on average than individuals born in the eastern UK), our estimate of %G was 85.2% (s.e.=16.4%; p=2.4×10^{-7}), implying that differences in systolic blood pressure along East-West birth coordinate are primarily due to selection and cannot be explained by genetic drift. The %G for systolic blood pressure remained highly significant when computing association statistics using height as a covariate (Table S22), although we caution that adjusting association statistics for heritable covariates can introduce collider bias[31]. In secondary analyses at γ=1, %G estimates were lower, but remained statistically significant (Table S23).

Discussion

We developed a method, PopDiff, that quantifies the genetic component of population differences in phenotype to detect signals of polygenic selection. The method was well-powered in simulations and analyses of real UK Biobank traits in detecting polygenic selection within British-ancestry samples, which have very subtle population structure. We identified several traits under polygenic selection, including traits previously reported to be under polygenic selection (height and BMI) and other traits (red hair pigmentation, systolic blood pressure and basal metabolic rate).

PopDiff is the first method that we are aware of that produces approximately unbiased estimates of the genetic component of population differences in phenotype (%*G*). In particular, estimating %*G* using only genome-wide significant associated SNPs is expected to produce downward biased estimates. Indeed, when we estimated %*G* for height along PC1 in UK Biobank data using a set of 1,131 genome-wide significant SNPs ($P < 5 \times 10^{-8}$; LD-pruned to $r^2<0.01$[26]), we obtained an estimate of 27.6% (s.e.=0.1% using s.e. of effect size estimates of each SNP, which does not account for effects of genetic drift; s.e.=7.2% using block-jackknife, which account for drift but may not be valid for small sets of SNPs), which is much lower than the estimate produced by PopDiff. When we estimated %*G* for red hair pigmentation along PC1 using a set of 47 genome-wide significant SNPs ($P < 5 \times 10^{-8}$; LD-pruned to $r^2<0.01$), we obtained an estimate of 5.6% (s.e.=0.1% using s.e. of effect size estimates of each SNP; s.e.=8.2% using block-jackknife). More generally, there exist several methods that shrink estimated effect sizes for the purpose of maximizing polygenic prediction accuracy[3,32-35], but these approaches are also expected to produce downward biased estimates of %*G*.

Our work has several limitations. First, all methods for detecting polygenic section may produce false-positive signals if association statistics are confounded by uncorrected population stratification[28,29], thus careful consideration of possible stratification is required. Our secondary analyses involving homogenous subsets of samples suggest that our results are robust to population stratification (Table S14). Second, the approach for LD matrix regularization[24] employed by PopDiff introduces a bias-variance tradeoff in

estimates of %*G* (Figure 1). Estimates of %*G* may not be perfectly unbiased, although they are close to unbiased in our simulations across a broad set of genetic architectures. However, the regularization parameter that optimizes this bias-variance tradeoff may vary across different data sets (e.g. depending on the sample size and SNP set), such that analyses of new data sets may require revisiting the choice of regularization parameter; investigating other LD matrix regularization approaches may also prove useful[36,37]. Third, we restricted our analyses to genotyped SNPs (due to complexities of LD matrix regularization and to computational cost); analyses of the %*G* explained by a larger set of genotyped and imputed SNPs might yield slightly larger estimates, consistent with the slightly larger heritability that they explain[38]. Fourth, we focused on British-ancestry samples in UK Biobank, which have very subtle structure. We did not apply PopDiff to estimate continental-level population differences in phenotype, because the much larger amount of drift between continental populations (and possible effects of differential LD[39]) will lead to large jackknife s.e., limiting power to detect polygenic section. Fifth, when polygenic selection is detected, we are unable to infer when the selection occurred, as the population structure of British-ancestry samples may often reflect differing proportions of ancestry from more deeply diverged source populations in which selection might have occurred[13,20]. Despite these limitations, PopDiff is a powerful method for quantifying the genetic component of population differences in phenotype to detect signals of polygenic selection.

URLs:

Software implementing the PopDiff method will be released prior to publication as a publicly available, open-source software package at https://www.hsph.harvard.edu/alkes-price/software; UK Biobank www site, http://www.ukbiobank.ac.uk/; LDSC software, https://github.com/bulik/ldsc/; PLINK2.0, https://www.cog-genomics.org/plink/2.0/; EIGENSOFT, https://www.hsph.harvard.edu/alkes-price/software.

Acknowledgements:

We are grateful to A. Dahl, M. Sohail, R. Maier, D. Reich and S. Sunyaev for helpful discussions. This research was conducted using the UK Biobank Resource under Application #14292 and was funded by NIH grants R01 HG006399 and R03 ES027902.

Methods

PopDiff method

The PopDiff method estimates the genetic component ΔG of population differentiation ΔY (defined as the correlation between continuous ancestry gradient and phenotype Y), where "genetic component" refers to the component explained by a specified set of SNPs. A genetic component that is significantly different from 0 after accounting for effects of genetic drift is indicative of polygenic selection.

We assume a simple linear model,

$$Y = X\beta + \epsilon, \tag{2}$$

where Y is standardized phenotype of N samples, X is the standardized genotype (N individuals × M SNPs), β is the vector of causal effect sizes, and ϵ is noise.

Let η denote a standardized continuous gradient of genetic ancestry, *e.g.* birth coordinates or values of a top PC (estimated in a finite sample). The population differentiation ΔY and its genetic component ΔG are defined as follows:

$$\Delta Y = E\left(\frac{\eta^T Y}{N}\right) \text{and}$$
$$\Delta G = E\left(\frac{\eta^T X\beta}{N}\right). \tag{3}$$

The genetic proportion of population differentiation ($\%G$) is estimated as $\frac{\widehat{\Delta G}}{\Delta Y}$.

Letting $\lambda^T = E(\frac{\eta^T X}{N})$ denote SNP loadings along the genetic ancestry gradient η, it follows that

$$\Delta G = \lambda^T \beta, \tag{4}$$

so that ΔG can be estimated using estimated causal effect sizes and SNP loadings.

Given marginal effect size estimates $\hat{\alpha} = \frac{X^T Y}{N}$ and an LD matrix estimated as $\hat{D} = \frac{X^T X}{N}$, an unbiased estimate of causal effect sizes can be computed as:

$$\hat{\beta} = \hat{D}^{-1}\hat{\alpha}. \quad (5)$$

To improve computational efficiency, we divide the genome into non-overlapping blocks of 10,000 SNPs. To reduce noise in LD estimates for SNPs that are far apart, we band the LD matrices to bands of 200 SNPs, with LD estimates outside the bands set to zero. We compute local banded LD matrices $\hat{D}_l$ for each block. To further reduce estimation noise, we regularize LD estimates as previously described[24] (using a regularization parameter γ), such that regularized causal effect size estimates for each block are computed as:

$$\tilde{\beta}_l = (\hat{D}_l + \gamma I)^{-1}\hat{\alpha}_l. \quad (6)$$

Regularized causal effect size estimates $\tilde{\beta}$ are computed genome-wide and the genetic component is estimated as

$$\widehat{\Delta G} = \lambda\tilde{\beta}. \quad (7)$$

Standard errors of $\widehat{\Delta G}$ are estimated via block-ackknife, partitioning the genome into 200 blocks of non-overlapping SNPs.

We note that if the ancestry gradient η is a genetic PC (e.g. PC1), and summary association statistics $\hat{\alpha}$ are computed by including genetic PCs as covariates, then this choice of $\hat{\alpha}$ would be mathematically guaranteed to produce a $\widehat{\Delta G}$ estimate of 0 along η. In detail, the singular value decomposition of $X = U\Sigma V^T$. When η is PC1, $\eta^T X = \eta^T U\Sigma V^T = \theta_1 v_1^T$, where θ_i is the ith diagonal entry of Σ, and v_i^T is the ith row of V^T. Thus, $(\hat{D} + \gamma I)^{-1} = V(\Sigma^2 + \gamma I)^{-1}V^T$. After correcting for the top 10 PCs,

$\hat{\alpha} = \frac{(u_{11}\theta_{11}v_{11}^T + u_{12}\theta_{12}v_{12}^T + \ldots + u_N\theta_N v_N^T)^T y}{N}$, where u_i is the i[th] column of U and y is phenotype Y corrected for 10 PCs. Following Equation 1, $\widehat{\Delta G} = \frac{\eta^T X}{N}(\hat{D} + \gamma I)^{-1}\hat{\alpha} = \theta_1 v_1^T V(\Sigma^2 + \gamma I)^{-1}V^T\hat{\alpha} = \frac{\theta_1}{N}(\Sigma^2 + \gamma I)^{-1}v_1^T\hat{\alpha} = 0$.

To overcome the problem of $\widehat{\Delta G}$ estimates being mathematically guaranteed to equal 0, when applying PopDiff to genetic PCs we partition the complete set of samples into 10 random non-overlapping subsets, compute summary association statistics for each subset using PCs of each subset as covariates, compute $\widehat{\Delta G}$ estimates using Equation 1, where η is the genetic PC of the complete set of samples and averaged $\widehat{\Delta G}$ estimates across subsets. Our simulations showed that this approach produces close to unbiased $\widehat{\Delta G}$ estimates (Figure 1). In the UK Biobank data that we analyzed, the average correlation between SNP loadings for PC1 of random subsets and PC1 of the entire set of 337,536 unrelated British-ancestry samples was equal to 0.93.

In secondary analyses (see Population stratification section), we instead partitioned the complete set of samples into 10 non-overlapping subsets ordered by PC1 values (similar values of PC1 within each subset, so as to minimize stratification). In the UK Biobank data that we analyzed, the average correlation between SNP loadings for PC1 of these subsets and PC1 of the entire set of 337,536 unrelated British-ancestry samples was equal to 0.027, confirming that stratifying samples by PC1 values largely eliminates stratification along PC1 (although this may not eliminate all population stratification).

UK Biobank data set

The UK Biobank data set contains 805,426 genotyped SNPs and 488,377 samples. We removed SNPs that were multi-allelic, had a genotyping rate less than 99%, had a minor allele frequency (MAF) less than 1%, or were not in Hardy-Weinberg equilibrium ($p < 10^{-6}$). We removed samples of non-British ancestry, samples with a genotyping rate less than 98% were removed, and related samples. After these QC filters, 516,086 SNPs and 337,536 samples remained.

When computing PCs, we LD-pruned the set of SNPs to r^2<0.2 and removed regions of long-range LD and regions with significant or suggestive selection signals, as previously described[21]. We computed PCs using the FastPCA software implemented in EIGENSOFT (see URLs).

Simulations

We performed simulations to evaluate the bias, type I error and power of Popdiff, using real UK Biobank genotypes (M=516,086 SNPs). Phenotypes were simulated using an additive model

$$Y = X\beta + C\eta + \varepsilon, \quad (9)$$

where $X\beta$ represents genetic effects (including genetic effects that are correlated to ancestry) and $C\eta + \varepsilon$ (corresponding to ϵ in Equation 2) represents environmental effects (including environmental effects that are correlated to ancestry: $C\eta$). Thus, in these simulations, population differentiation ΔY may be due to genetic and/or environmental effects. We specified η using either PC1 (computed using 337,536 British samples) or North-South birth coordinates. We simulated phenotypes for either 33,754 randomly selected samples (PC1 simulations) or all 337,536 British samples (North-South birth coordinate simulations), consistent with our analyses of real traits. Causal effect sizes β were specified with the proportion of causal SNPs (p) set to 1%, 10%, or 100%, via a point-normal distribution: $\beta_i \sim N\left(c_1\lambda_i, \frac{h_g^2}{Mp}\right)$ with probability p and 0 otherwise, where λ_i is the SNP loading of SNP i along η, h_g^2 was set to 0.2, and M is the number of SNPs. In null simulations, c_1 was set to 0, so that population differentiation ΔY was entirely non-genetic. In causal simulations, c_1 was set to values such that $c_1 p = 0.006$. Values of ΔY, ΔG and $\%G$ for each simulation are provided in the Results section.Marginal effect sizes $\hat{\alpha}$ were estimated by linear regression using the top three in-sample PCs as covariates. LD matrices $\hat{D}$ were computed using the complete set of UK Biobank British samples (N=337,536, M=516,086). $\widehat{\Delta G}$ estimates were computed using Equation 1. For each type

of simulation, we performed 1,000 simulations using PC1 (33,754 samples simulated) and 100 simulations using North-South birth coordinates (337,536 samples simulated).

Analyses of UK Biobank traits

We considered 43 UK Biobank traits, restricting to 337,536 unrelated British-ancestry samples (average *N*= 321,389 phenotyped samples; Table S10). We estimated genetic components of population differentiation along three continuous gradients of genetic ancestry: PC1, North-South birth coordinate, and East-West birth coordinate (correlations between these ancestry gradients are reported in Table S24). In all analyses, banded LD matrices were computed using the complete set of British samples (N=337,536 and M=516,086). In the PC1 analysis, we divided the British samples into 10 random non-overlapping subsets (6 subsets with *N*=33,754 and 4 subsets with *N*=33,753). We computed summary association statistics for each of the 10 random subsets using PLINK 2.0 (see URLs) and included the top 10 PCs of each subset, age, sex, genotyping array and assessment center as covariates. In the analysis of North-South and East-West birth coordinates, we computed summary association statistics of the complete set of British samples (N=337,536) using PLINK 2.0 and included the top 10 PCs, age, sex, genotyping array and assessment center as covariates. $\widehat{\Delta G}$ estimates were computed using Equation 1.

Figure legends

Figure 1. Null and causal PC1 simulations to evaluate bias, type I error and power. (A) No bias in null simulations. We report the bias in %G estimates for different values of γ. (B) Type I error. We report type I error at $p<0.05$ and type I error at $p<0.005$ for different values of γ. (C) Bias in causal simulations. We report the bias in %G estimates for different values of γ. (D) Power. We report power at $p<0.05$ and power at $p<0.005$ for different values of γ. Error bars represent 95% confidence intervals. Numerical results are reported in Table S2 and Table S6.

Figure 2. Estimates of genetic components of population differentiation (%*G*) and inference of polygenic selection along PC1 in UK Biobank. We report point estimates and standard errors for %*G* along PC1 for 22 UK Biobank traits for which ΔY (correlation between PC1 and phenotype *Y*) had absolute value greater than 0.01. Traits are ranked by statistical significance of nonzero %*G*. Traits with Bonferroni-significant nonzero %*G* ($p < 0.05/67$), indicative of polygenic section, are denoted via orange bars. Numerical results are reported in Table S11.

Figure 3. Estimates of genetic components of population differentiation (%*G*) and inference of polygenic selection along North-South birth coordinate in UK Biobank. We report point estimates and standard errors for %*G* along North-South birth coordinate for 24 UK Biobank traits for which ΔY (correlation between North-South birth coordinate and phenotype *Y*) had absolute value greater than 0.01. Traits are ranked by statistical significance of nonzero %*G*. Traits with Bonferroni-significant nonzero %*G* ($p < 0.05/67$), indicative of polygenic section, are denoted via orange bars. Numerical results are reported in Table S17.

Figure 4. Estimates of genetic components of population differentiation (%*G*) and inference of polygenic selection along East-West birth coordinate in UK Biobank. We report point estimates and standard errors for %*G* along East-West birth coordinate for 21 UK Biobank traits for which ΔY (correlation between East-West birth coordinate and phenotype *Y*) had absolute value greater than 0.01. Traits are ranked by statistical

significance of nonzero %*G*. Traits with Bonferroni-significant nonzero %*G* ($p < 0.05/67$), indicative of polygenic section, are denoted via orange bars. Numerical results are reported in Table S21.

Figure 1. Null and causal PC1 simulations to evaluate bias, type I error and power.

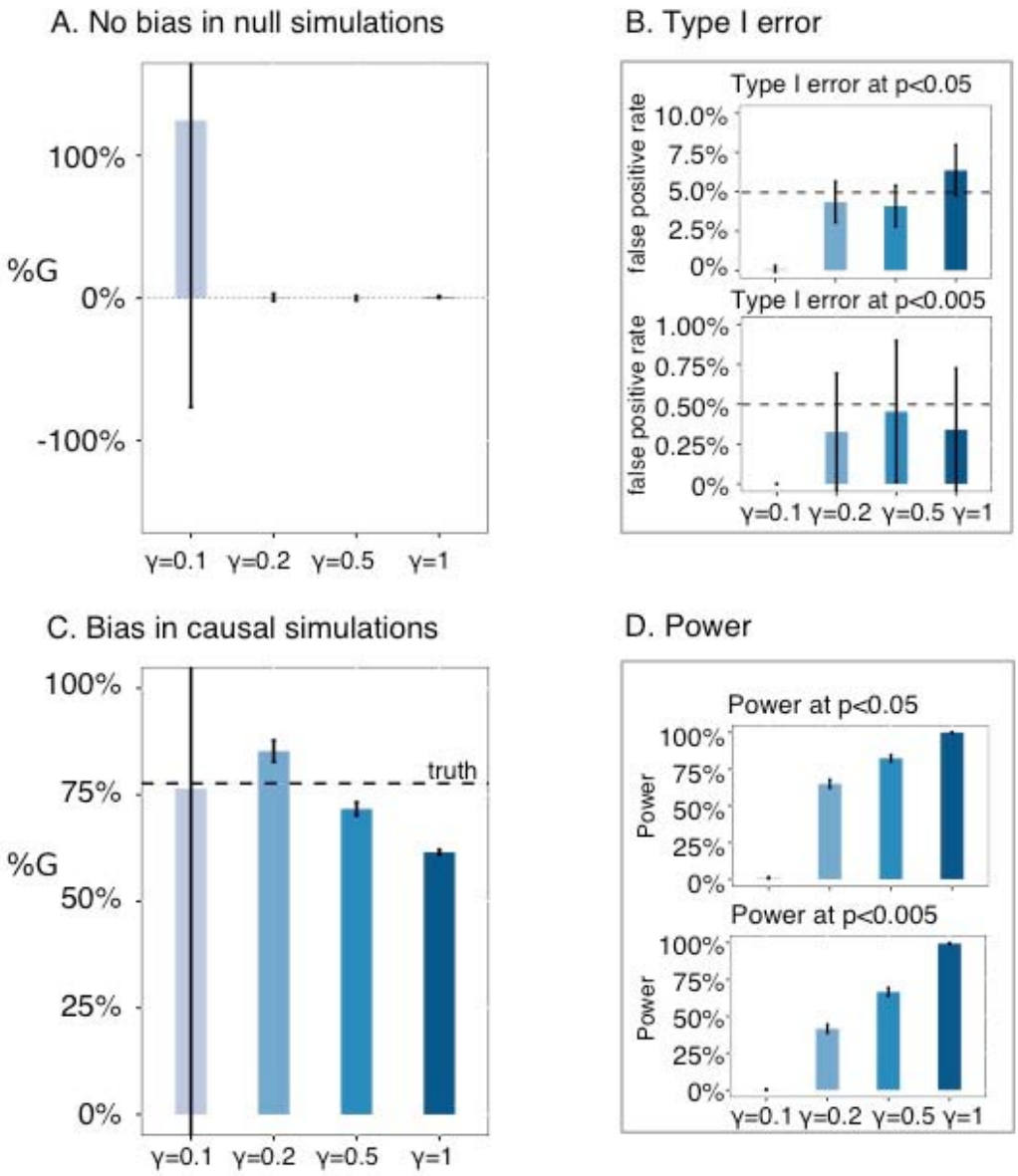


Figure 2. Estimates of genetic components of population differentiation (%*G*) and inference of polygenic selection along PC1 in UK Biobank.

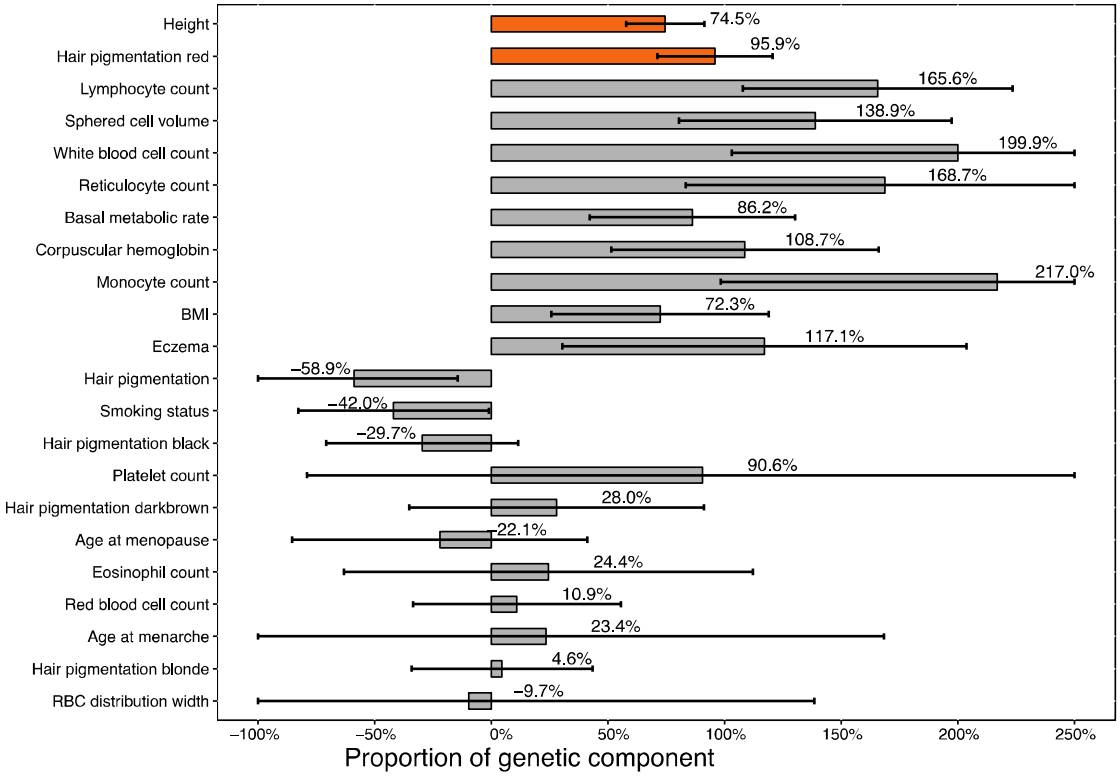


Figure 3. Estimates of genetic components of population differentiation (%*G*) and inference of polygenic selection along North-South birth coordinate in UK Biobank.

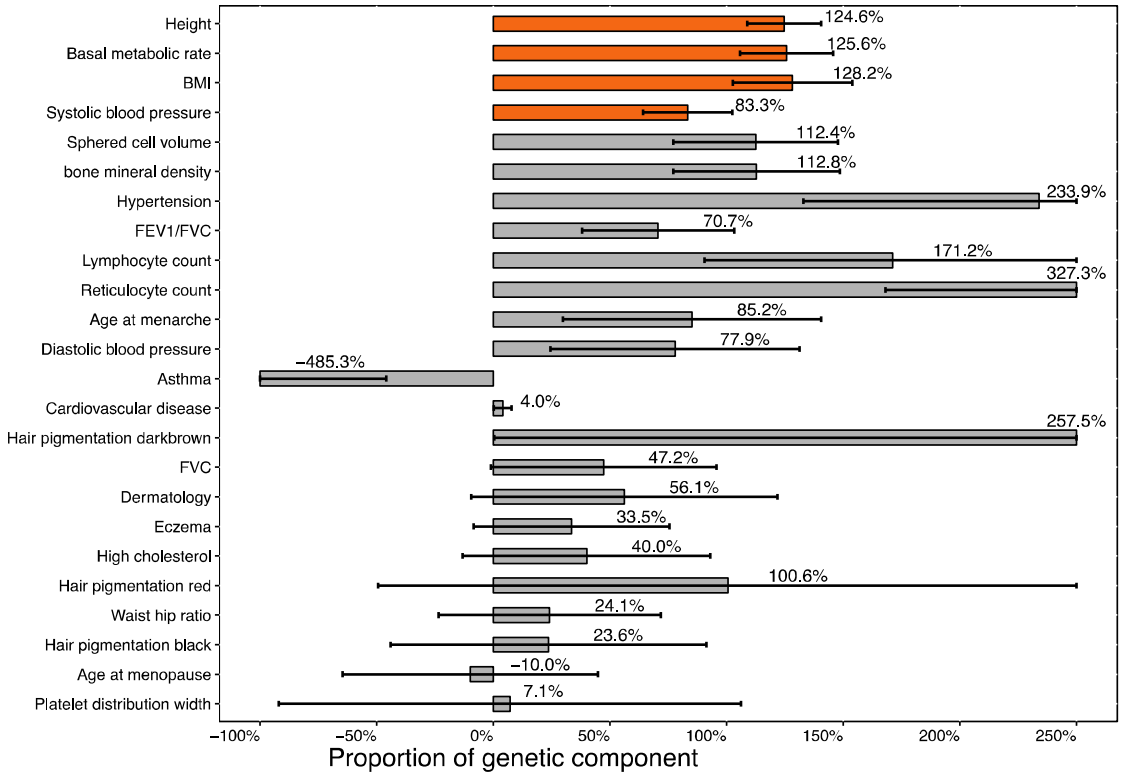
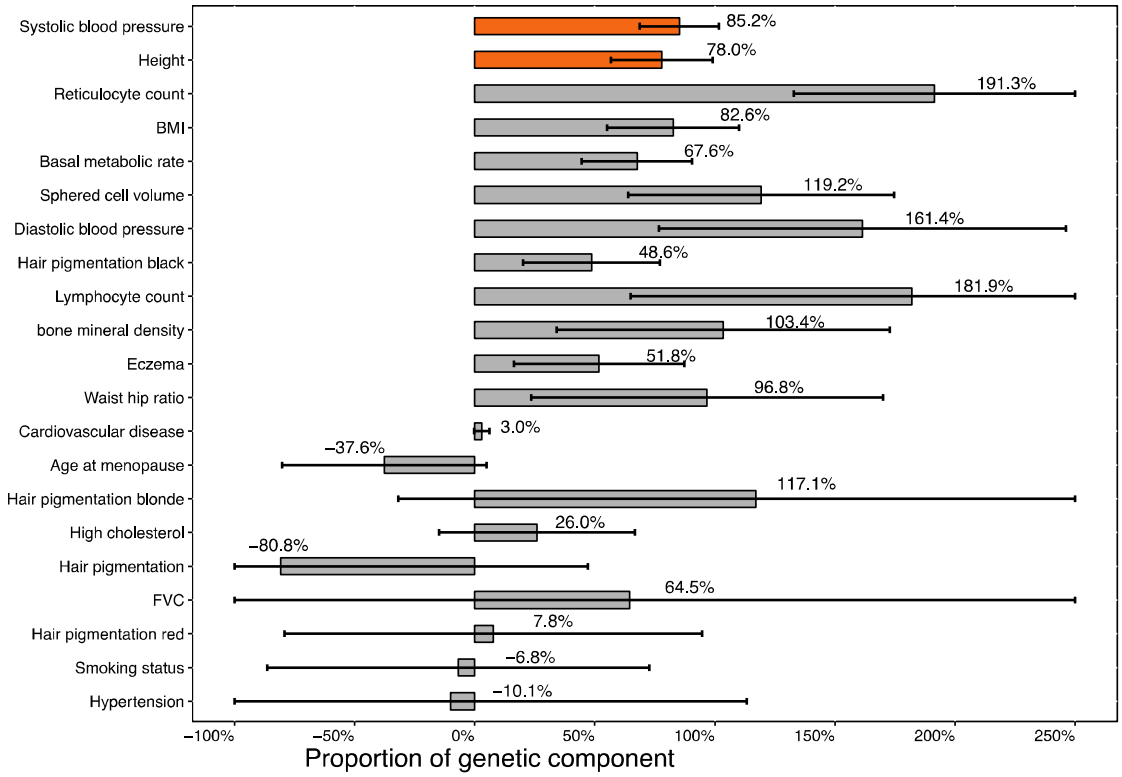


Figure 4. Estimates of genetic components of population differentiation (%*G*) and inference of polygenic selection along East-West birth coordinate in UK Biobank.



References

1. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42,** 565–569 (2010).
2. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43,** 519–525 (2011).
3. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet* **44,** 483–489 (2012).
4. Palla, L. & Dudbridge, F. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. *Am. J. Hum. Genet.* **97,** 250–259 (2015).
5. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47,** 1385–1392 (2015).
6. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).
7. Zeng, J. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* **50,** 746–753 (2018).
8. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20,** R208–15 (2010).
9. Pritchard, J. K. & Di Rienzo, A. Adaptation - not by sweeps alone. *Nat Rev Genet* **11,** 665–667 (2010).
10. Turchin *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet* **44,** 1015–1019 (2012).
11. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet* **10,** e1004412 (2014).
12. Robinson, M. R. *et al.* Population genetic differentiation of height and body mass index across Europe. *Nat Genet* **47,** 1357–+ (2015).
13. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (2015).
14. Zoledziewska, M. *et al.* Height-reducing variants and selection for short stature in Sardinia. *Nat Genet* **47,** 1352–1356 (2015).
15. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354,** 760–764 (2016).
16. Berg, J. J., Zhang, X. & Coop, G. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. *bioRxiv* 167551 (2017). doi:10.1101/167551
17. Guo, J. *et al.* Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat Commun* **9,** 1865 (2018).
18. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101,** 5–22 (2017).
19. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
20. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519,** 309–314 (2015).
21. Galinsky, K. J. *et al.* Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure. *The American Journal of Human Genetics* **99,** 1130–1139 (2016).
22. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298 (2017). doi:10.1101/166298
23. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44,** 369–75– S1–3 (2012).
24. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30,** 2906–2914 (2014).
25. Ge, T. *et al.* Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet* **13,** e1006711 (2017).
26. Loh, P.R. *et al.* Mixed model association for biobank-scale data sets. *Nat Genet* **38**, 203 (2018).
27. Lin, B. D. *et al.* Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. *Genes (Basel)* **6,** 559–576 (2015).

28. Sohail, M.*, Maier, R.* *et al.* Polygenic adaptation signals for height are confounded by population structure. *bioRxiv*. (2018).
29. Berg J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *bioRxiv*. (2018).
30. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47,** 1236–1241 (2015).
31. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for Heritable Covariates Can Bias Effect Estimates in Genome-Wide Association Studies. *The American Journal of Human Genetics* **96,** 329–339 (2015).
32. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
33. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).
34. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47,** 284–290 (2015).
35. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* **17,** 392–406 (2016).
36. Wen, X. & Stephens, M. USING LINEAR PREDICTORS TO IMPUTE ALLELE FREQUENCIES FROM SUMMARY OR POOLED GENOTYPE DATA. *Ann Appl Stat* **4,** 1158–1182 (2010).
37. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* **99,** 139–153 (2016).
38. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47,** 1114–1120 (2015).
39. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100,** 635–649 (2017).