Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies

Aaron J. Stern,[1,*] Leo Speidel,[2] Noah A. Zaitlen,[3] and Rasmus Nielsen[4,5]

Summary

We present a full-likelihood method to infer polygenic adaptation from DNA sequence variation and GWAS summary statistics to quantify recent transient directional selection acting on a complex trait. Through simulations of polygenic trait architecture evolution and GWASs, we show the method substantially improves power over current methods. We examine the robustness of the method under stratification, uncertainty and bias in marginal effects, uncertainty in the causal SNPs, allelic heterogeneity, negative selection, and low GWAS sample size. The method can quantify selection acting on correlated traits, controlling for pleiotropy even among traits with strong genetic correlation ($|r_g| = 80\%$) while retaining high power to attribute selection to the causal trait. When the causal trait is excluded from analysis, selection is attributed to its closest proxy. We discuss limitations of the method, cautioning against strongly causal interpretations of the results, and the possibility of undetectable gene-by-environment (GxE) interactions. We apply the method to 56 human polygenic traits, revealing signals of directional selection on pigmentation, life history, glycated hemoglobin (HbA1c), and other traits. We also conduct joint testing of 137 pairs of genetically correlated traits, revealing widespread correlated response acting on these traits (2.6-fold enrichment, $p = 1.5 \times 10^{-7}$). Signs of selection on some traits previously reported as adaptive (e.g., educational attainment and hair color) are largely attributable to correlated response ($p = 2.9 \times 10^{-6}$ and 1.7×10^{-4}, respectively). Lastly, our joint test shows antagonistic selection has increased type 2 diabetes risk and decrease HbA1c ($p = 1.5 \times 10^{-5}$).

Introduction

Genome-wide association studies (GWASs) have shown that many human complex traits spanning anthropometric, behavioral, metabolic, and many other domains are highly polygenic.[1–3] GWAS findings have strongly indicated that the action of purifying and/or stabilizing selection acts pervasively on complex traits.[4–7] Some work has also utilized biobank data to measure the fitness effects of phenotypes by using direct measurements of reproductive success.[8] However, there are few, if any, validated genomic signals of directional polygenic adaptation in humans.

Several factors have contributed to this uncertainty. Chief among them, polygenicity can allow adaptation to occur rapidly with extremely subtle changes in allele frequencies.[9] Classic population genetics-based methods to detect adaptation via nucleotide data have historically been designed to detect selective sweeps with strong selection coefficients unlikely to occur under polygenic architecture.[10] Polygenic adaptation, after a shift in the fitness optimum, can occur rapidly while causal variants only undergo subtle changes in allele frequency.[11] After a transient period during which the mean of the trait changes directionally, a new optimum is reached and the effect of selection will then largely be to reduce the variance around the mean.[12] However, identifying the genomic footprints of the transient period of directional selection is of substantial interest because it provides evidence of adaptation.

To this end, the advent of GWASs has ushered in a series of methods that take advantage of the availability of allele effect estimates by aggregating subtle signals of selection across association-tested loci. For example, some methods (e.g., the Q_X test) compare differences in population-specific polygenic scores—an aggregate of allele frequencies and allele effect estimates—across populations and test whether they deviate from a null model of genetic drift.[13] Other methods have generalized this test, e.g., to identify and map polygenic adaptations to branches of an admixture graph.[14] Whereas the aforementioned methods exploit between-population differentiation to detect polygenic adaptation, another class of methods is based on within-population variation. For example, selection scans based on singleton density score (SDS) have demonstrated utility in detecting polygenic adaptation via the correlation of SNPs' effect estimates and their SDSs.[15] Another test looks for dependence of derived allele frequencies (DAF) on SNP effect estimates.[16]

Several powerful tests for selection were developed to take advantage of recent advances in ancestral recombination graph (ARG[17]) and whole-genome genealogy inference. Such methods enjoy better power in detecting selection because the ARG, if observed directly, fully summarizes the effects of selection on linked nucleotide data. We note that several methods, notably ARGweaver,[18] infer the strictly defined ARG; by contrast, methods such as Relate[19] infer a series of trees summarizing ancestral histories spanning chunks of the genome. For example, the

[1]Graduate Group in Computational Biology, UC Berkeley, Berkeley, CA 94703, USA; [2]Department of Statistics, University of Oxford, Oxford, UK; [3]David Geffen School of Medicine, UC Los Angeles, Los Angeles, CA 90095, USA; [4]Department of Integrative Biology, UC Berkeley, Berkeley, CA 94703, USA; [5]Department of Statistics, UC Berkeley, Berkeley, CA 94703, USA
*Correspondence: ajstern@berkeley.edu
https://doi.org/10.1016/j.ajhg.2020.12.005.

T_X test estimates changes in the population mean polygenic score over time by using the coalescent tree at a polymorphic site as a proxy for its allele frequency trajectory; the sum of these trajectories weighted by corresponding allelic effect estimates forms an estimate of the polygenic score's trajectory.[20] Speidel, et al. (2019) also designed a non-parametric test for selection based on coalescence rates of derived- and ancestral-allele-carrying lineages calculated empirically from the coalescent tree inferred by Relate.[19] However, these methods ultimately treat the coalescent tree as a fixed, observed variable, where it is actually hidden and highly uncertain. Furthermore, most methods infer the tree under a neutral model and thus provide biased estimates of the genealogy under selection.

To address these issues, we recently developed a full-likelihood method, CLUES, to test for selection and estimate allele frequency trajectories.[21] The method works by stochastically integrating over both the latent ARG with Markov Chain Monte Carlo and the latent allele frequency trajectory with a dynamic programming algorithm and then using importance sampling to estimate the likelihood function of a focal SNP's selection coefficient, correcting for biases in the ARG due to sampling under a neutral model.

Beyond the issue of statistical power, tests for polygenic adaptation can in general be biased when they rely on GWASs containing uncorrected stratification. For example, several groups found strong signals of height adaptation in Europe;[13–15,22–24] later, it was shown that summary statistics from the underlying meta-analysis (GIANT [Genetic Investigation of ANthropometric Traits]) were systematically biased because of uncorrected stratification, and subsequent tests for selection on height failed to be reproduced via properly corrected summary statistics.[20,25,26] However, beyond this case, the extent to which other signals of polygenic selection may be inflated by uncorrected stratification is an open question. Here, we investigate the robustness of the new likelihood method to uncorrected stratification and compare it to another state-of-the-art method (trait SDS [tSDS]), showing that our likelihood method is less prone to bias but has substantially improved power.

Another issue faced by current methods to detect polygenic adaptation is confounding due to pleiotropy. For example, direct selection on one trait may cause a false signal of selection on another, genetically correlated trait. While a variant of the Q_X test has been proposed for the purpose of controls for pleiotropy, this method relies on signals of between-population differentiation to test for selection and is not directly applicable to test multiple traits jointly.[24]

Here, we present a full-likelihood method (Polygenic Adaptation Likelihood Method [PALM]) that uses population DNA sequence data and GWAS summary statistics to estimate direct selection acting on a polygenic trait. We demonstrate robustness by exploring potential sources of bias, including uncorrected GWAS stratification. We also introduce a variant on our method, which controls for pleiotropy by testing ≥ 2 traits for selection jointly. We show our method not only fully controls for this bias but retains high power to distinguish direct selection from correlated response, even in traits with strong genetic correlation (up to 80%), and has unique power to detect cases of antagonistic selection on genetically correlated traits. We explore the behavior of the test when traits with causal fitness effects are excluded to illustrate limitations and proper interpretation of these selection and correlated response estimates.

Material and methods

Model

Linking SNP effects to selection coefficients

Let β be the effect of a SNP on a trait. We model the selection coefficient acting on this SNP by using the Lande approximation[27] $s \approx \beta\omega$, where ω is the selection gradient, the derivative of fitness with respect to trait value. If β is measured in phenotypic standard deviations, then ω is the so-called selection intensity. Chevin et al. (2008) showed that a linked neutral SNP effectively undergoes selection with $s_{tag} \approx rs\sqrt{pq}/\sqrt{p'q'}$, where the neutral "tag" SNP has frequency $p' = 1 - q'$ and genotypic correlation r to a SNP with selection coefficient s and allele frequencies p and $q = 1 - p$.[28] Applying this principle to the multivariate Lande approximation, we find that $s_{tag} \approx \beta_{tag}\omega$, where $\beta_{tag} = \beta \cdot r\sqrt{pq}/\sqrt{p'q'}$ is the marginal effect of the tag SNP, assuming no linkage disequilibrium (LD) between the tag SNP and any other causal SNP.

So long as the effect size is moderate and the population is displaced significantly from its optimal trait value, the Lande approximation is an accurate model for the allelic dynamic of polygenic adaptation under stabilizing selection on short timescales (Figure S3B). Even when the population is close to its optimal trait value (i.e., at ~90% mean absolute fitness), the Lande model has $\leq 15\%$ relative error, with more moderate error at higher minor allele frequencies (MAFs), where SNPs are more likely to be ascertained (Figure S3B). As the population reaches stabilizing selection at equilibrium, relative error of the Lande model becomes more pronounced.

Inferring the selection gradient with a full-likelihood model

Our likelihood model builds heavily on Stern et al. (2019), which developed importance sampling approaches for estimating the likelihood function of the selection coefficient acting on a SNP, $L^{SNP}(s)$.[21] Let $\beta_{(i)}$be the effect of SNP i on the trait. On the basis of these SNP-level selection likelihoods, we model the likelihood function for the selection differential acting on a trait as the product of the SNP likelihoods evaluated at selection coefficients under the Lande approximation:

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}\omega\right), \qquad \text{(Equation 1)}$$

where M is the number of causal SNPs. We provide details for calculating this likelihood function with an importance sampling approach based on Stern et al. (2019) in Appendix A.[21] Given this likelihood function, we estimate ω by using its maximum-likelihood estimate. This model is used by our so-called marginal test, PALM.

Fitness effects of multiple traits

To model fitness effects of multiple traits jointly, here, we propose a multivariate extension of the Lande approximation that links pleiotropic SNP effects to the selection coefficient. Let β be a vector of a particular SNP's effects on d distinct traits. We assume the selection coefficient acting on this SNP follows a multivariate version of the Lande approximation,[27]

$$s \approx \sum_{j=1}^{d} \beta_j \omega_j, \qquad \text{(Equation 2)}$$

where ω now is a vector of selection gradients for each of the d traits. The results of Chevin et al. (2008) apply directly given this approximation for the selection coefficient, and we now express the likelihood of the selection gradient by using Equation 2: $L(\omega) = \prod_{i=1}^{M} L_i^{SNP}(\beta_{(i)}^{\top}\omega)$. We can solve for the maximum-likelihood estimate of ω by jointly using standard optimization. This model is used by our joint test, J-PALM.

Simulations

Pleiotropic polygenic trait architecture

We sample effect sizes jointly for $d = 23$ polygenic traits with previously estimated SNP heritability and genetic correlations.[29,30] We consider different values of polygenicity (M, the number of causal SNPs) and degrees of pleiotropy (ϱ, the probability that a causal SNP is pleiotropic). Let G be the additive genetic covariance matrix (diagonal entries are the SNP heritabilities h_i^2 for each trait i). Then the genetic correlation of traits i, j is $r_{g,ij} = g_{ij}/\sqrt{g_{ii}g_{jj}} = g_{ij}/\sqrt{h_i^2 h_j^2}$. Under our simulation model, we assume that if a SNP is pleiotropic, then $\beta \sim MVN(0, G^*/M\nu)$, where $g^*_{ii} = g_{ii}\cdot(1-(1-\varrho)/d)/\varrho$, $g^*_{i\neq j} = g_{i\neq j}/\varrho$. If a SNP is non-pleiotropic and is causal for trait j, then $\beta_j \sim N(0, h_j^2/M\nu)$, where $h_j^2 := g_{jj}$, and $\beta_{-j} = 0$. We assume that if a SNP is non-pleiotropic, it is causal for a particular trait j with uniform probability $1/d$. Under this model, we can see that averaging over pleiotropic and non-pleiotropic loci, we recover the overall genetic covariance G:

$$\sigma^2_{\beta_j} = (1-\varrho)\Big/ d\cdot h_j^2 + \varrho\cdot(1-(1-\varrho)/d)\Big/\varrho\cdot h_j^2 = h_j^2 = g_{jj}, \qquad \text{(Equation 3)}$$

$$\sigma_{\beta_i,\beta_j} = 0 + \varrho\cdot 1\Big/\varrho\cdot g_{i\neq j} = g_{i\neq j}. \qquad \text{(Equation 4)}$$

Note that, here, β is standardized by the phenotypic variance but not the genotypic variance. Thus, we normalize the variance by a factor of $\nu = 2\cdot E[pq]$, assuming some stationary distribution for the allele frequency $p = 1-q$. Assuming the neutral stationary distribution $f(p) \propto 1/p$ yields $\nu = 4logN_e$, where N_e is the diploid effective population size. This choice of ν ensures $E[\sum_{k=1}^{M} 2\beta_k^2 p_k q_k] = h^2$ under the nominal allele frequency spectrum. The equation holds because we assume independence of effects and allele frequencies; we also performed simulations where β and p are allowed to depend strongly on each other due to purifying selection.

Simulation of confounding due to population structure and uncorrected GWAS stratification

Previous estimates of selection to increase height in Europe have been biased by a combination of uncorrected stratification and GWAS and systematic differences in the coalescent rate at SNPs that depended on their allele frequency difference in 1000 Genomes (1KG) British (GBR) versus Southern Italy (TSI) populations.[25,26] We developed a simulation model based on empirical data from the 1KG data in order to assess the robustness of our method compared to tSDS-based tests for polygenic selection.[15] We model uncorrected stratification in summary statistics for a simulated polygenic trait architecture by drawing random SNP effects

$$\beta \sim N\big(0, h^2/(M\nu)\cdot I\big), \qquad \text{(Equation 5)}$$

where I is the identity matrix. We assume that the phenotype follows the form

$$\phi = X\beta + S + \epsilon, \qquad \text{(Equation 6)}$$

where S is some environmentally determined stratified effect experienced by an individual on the basis of whether they belong to a subpopulation. If N_1, N_2 individuals ($N_1 + N_2 = N$) belong to subpopulations 1 and 2 (e.g., GBR and TSI), respectively, then $S_i = +\sigma_s/\sqrt{N_1/N_2}$ if $i = 1$, $S_i = -\sigma_s/\sqrt{N_2/N_1}$ if $i = 2$. (It can be shown then that phenotypic mean remains 0, and variance due to stratification is σ_s^2.) Under this form of stratification, assuming random mating of genotypes, the expected effect estimate is biased:

$$E\Big[\widehat{\beta}\big|X\Big] = \beta + X^{\top}S\Big/(2Npq) \qquad \text{(Equation 7)}$$

$$= \beta + 2\sigma_s\Big(\sqrt{N_1N_2}f_1 - \sqrt{N_1N_2}\cdot(N/N_2\cdot p - N_1/N_2\cdot f_1)\Big)\Big/(2Npq) \qquad \text{(Equation 8)}$$

$$= \beta + \sqrt{N_1/N_2}\sigma_s(f_1-p)\Big/(pq), \qquad \text{(Equation 9)}$$

where $p = 1-q = (N_1 f_1 + N_2 f_2)/N$ is the overall frequency of the SNP and f_1 is the frequency of the SNP in subpopulation 1. The nominal standard error of $\widehat{\beta}$ is the usual $se(\widehat{\beta}) = 1/\sqrt{2Npq}$.

Hence, we can simulate GWAS-estimated SNP effects with uncorrected stratification by using

$$\beta \sim MVN\big(0, h^2/(M\nu)\cdot I\big) \qquad \text{(Equation 10)}$$

$$\widehat{\beta}\big|\beta \sim N\Big(\beta + \sqrt{N_1/N_2}\sigma_s(f_1-p)\Big/(pq), \sigma_e^2\Big/N\cdot I\Big), \qquad \text{(Equation 11)}$$

where $Z = \sqrt{2Npq}\,\widehat{\beta}$ and $\sigma_e{}^2 := 1 - h^2 - \sigma_S{}^2$. Although in this simple model of GWASs with uncorrected stratification, we assume no LD between causal sites, the bias in the effect estimates does not depend on LD. We note that this is equivalent to the model of Bulik-Sullivan et al. (2015a),[29] generalized to uneven sample sizes from subpopulations.

Population genetic model of selection and ascertainment bias via GWASs

Given β, we simulate selection following the multivariate Lande approximation (see Model). Because we simulate polygenic architectures of $M \geq 100$ assuming no linkage between causal loci, our assumption of infinitesimal genetic architecture is appropriate. (We also explore the performance of our model when we allow LD between causal SNPs; see Figure S4). We then simulate the trajectory of the allele forward in time by using a normal approximation to the Wright-Fisher model with selection, i.e., $p_{t+1} \sim N(p_t + sp_t(1-p_t), p_t(1-p_t)/4N_e)$, where s is calculated with the multivariate Lande approximation. For most of our simulations, we simulate forward for 50 generations (i.e., we assume

selection began 50 generations before the present), unless otherwise stated. Let p be the present-day allele frequency. We simulate the ascertainment of this SNP in a GWAS by simulating the SNP Z scores $Z \sim MVN(\sqrt{2Npq}\,\beta, E)$, where $E_{ii} = 1, E_{i \neq j} = \rho_e$, where ρ_e is a term that allows for cross-trait correlations in environmental noise. (Note that, here, Z is the usual Z score of $\widehat{\beta}$, not to be confused with the selection Z score we present earlier.) Unless stated otherwise, we set $N = 10^5, \rho_e = 0.1$ in all simulations. We use a p value threshold of 5×10^{-8} to ascertain a SNP; this must be surpassed by at least one trait. If a SNP is ascertained, we simulate its trajectory backward in time by using the normal approximation to the neutral Wright-Fisher diffusion conditional on loss, $p_{t-1} \sim N(p_t(1 - 1/4N_e),\ p_t(1 - p_t)/4N_e)$. We use the coalescent simulator mssel to simulate a sample of haplotypes conditional on this allele frequency trajectory.[20] We use $n = 400$ haplotypes and $\mu = r = 10^{-8}$/bp/gen and simulate regions of 1 Mbp, centered on the causal SNP at the position 5×10^5.

To simulate ascertainment of non-causal SNPs in a GWAS, we take the trait with the top Z score at the causal SNP and jointly simulate Z scores for that trait for all linked SNPs within a 200 kbp window centered on the causal SNP and surpassing an MAF threshold (MAF $\geq$ 0.01). We ascertain the SNP with the top Z score (sometimes the causal SNP) and then simulate the Z scores for all traits, conditioned on the Z score for the one aforementioned trait. We simulate this way rather than jointly simulating Z scores for all traits at all SNPs for two reasons: the top SNP will typically have the same top trait association as the causal and jointly simulating all trait-by-SNP Z scores increases computational time by >400 for the parameters we used.

To further reduce computational burden, we simulated libraries of $10 \times M$ causal loci and resampled sets of M loci without replacement (some proportion of which meet the ascertainment criteria) in order to model sampling variation in the test statistics.

Inference of local genealogies

Given a set of simulated haplotypes, we use the software package Relate[19] to infer local genealogies along the sequence. Using positions of the SNPs ascertained through GWASs, we use the add-on module SampleBranchLengths to draw $m = 5,000$ Markov chain Monte Carlo (MCMC) samples of the branch lengths of the local tree at the ascertained sites. We then extract coalescence times from these MCMC samples (thinned down to $m = 500$ approximately independent samples) and partition the coalescence times for each sample tree on the basis of whether they occur between lineages subtending the derived/ancestral alleles. We note that Relate, unlike ARGweaver, does not sample over different ARG or tree topologies and it samples branch lengths for two distinct local trees independently, conditional on the observed data.

Comparisons to tSDS in simulations

In order to calculate tSDS values for our simulated polygenic traits, we computed the Gamma shape parameters for a model with constant $N_e = 10^4$ by using 250 simulations at a range of DAFs from 1% to 99%, with 2% steps between frequencies, and a sample size of $n = 400$ haplotypes. We randomly paired haplotypes in the sample to form diploid individuals and found singletons carried by each diploid. We then calculate raw SDS by using the compute_SDS.R script with our custom Gamma-shapes file. To calculate SDS, we find the Z score of a SNP's raw SDS value, where the mean and standard deviation are estimated from an aggregated set of 29,478 completely unlinked SNPs from our neutral trait simulations. To calculate tSDS, we calculate the p value of the Spearman correlation of $(\text{sign}(\widehat{\beta}), SDS)$.

Results

Simulations

Overview of simulations

We conducted evolutionary simulations of polygenic adaptation acting on a wide range of multi-trait polygenic architectures. Our simulated traits are based on SNP heritability and genetic correlation estimates for 23 real human traits;[29,30] unless otherwise stated, we simulate positive selection on/test for selection on a trait modeled after the heritability of schizophrenia ($h^2 = 0.45$), and in most of our pleiotropy analyses, we used parameters based on schizophrenia and its genetic correlation with three other traits: bipolar disorder, major depression, and anorexia. In most of our analysis, we refer to these traits as trait I/II/III/IV (corresponding to models of schizophrenia/bipolar/depression/anorexia, respectively). As our method is based on aggregating population genetic signals of selection with GWAS summary statistics, we also simulated GWASs in samples of modern-day individuals ($N = 10^5$). Since our method works by taking the product of likelihoods across independent sites, we simulate LD pruning in independent LD blocks, each of 2 Mb in length; LD blocks with a minimum Wald test p value of greater than genome-wide significance are excluded from further analysis. Our simulated summary statistics incorporate all of the following sources of bias found in GWASs, unless stated otherwise: random noise in the effect estimates; Winner's Curse bias in the effect estimates (unless stated otherwise, we ascertain SNPs at genome-wide significance with associations $p < 5 \times 10^{-8}$ for at least one trait analyzed; we perform this ascertainment in part to mitigate the relative bias of effect size estimates due to uncorrected stratification); uncertainty in the location of the causal SNP (we ascertain the top GWAS hit throughout the linked region); and environmentally correlated noise across traits (only relevant to simulations of pleiotropic architectures). Average selection coefficients, allele frequency changes, and population phenotype changes are detailed in Table S1. Furthermore, we also simulate a number of scenarios that violate our model assumptions to assess our method's robustness: these include uncorrected GWAS stratification, purifying/stabilizing selection, underpowered/uneven GWAS sample sizes, and allelic heterogeneity (i.e., multiple linked causal SNPs).

For each causal locus, we simulate haplotype data for a sample of $n = 400$ 1 Mbp-long chromosomes (mutation and recombination rates $\mu = r = 10^{-8}$ and effective population size $N_e = 10^4$ unless stated otherwise) on which we applied Relate, a state-of-the-art method for tree inference,[19] to infer the coalescent tree at SNPs ascertained in this GWAS. However, we point out that our approach can be applied to any pre-existing method for estimating/sampling these trees (e.g., ARGweaver[18]). We then conduct importance sampling to estimate the likelihood function of the selection gradient—i.e., the effect of a unit increase in phenotypic values on fitness—for

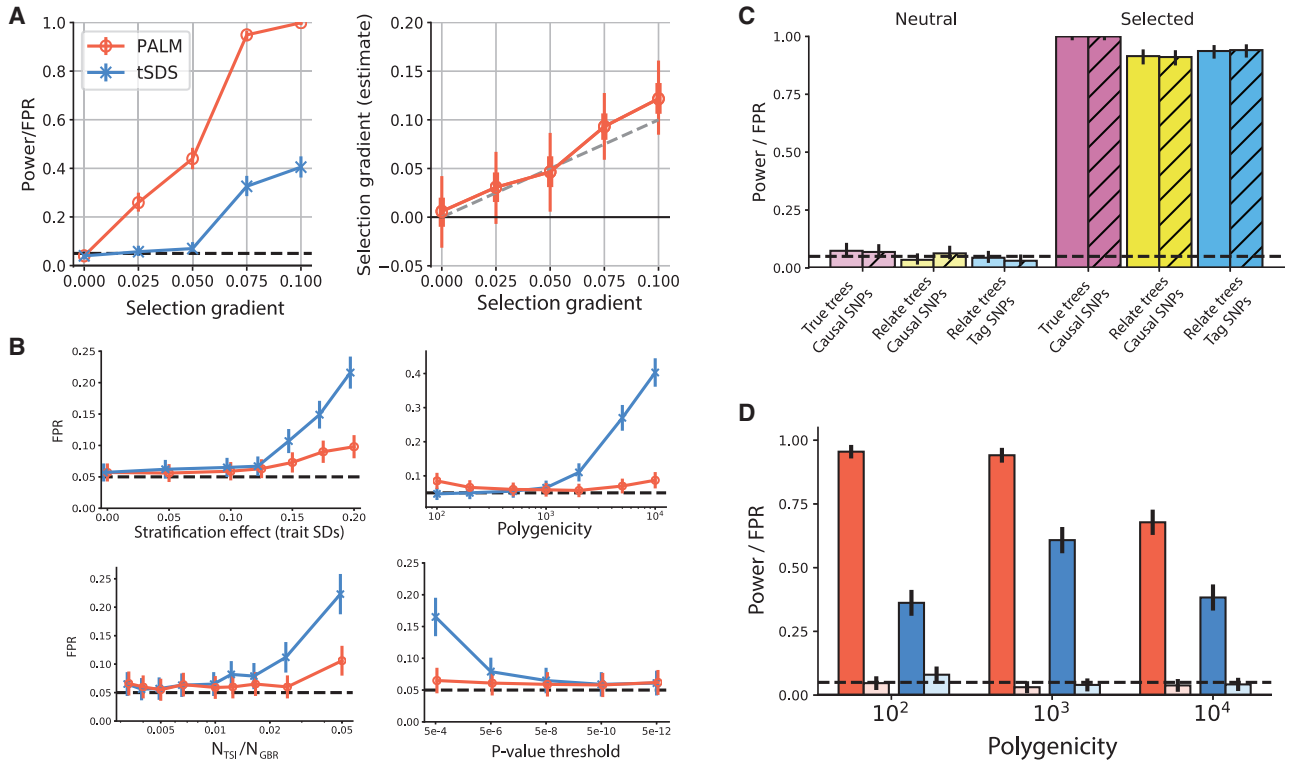

Figure 1. PALM power, calibration, and robustness to uncorrected stratification and ascertainment

(A) Left: Power/false positive rate (FPR) of PALM and tSDS. Error bars denote 95% Bonferroni-corrected confidence intervals. Right: PALM selection gradient estimates ($\widehat{\omega}$). Error bars denote 25th–75th percentiles (thick) and 5th–95th percentiles (thin) of estimates; see Table 1 for more details of $\widehat{\omega}$ moments and error. Markers and colors in (A) also apply to (B) and (D).

(B) FPR of PALM and tSDS applied to neutral simulations with uncorrected population stratification, simulated with 1000 Genomes data. We used baseline values of $\sigma_S = 0.1$, $N_{TSI}/N_{GBR} = 1\%$, $M = 10^3$, $h^2 = 50\%$ by using SNPs ascertained at $p < 5\times10^{-8}$. Error bars denote 95% Bonferroni-corrected confidence intervals.

(C) Comparison of PALM with true versus Relate-inferred trees, causal versus GWAS-ascertained tag SNPs, and true marginal SNP effects (solid) versus GWAS-estimated SNP effects (hatched). Error bars denote 95% Bonferroni-corrected confidence intervals.

(D) Varying polygenicity (M) of the polygenic trait. Error bars denote 95% Bonferroni-corrected confidence intervals. Baseline parameters for all simulations except (C) were our constant-size model with $M = 10^3$, with Scz under positive selection and testing Scz for selection. In (A) and (B), we use Relate-inferred trees and estimated SNP effects at the causal SNPs; in (D), we use Relate-inferred trees and estimated effects at tag SNPs. In all panels, we use a 5% nominal FPR (dashed horizontal line) and simulated 10^3 replicates. In (D), light/saturated colors signify neutral/selected simulations.

individual traits (i.e., estimated marginally), as well as sets of genetically correlated traits (i.e., estimated jointly). Our method, Polygenic Adaptation Likelihood Method (PALM), can be used to estimate ω for polygenic traits.

Improved power to detect selection and estimates of the selection gradient

We ran PALM to test for selection on our simulations of polygenic trait architectures, described above (and in more detail in Appendix A). We estimate the selection gradient and standardize this quantity by its standard error, estimated through block bootstrap, to conduct a Wald test on whether the selection gradient is non-zero.

First, we conducted simulations at different values of the selection gradient, ranging from neutral ($\omega = 0$) to strong ($\omega = 0.1$, average change of mean phenotype of ~2 standard deviations) and compared the statistical power of PALM to that of tSDS (Figure 1A). Summaries of SNP selection coefficients, allele frequency changes, and phenotypic changes are detailed in Table S1. We simulate 5 Mb haplotypes for a trait with polygenicity (i.e., number of causal SNPs) $M = 1,000$; we sample $n = 178$ haplotypes for PALM and $n = 6,390$ for tSDS, corresponding to the sample sizes we used in our application to 1000 Genomes British (GBR) individuals versus the sample used by Field et al. (2016)[15] from the UK10K. Here we ascertain only causal SNPs, but SNP effects are still estimated through an association test (unless otherwise stated, all other simulations incorporate uncertainty in the causal SNP). Both methods are well calibrated under the null ($\omega = 0$, Figure 1A). But we find that despite having a much smaller sample size, PALM has substantially improved power to detect selection at all levels (Figure 1A), especially at weaker values of the selection gradient, where tSDS has essentially no power ($\omega \leq 0.05$). PALM is also capable of estimating the selection gradient (Figure 1A, Table 1). These estimates are well calibrated, and empirical standard errors closely match estimated standard errors, except when the selection gradient is exceptionally strong ($\omega \geq 0.1$) (Table 1).

Table 1. Selection gradient estimates and standard errors

ω	Mean $\widehat{\omega}$	SD($\widehat{\omega}$)	MSE($\widehat{\omega}$)	Mean SE($\widehat{\omega}$)
0	0.0053	0.0226	0.0232	0.0246
0.025	0.0306	0.0225	0.0232	0.0243
0.05	0.0465	0.0243	0.0245	0.0266
0.075	0.0931	0.0211	0.0278	0.023
0.1	0.1223	0.0236	0.0325	0.0255

Summary statistics for the accuracy and calibration of estimates also used in Figure 1 (see caption for simulation details). Mean SE is the mean nominal standard error. Simulations are the same as used in Figure 1A. SD, standard deviation; MSE, mean squared error.

We also examined the calibration and power of the marginal test in simulations of a polygenic trait with varying polygenicity (Figure 1D). Across a wide range of polygenicities, PALM is well powered to detect selection (>90% for $100 \leq M \leq 1,000$) and the false positive rate (FPR) was well calibrated in all circumstances (Figure 1D). Although PALM had slightly lower power for extremely polygenic architectures (65% ±3% for $M = 10^4$), we found that when the causal effect sizes and true genealogies are known, this power significantly increases (88% ± 2%), suggesting that ability to detect selection on extremely polygenic traits is limited by ARG and GWAS effect size estimates. In comparisons to tSDS, we found substantially improved statistical power across this range of polygenicity values (Figure 1D). We also conducted similar tests for a short pulse of selection ($\omega = 0.05$ for 35 generations, or ~1,000 years assuming 29 years/generation) under a model of British demography;[19] we found that overall power was comparable to that of constant population size simulations with $\omega = 0.025$, consistent with previous work showing that the product of selection strength and time span largely determines statistical power (Figure S2). Lastly, we conducted a test for a 50-generation-long pulse of selection starting 250 generations before the present (~7.3 kya, assuming 29 years/generation), which had 23% (±3%) power to detect the pulse, whereas the test was well calibrated under the null (FPR of 4% ± 1%); this illustrates that our method is well calibrated and has (although attenuated) power to detect selection on more ancient timescales. However, we caution that certain model violations, such as changes in environment, LD and marginal effect sizes, and unmodeled population structure, may potentially cause biases.

Robustness to uncorrected GWAS stratification

We compared the power curve to the FPR of both methods under a model of uncorrected GWAS stratification (Figure 1B). We simulated polygenic trait architectures and GWASs such that estimated SNP effects ($\widehat{\beta}$) were both systematically biased and correlated with differences in the coalescence rate, stratified by DAF (e.g., SDS), matching the findings of Berg et al. (2019)[25] and Sohail et al. (2019)[26] that allele frequency differentiation between British (GBR) and Toscani in Italia (TSI) individuals was positively correlated with both $\widehat{\beta}$ and SDS (Figure S1).

To model this scenario, we ascertained a set of 40,320 SNPs with MAF > 0.5% in the UK Biobank (UKBB) and SDS calculated by Field et al. (2016) using the UK10K cohort.[15] We then sampled coalescence times at these SNPs in 1KG Phase 3 GBR individuals by using Relate. For each SNP, we simulated GWAS summary statistics by assuming that the GWAS cohort is comprised of some ratio, N_{TSI}/N_{GBR}, of TSI to GBR individuals, where population identity determines an individual's stratified effect. This induces a correlation between SNP effects and the difference in allele frequency between TSI and GBR. Baseline parameter values were $\sigma_S = 0.1$, $N_{TSI}/N_{GBR} = 1\%$, $M = 1,000$, and $p = 5 \times 10^{-8}$. We varied the strength of the stratified effect (σ_S, in phenotypic standard deviations) and found that both methods are well calibrated when σ_S is sufficiently small, but as σ_S grows past 0.1, the FPR of tSDS was inflated over 100% more than that of PALM (Figure 1B).

We stress that this disparity is most likely not caused by higher sensitivity of tSDS, as we simulated polygenic adaptation under similar parameters and found PALM was better powered to detect selection, with up to 8× improvement in power for smaller values of the selection gradient (Figure 1A). We also found that for highly polygenic traits (e.g., $M = 2 \times 10^3$), the tSDS test is overconfident (>10% at 5% nominal), while PALM remains well calibrated (Figure 1B). We observe the same pattern as we increase the size of the cohort subgroup receiving the stratified effect (N_{TSI}/N_{GBR}); at $N_{TSI}/N_{GBR} = 2.5\%$, the tSDS test is overconfident (>10% at 5% nominal), while PALM remains well calibrated (Figure 1B).

Lastly, we tested the sensitivity of these methods to the stringency of the p value threshold used and found that the tSDS test was increasingly overconfident as the threshold was relaxed, whereas PALM was well calibrated regardless of p value threshold (Figure 1B). These results suggest that PALM is more robust to uncorrected stratification than the tSDS test and obtains superior statistical power even at lower sample sizes. However, we emphasize that PALM, like any other available test, is not fully robust to the effects of uncontrolled population stratification. Sufficiently strong uncorrected population stratification can lead to false inferences of polygenic selection when there is none. These results further demonstrate that cryptic population structure (e.g., systematic differences in the

allele frequency spectrum between TSI and GBR) in the absence of a stratification effect (i.e., $\sigma_S = 0$) does not incur a bias in PALM's test for selection.

We note, however, that this experiment does not exhaust the possible scenarios of uncorrected stratification that might bias a test for polygenic adaptation. For example, it may be that more recently formed population structure is harder to correct with standard tools; this might correspond to population structure within populations (e.g., within GBR), represented by higher-order principal components of Europeans in 1000 Genomes.

Robustness to ascertainment bias and uncertainty in GWAS estimates

Next, we considered the effects of different levels of uncertainty and ascertainment on the performance of PALM (Figure 1C). We considered the effects of conditioning on the true local tree versus with Relate-inferred trees combined with importance sampling, conditioning on the true marginal SNP effect versus estimating this effect with noise in a GWAS, and conditioning on the causal SNP versus taking the top tag SNP in a local GWAS on linked SNPs. PALM was well calibrated with both true trees and importance sampling, with highest statistical power (100%) with true trees and a slight drop in power under importance sampling (90%–92%) (Figure 1C). Our test was well calibrated despite bias (from winner's curse) and noise in the estimated SNP effects, with no discernible difference from using the true SNP effects (Figure 1C); however, for smaller sample sizes ($N << 10^5$) this may not be the case. Lastly, using the causal SNPs versus GWAS-ascertained tag SNPs did not diminish test power, and FPR remained well calibrated (Figure 1C). We also explored the effects of GWAS sample size, which will affect the ascertainment process and, hence, the degree of bias and uncertainty in ascertained SNP effect estimates (Table S2). We considered two different GWAS sizes: $N = 10^4$ and 10^5. We found that under lower sample size, the test was slightly inflated (e.g., empirical FPR of 3.1% [±1.4%] and 7.0% [±1.6%] at $N = 10^5$ and 10^4 for trait II, respectively (brackets denote 95% CIs; Table S2). In terms of power, the test is still well powered at lower sample sizes, but there is a noticeable drop (94.1% [±1.4%] and 69.0% [±3.0%] at $N = 10^5$ and 10^4, respectively; Table S2).

Robustness to model violations

We also conducted simulations of polygenic trait architectures under purifying selection based on the model proposed by Schoech et al. (2019)[7] (Figure S3). Under such a scenario, an inverse relationship between effect size magnitude and DAF is expected, in contrast to our baseline simulation model in which effect size is independent of frequency prior to the onset of selection. We found that across a range of polygenicities ($M = 3\times10^3, 10^4, 3\times10^4$) and selection strengths ($2N\bar{s} = 3, 10, 30$, where $\bar{s}$ denotes mean selection coefficient of causal SNPs), PALM is not confounded by purifying selection and is well calibrated to a nominal FPR of 5% (Figure S3); in fact, under very strong selection and/or low polygenicities, PALM is slightly conservative (Figure S3).

As our model and baseline simulations assume a single causal SNP per linked locus, we conducted simulations of allelic heterogeneity (Figure S4) by using forward simulations in SLiM.[31] We simulated a trait architecture with $h^2 = 50\%$ and a mutational target of 100×1 Mbp linked loci, considering two cases: (1) 5% of incoming mutations are causal and (2) 50% of incoming mutations are causal. In each of these scenarios, we conducted simulations with neutral evolution and adaptation. We found that in each case, the test is well calibrated under the null and well powered to detect selection (Figure S4).

We also explored the time specificity of PALM's test for selection. Testing under a nominal model of selection in the last 50 generations, we consider the rate at which PALM's estimate of selection timing can be biased by older selection (Figure S5). We found that as selection recedes into the past, the FPR decays toward the nominal rate, with limited confounding when the pulse of selection occurred 200–250 generations ago. Although we did not directly test the ability for PALM to detect more ancient selection under the correct nominal model (i.e., assuming ancient selection), one can interpret the FPR in Figure S5 as a lower bound on the power.

Lastly, we tested the robustness of PALM under mutational biases (Figure S15). We used a model where, with probability $0 \leq b \leq 1$, the sign of the effect of the derived allele is set to +1. (In other words, $b = 1$ creates 100% mutational bias for trait-increasing alleles, whereas $b = 0$ returns the original model.) We simulate no selection in this scenario. We find that for moderate values of b, the test is well calibrated; however, substantial mutational bias can bias the test (Figure S15).

We summarize all of these various robustness measurements and previous tests (e.g., stratification) of PALM in Table 2.

Pleiotropy can cause bias in tests for polygenic adaptation

Traits with no fitness effect can undergo correlated response due to direct selection on pleiotropically related traits. Without accounting for pleiotropy, standard tests for polygenic adaptation cannot be interpreted as statements regarding direct selection. To illustrate how pleiotropy can affect tests for polygenic adaptation, we simulated pleiotropic trait architectures for 23 traits based on estimates of SNP heritability and genetic correlation for real human traits.[30] This builds largely off our aforementioned simulation approach, with the introduction of a parameter ϱ, the degree of pleiotropy, i.e., the probability that a causal SNP is pleiotropic. As a brief illustration of how pleiotropy causes bias in polygenic selection estimates, we used our pleiotropic traits simulations to estimate maximum likelihood selection coefficients for SNPs ascertained for associations to two genetically correlated traits, trait I and II, modeled after schizophrenia and bipolar disorder ($r_g \approx 80\%$; Figure S6). We simulate a pulse of selection to increase trait I ($\omega = 0.05$, approximately +1 standard

Table 2. Summary of PALM/J-PALM robustness tests

Scenario	Well calibrated?
Uncorrected stratification	~
Uncertainty in causal SNPs	✓
Uncertainty in marginal effect sizes	✓
Uncertainty in genealogies	✓
Low GWAS sample size	✓
Polygenicity	✓
Allelic heterogeneity	✓
Purifying/stabilizing selection	✓
Non-equilibrium demography	✓
Mutational bias	~
Time specificity	~
Environmental covariance	✓
Varying pleiotropy	✓

Here, we summarize robustness to various scenarios and model violations based on our tests of empirical versus nominal false positive rates. Check marks (✓) signify uniform/overwhelming robustness; tildes (~) signify robustness under moderate conditions.

deviation change in population mean over 50 generations, Table S1); trait 2 has no causal effect on fitness. We estimated selection coefficients by taking the maximum likelihood estimate of s for each SNP independently, where the likelihood is estimated with our importance sampling approach. Here, we show results for polygenicity $M = 1,000$ and degree of pleiotropy $\varrho = 60\%$ (Figure S6).

Under the Lande approximation $s \approx \beta^\top \omega$, we expect a non-constant linear relationship between $\widehat{\beta}$ and $\widehat{s}$ for traits under selection. But as a result of the strong correlation between these two traits, it is difficult to disentangle which of the traits has a causal effect on fitness (Figure S6A). We performed an ad hoc test for a systematic relationship between $\widehat{\beta}$ and $\widehat{s}$ (Spearman test) to detect polygenic adaptation; while this test is well powered to detect selection on trait I, it is prone to spurious hits for selection on trait II, which has no effect on fitness (Figure S6B). Thus, marginal tests for selection on traits can be significantly biased because of pleiotropy (in this case, genetic correlation).

Joint test for polygenic adaptation controls for pleiotropy

We also introduce a variant on our method, J-PALM, which is designed to disentangle correlated traits under selection and control for confounding due to pleiotropy. Briefly, J-PALM uses the same likelihood approach as PALM, but we jointly infer the selection gradient ω on a set of d traits jointly, rather than inferring the selection gradient on a single trait marginally (see Model and Appendix A for details). Under the joint model, the likelihood is still a function of the selection coefficient of each SNP, but we allow that these selection coefficients depend on the fitness effects of d traits jointly (see Model and Equation 2).

We applied both our marginal test, PALM, and our joint test, J-PALM, to our cluster of four simulated traits, traits I–IV, modeled after SNP heritabilities and genetic correlations for four psychiatric traits: schizophrenia, bipolar disorder, major depression, and anorexia (Figure 2A). All traits have significantly positive genetic correlation to one another; here, we highlight their genetic correlations to the selected trait, trait I (Figure 2A; genetic correlations and SNP heritabilities directly from Bulik-Sullivan et al. [2015a][29] and Bulik-Sullivan et al. [2015b][30]). We assume a pulse of recent selection for increased trait I prevalence, with all other traits selectively neutral. We tested traits marginally and jointly (i.e., all four simultaneously) for selection (Figures 2B and 2C). We found that marginal estimates are biased and cause inflation of the FPR when testing for selection (Figures 2B and 2C). This bias largely follows the genetic correlation of the estimand trait to the selected trait (Figures 2A and 2B). Here, we show results for polygenicity $M = 1,000$ and degree of pleiotropy $\varrho = 100\%$ (Figure 2), but the results are similar for differing degrees of pleiotropy (holding r_g constant), such as $\varrho = 60\%$ (Figure S7). This highlights that genetic correlation, regardless of the degree of pleiotropy, is the main cause of bias in marginal estimates of the selection gradient.

Furthermore, our results show that if any trait in a genetically correlated cluster is under selection, marginal estimates of the selection gradient for the other traits is typically highly inflated. For example, a genetic correlation as low as $r_g = 19\%$ is sufficient to inflate the FPR for a neutral trait by nearly 150% (Figures 2A and 2C). Most traits studied in GWASs have large genetic correlations; Watanabe et al. (2019) found an average $|r_g| = 16\%$ across 155,403 human trait pairs, with 15.5% of trait pairs significant (average $|r_g| = 38\%$).[32] The extent of strong genetic correlation suggests that if any single heritable trait has evolved under selection, it is likely to cause substantial ripple effects in terms of bias of selection estimates on other heritable traits. By contrast, estimates of selection obtained via our joint test fully correct for these biases if the relevant selected trait is included in the analysis (Figures 2B and 2C). We applied the joint test to the same set of simulations and found it can reliably detect and attribute selection to trait I (Figures 2B and 2C). The joint test preserved ~80% power even with the leading genetic correlate's, trait II, having $r_g = 79.4\%$ to Trait I and produces well-calibrated FPR regardless of r_g (Figure 2C).

We explored performance of J-PALM under a wide array of simulation scenarios of different polygenic architectures and types of selection (Figure 5), varying the degree of pleiotropy ϱ (Figure 3A), r_g to the selected trait (Figure 3B), polygenicity M (Figure 3C), and antagonistic selection (Figure 3D). Baseline values of parameters used were positive selection on trait I with other traits neutral, jointly testing trait I and trait III ($r_g = 51\%$, $\varrho = 60\%$, and $M = 1,000$). All of our pleiotropic simulations include an environmental noise correlation across traits of $\rho_e =$

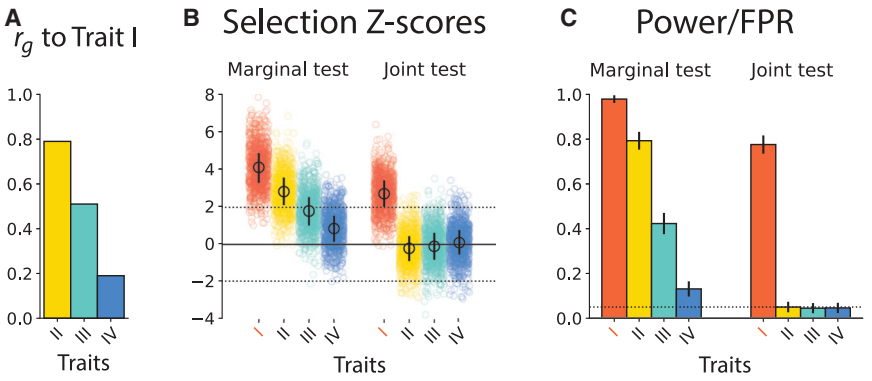

Figure 2. Joint testing for polygenic adaptation controls for pleiotropy

(A) We simulated a cluster of four traits (I–IV) modeled after real human heritability and genetic correlation estimates for schizophrenia (I), bipolar disorder (II), major depression (III), and anorexia (IV), with selection to increase trait I in the last 50 generations.

(B and C) We ran marginal and joint tests for selection on these four traits. While marginal selection tests were well powered, they were strongly biased by even fairly low genetic correlations. Conducting a joint test fully controls for genetic correlations while retaining high power to detect and isolate selection on trait I. Simulations (1,000 replicates) were done under our constant effective population size model with $\varrho = 60\%$, $M = 1,000$, with trait I under positive selection.

10%. Across this range of pleiotropic and polygenic architectures, we established that the joint test is well calibrated when no traits are under selection (Figure S8). Across different degrees of pleiotropy ($40\% \leq \varrho \leq 100\%$), we found J-PALM was well calibrated and had good power to detect and attribute selection to trait I (Figure 3A).

Across a range of levels of polygenicity ($100 \leq M \leq 10,000$), PALM was well calibrated and had good power to detect and attribute selection to trait I (>75% for $M \leq 3,000$), although the power is somewhat attenuated for extremely polygenic architectures (~40% for $M = 10,000$) (Figure 3B). This pattern is also found in the marginal tests on the same data, and there is only a modest reduction in power when switching to the joint test (Figures 1C and 3B). We note that the reduction in power is sensitive to the strength of genetic correlation; the joint test of trait I versus trait II ($r_g = 79\%$) had greater reduction in power from the marginal test than that of trait I versus trait III (Figures 1C, 3B, 3C, and S9). Our method fully corrects the biases suffered by marginal tests for polygenic adaptation while retaining good power to detect adaptation even when genetic correlation is strong.

We also examined what happened when selection acted on different traits in the cluster, jointly testing each selected trait with trait II (Figure 3C). The test is well calibrated for all traits, but has less power to attribute selection to traits with a high genetic correlation to trait II (e.g., trait I, $h^2 = 45\%$, $r_g = 79\%$) or low heritability (e.g., trait III, $h^2 = 17\%$, $r_g = 48\%$) (Figures 1E and 3C). By contrast, traits with high heritability and/or low genetic correlation to trait II (e.g., trait IV, $h^2 = 49\%$, $r_g = 11\%$) have little loss in power in the joint test (Figures 1E and 3C). We summarize the calibration of J-PALM under these various conditions in Table 2.

Detecting antagonistic selection

We also considered the possibility of antagonistic selection (i.e., selection to both increase trait I and decrease trait II, Figure 3D). We hypothesized that marginal tests would be underpowered to detect this mode of selection acting on traits with strong genetic correlation and that joint testing might uncover this signal. Indeed, power to detect selection in this regime is quite low with marginal testing, with 3%–13% power at a 5% threshold (Figure 3D). However, the joint testing boosts power significantly, with 40%–51% power at a 5% threshold (Figure 3D). We also tested the opposite scenario, where trait I and trait II are both under positive (complementary) selection; we found the joint test is well powered to detect that multiple genetically correlated traits are under selection (Figure S10). Thus, J-PALM provides several gains in power over the marginal test, such as uncovering antagonistic selection that is "cancelled out" by genetic correlation or confirming multiple traits are under selection.

Interpretation and limitations of the joint test

We also considered how our joint test performs when the causal trait (i.e., a trait with a causal effect on fitness) is excluded from the model. We conducted pairwise joint tests on each pair of traits I–IV in simulations with trait I under selection and all other traits neutral (Figure 3E). Rows correspond to the trait for which the selection test is performed (the focal trait), and columns correspond to the other trait included in the joint model (the conditional trait). We also considered other scenarios, such as all traits neutral, complementary selection, and antagonistic selection (Figure S11).

As we demonstrated previously, when the causal trait (trait I) is included, the selection test is well calibrated for neutral traits (Figure 3E). However, we find that when trait I is excluded, the selection test has high positive rates for traits that have no causal fitness effect but are strongly genetically correlated with the causal trait (e.g., trait II). In general, our results demonstrate that selection tends to be attributed to the trait with the strongest genetic correlation to the causal trait (e.g., trait II); other traits with genetic correlation to the causal trait (e.g., trait III) have some minor inflation of the positive rate, but selection is predominantly attributed to the closest proxy for the causal trait. These results highlight an important limitation of our model: namely, the selection gradient estimates are not to be interpreted as causal fitness effects. As our simulated results show, this proposition is generally false when a trait with causal fitness effect and nonzero genetic correlation is excluded.

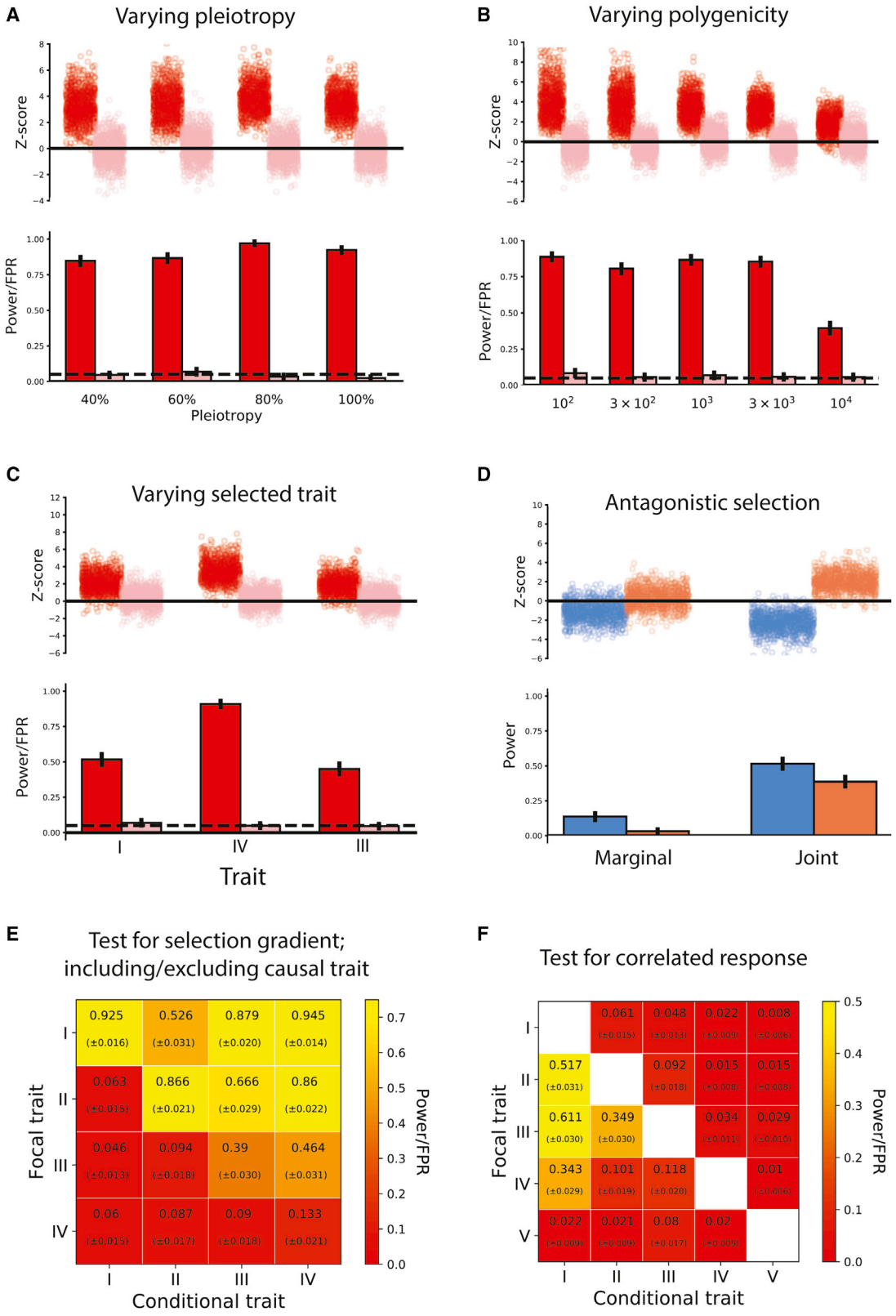

Figure 3. Simulations of joint testing power and calibration
(A–F) Differing the degree of pleiotropy ϱ (A), the trait truly under selection (B), the polygenicity M of the traits (C), antagonistic selection on two traits with positive genetic correlation (D), pairwise tests for selection (trait I under selection) (E), and pairwise tests for correlated response (trait I under selection) (F). In (A)–(D), red/pink/blue bars indicate estimates of selection for traits under positive selection/neutrality/negative selection. In (E) and (F), heatmap is colored by positive rate (also text in boxes; standard errors in parentheses).

(legend continued on next page)

Interpretation and limitations due to GxE

We also modeled gene-by-environment (GxE) interactions. We considered the worst-case GxE scenario in which it is statistically impossible to distinguish GxE from direct genetic effects (Figure 4). Such is the case when, e.g., a study is conducted in a completely homogeneous environment in which the environment causes a trait to become heritable. For example, suppose a study is conducted on Planet A, where a social policy dictates that admission to college is contingent on one's skin color (e.g., 80% heritable) exceeding a threshold. In the absence of the policy, the heritability of "years of education" (EduYears) might be 0%; however, the environment incurs heritability >> 0%, and in a strict (although not biological) sense, skin color is causal for EduYears on Planet A. Furthermore, unless the study is heterogeneous with respect to the relevant environment (e.g., extending to some Planet B without such a policy), it is intractable to correctly attribute to GxE the heritability of EduYears (and its genetic correlation to skin color).

We note that such "intractable" GxE reduces to the main simulation model we have used, since direct effects on one trait (e.g., trait I) are consistently translated into effects on another trait (e.g., trait II) via a homogeneous environmental condition. Following our previous simulations, we find that under neutrality, such intractable GxE does not incur an excess of false positives on the marginal test (Figure 1A). Supposing direct selection on trait I (on which the environment acts on to incur heritability of/genetic correlation with trait II), the marginal test will ascribe selection to trait II (Figure 2B); however, provided the traits are genetically non-collinear (e.g., $r_g \leq 80\%$), then the joint test will correctly ascribe direct selection to trait I in a joint test where this trait is included (Figure 3E). In the case that selection is acting directly on trait II (which is only heritable via GxE), the model will detect selection on this trait in both marginal and joint tests; thus, we encourage caution in interpreting estimates of the selection gradient, as it is not capable of distinguishing selection induced by GxE versus selection on a non-environmentally contingent trait.

Testing for correlated response

Our method can also test for correlated response to selection, i.e., whether a trait has evolved (at least in part) as a result of selection on some other genetically correlated trait. We introduce the notion of an "effective selection gradient" $(\omega_{trait,model})$, which measures attributable amounts of selection to each trait included in a model. Consider two traits, A and B. Suppose trait A is under selection and trait B is neutral. If $r_g = 0$, the effective selection gradient of B is 0, regardless of selection on trait A or whether we include trait A in the model, because no selection on A is attributable to B. Hence, $\omega_{B,marginal} = \omega_{B,joint}$. By contrast, if $|r_g| > 0$, marginally trait B has a nonzero effective selection gradient; however, in a joint model with trait I, the effective selection gradient of trait II is 0, since all direct selection can be attributed to trait I. Hence, because of correlated response, there is a difference in the effective selection gradient in the two models: $\omega_{B,marginal} \neq \omega_{B,joint}$. However, the converse is not true for trait I; both marginally and jointly with trait II, all selection can be attributed to trait I, and so $\omega_{A,marginal} \neq \omega_{A,joint}$. We developed a test statistic R (see Equation A10 in Appendix A) that tests for correlated response under the null hypothesis $H_0 : \omega_{j,marginal} = \omega_{j,joint}$, i.e., that the marginal and joint effective selection gradients are equal.

We conducted tests of correlated response on each pair of traits I–V (we introduce trait V, which has $r_g = 0\%$ to trait I) (Figure 3F). We found that the test for correlated response of trait I is null, concordant with all other traits in the simulation's being neutral (Figure 3F). We also saw that for trait V, which has no genetic correlation to the directly selected trait, the test is null, concordant with the necessity of genetic correlation to drive correlated response (Figure 3F). We saw that tests for correlated response generally grew in their power as r_g to trait I increased. However, power is slightly lower for $r_g = 80\%$ than $r_g = 50\%$ (i.e., testing trait II versus trait III for correlated response to trait I) (Figure 3F). This may indicate that for strongly genetically correlated traits, it is often ambiguous which one of the traits is evolving in correlated response. The test is also well calibrated under neutral simulations (Figure S12A) and well powered to detect more complex forms of correlated response, such as antagonistic and complementary selection (Figures S12B and S12C). We also explored the performance of the correlated response test, along with the joint test for selection, in a K-way model with traits I–IV tested jointly (Figure S13). Our results indicate that our test statistic R can be used to detect whether a trait has been under correlated response; however, it is incorrect to make strongly causal interpretations of the test (e.g., "trait III is under correlated response to trait II").

Effect of small or uneven GWAS sample size

We tested the effect of GWAS sample size on the joint test, considering not only lower sample size but also uneven sample sizes (Table S2). Similar to the effect of lower sample size on the marginal test, we found that lower sample size for both traits reduced power and slightly inflated the FPR; e.g., testing for selection jointly on trait I versus trait II (simulating selection to increase trait I), we found that at $N = 10^4$ for trait I and trait II, the FPR for trait II reached 8.0% ($\pm$1.8%) (Table S2). However, this was not always the case; e.g., for $N_I = 10^5$, $N_{II} = 10^4$, the FPR for trait II was calibrated properly (4.6% $\pm$ 1.4%) (Table S2).

Dashed horizontal lines indicate 5% nominal significance level, and black lines indicate 95% Bonferroni-corrected confidence intervals. Baseline parameters for all simulations (1,000 replicates under each scenario) were our constant-size model with $\varrho = 60\%$, $M = 1,000$, with trait I under positive selection. In (A) and (B) and (D), joint tests are performed on trait I/trait III and trait I/trait II, respectively. (E) Diagonal elements correspond to marginal test for selection.

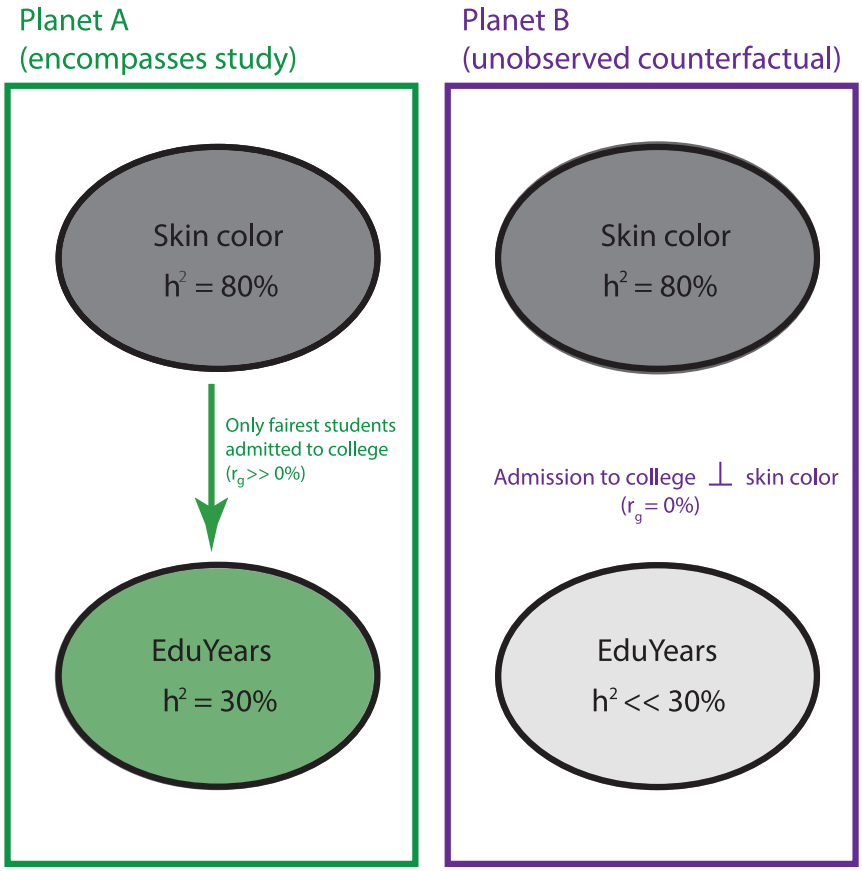

Figure 4. Schematic of "intractable" GxE and induced heritability/genetic correlation

Consider Planet A in which there is a homogeneous environment where a policy states that only the fairest (i.e., lightest-skinned) teenagers are admitted to college. If skin color is heritable, then so will be years of education (EduYears) (with some minor attenuation due to truncation with respect to skin color). Likewise, since genetic variants that modulate skin color will also modulate the likelihood of being admitted to college, there will be a genetic correlation between the two traits. However, on Planet B, such a policy might not exist, and accordingly, the heritability of EduYears will be attenuated (as well as its genetic correlation with skin color, which will go to 0%). Unless a study extends into Planet B, GxE effects are indistinguishable from genetic effects.

Power to assign selection to the causal trait was reduced when that trait's GWAS was underpowered; e.g., 51.6% (±1.6%) to 45.7% (±1.6%) when $N_I$was dropped from 10^5 to 10^4($N_{II} = 10^5$) (Table S2). Interestingly, we found an even bigger drop in power associated with reduced sample size for the correlated trait (trait II); when $N_{II}$was reduced from 10^5 to 10^4($N_I = 10^4$), power to detect selection on trait I dropped from 45.7% (±1.6%) to 27.7% (±1.4%) (Table S2). These results indicate that as long as sample size is reasonably large, estimates are well calibrated; furthermore, by increasing sample size of GWASs for one trait, the joint test is able to leverage that toward improving power to detect selection on other traits that have overlapping genetic architecture.

Empirical analysis of trait evolution in British ancestry

Data analysis and quality control

We analyzed 56 GWASs of metabolic, anthropometric, life history, behavioral, and pigmentation- and immune response-related traits in humans (54 from the UKBB; see Table S3 for details) for signs of polygenic adaptation. We used GWAS summary statistics that were nominally corrected for population structure via either a linear mixed model (LMM)[33] or fixed principal components (PCs, K = 20 PCs)[34] and, in some cases, a family history-based approach (LTFH)[35] to boost power for under-powered UKBB traits, such as type 2 diabetes (Table S3). For most traits, we avoided using meta-analyses and used summary statistics from the UKBB,[34–36] with the exception of schizophrenia and autism spectrum disorder, since these traits are underpowered in the UKBB (note: neither of the latter two traits showed any significant signals of selection or correlated response in our analyses). When available, we chose to use LMM-based summary statistics over those from fixed PCs in order to redress possible environmental effects on stratification on minor PCs, which are not corrected for by the latter approach. See Table S3 for full information on GWAS summary statistic sources/methods for each trait we analyzed. All traits used had at least 25 genome-wide significant (GWS) loci ($p < 5 \times 10^{-8}$) in independent LD blocks.[36] For all of our empirical analyses, we used coalescent trees sampled via Relate for a sample of British ancestry (GBR, $n = 89$) from the 1000 Genomes Project, assuming pre-established estimates of GBR demographic history.[19,37] We specifically tested for selection in the last 2,000 years (i.e., 68.95 generations, assuming a generation time of 29 years). The selection gradient (ω) was estimated with maximum likelihood, with standard errors estimated by block bootstrapping. We first tested traits marginally for polygenic adaptation (Figure 5). We include SNPs by pruning for LD via independent LD blocks, choosing the SNP with the lowest p value in each independent block and excluding blocks that do not have a SNP exceeding this threshold.[36]

Additionally, as a precaution, we analyzed all 56 GWASs for residual stratification (Figure S14A, *cf.* Sohail et al. (2019),[26] Figure 2A), finding no significant residual stratification on PCs 1–20 ($p > 0.005$ for all tests). Furthermore, for comparison, we tested PC-wise stratification in our simulations of uncorrected stratification (Figures S14B and S14C, *cf.* Figure 1B). We found that many of our simulation

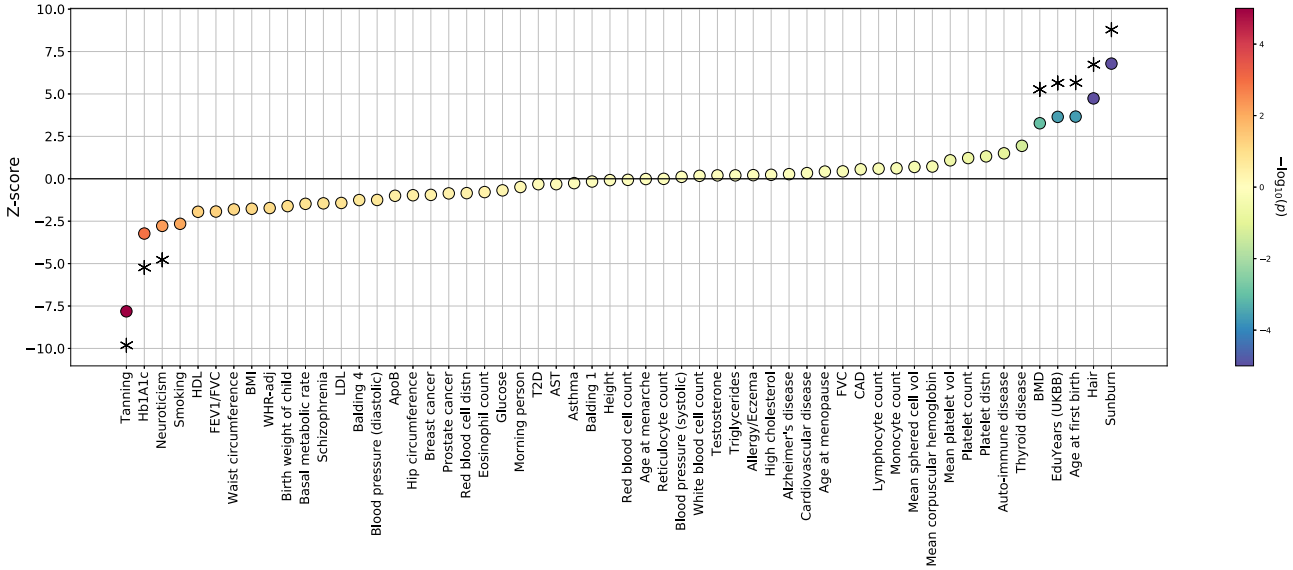

Figure 5. Estimates of the selection gradient on 56 human traits
The selection gradient ($\widehat{\omega}$) was estimated using 1000 Genomes Great British (GBR) individuals and summary statistics from various GWASs (see Table S4 for full results), with standard errors ($\widehat{se}_{\omega}$) estimated via block bootstrap ($Z = \widehat{\omega}/\widehat{se}_{\omega}$). Starred traits indicate significance at FDR = 0.05.

conditions exhibited significant stratification along PCs 1–20; nonetheless, our method was robust to levels of uncorrected stratification in the simulated data, which greatly exceed anything observed in the empirical GWAS (e.g., $\sigma_S \leq 0.1$, Figures 1B and S14B). Thus, we ruled out residual stratification along major PCs as a reasonable source of bias in our results.

Marginal tests for selection

We report our estimates of the selection gradient (Figure 5) normalized by their standard errors, highlighting significant traits (FDR = 0.05) and other traits of interest, with results also presented in Table S4. In the marginal tests with PALM, we found strong signals of selection acting to decrease pigmentation (Figure 5, Table S4). We reported traits with selection gradient p value exceeding a multiple testing-corrected threshold (FDR = 0.05, Benjamini-Hochberg). Tanning showed the strongest signal of directional (in this case, negative) selection among all tested traits ($\omega = -0.357$ [0.046], $p = 5.5 \times 10^{-15}$; standard errors in brackets). Sunburn ($\omega = +0.356$ [0.052], $p = 1.1 \times 10^{-11}$) and hair color ($\omega = +0.128$[0.027], $p = 2.2 \times 10^{-6}$) also showed significant positive selection. Several life history traits also showed significant selection; e.g., age at first birth ($\omega = +0.0546$ [0.0149], $p = 2.5 \times 10^{-4}$) and EduYears ($\omega = +0.389$ [0.0107], $p = 2.6 \times 10^{-4}$). We also found significant selection acting on an anthropometric trait, bone mineral density heel-T Z score (BMD, $\omega = +0.0728$ [0.0222], $p = 1.1 \times 10^{-3}$) and negative selection acting on glycated hemoglobin levels (HbA1c, $\omega = -0.0167$ [0.00518], $p = 1.2 \times 10^{-3}$) as well as neuroticism ($\omega = -0.0706$ [0.0254], $p = 5.5 \times 10^{-3}$).

Several traits of interest have no or inconclusive evidence of directional selection. We found no evidence for any recent directional selection on height ($\omega = -0.00148 \times 10^{-3}$ [± 0.0190], $p = 0.938$). We also find inconclusive evidence for selection on body mass index (BMI, $\omega = -0.0585$ [± 0.0331], $p = 0.077$), in contrast to previous findings that BMI has been under significant selection to decrease.[16]

Joint tests for selection

We analyzed 137 trait pairs (Bonferroni $p_{r_g} < 0.005$ and $|r_g| > 0.2$)[32] by using J-PALM to examine whether marginal signals of selection were due to a correlated response to selection on another trait (Table 3, Table S5). To aid clarity, we introduce the notion of focal versus conditional traits in a joint test. For example, if we estimate the selection gradient of trait 1 and trait 2, (ω_1, ω_2), then $\omega_1$is the estimate for trait 1 (the focal trait), jointly tested estimated with trait 2 (the conditional trait); similarly $\omega_2$is the estimate for trait 2 (the focal trait), jointly tested estimated with trait 1 (the conditional trait). We establish significance of correlated response by using a Wald test on the statistic R, the difference in the joint and marginal selection estimates for a focal trait, where the joint analysis is performed with some other conditional trait (see Testing for correlated response and Appendix A for more details). Selected results are presented in Table 3, and results for the full analysis of all 137 trait pairs are available in Table S5.

We found several significant signals (FDR = 0.05) of correlated response (Table 3, full results in Table S5). For example, although EduYears had strong evidence for selection in the marginal test ($p_{marginal} = 2.6 \times 10^{-4}$), we found after conditioning on sunburn ability ($r_g = 0.24$, $p = 2.3 \times 10^{-4}$)[32] a significant attenuation of this estimate ($p_{joint} = 0.020$, $p_R = 2.6 \times 10^{-6}$). These results suggest that a large part of the signal of selection on EduYears is

Table 3. Selected trait pairs under correlated response in Great British ancestry

Traits		Marginal test		Joint test			
Focal	**Conditional**	***Z***	p_Z	***Z***	p_Z	**R**	**pR**
Hair	Tanning	4.74	2.2×10^{-6}	1.91	0.056	−3.77	1.7×10^{-4}*
EduYears	Sunburn	3.65	2.7×10^{-4}	2.33	0.020	−4.68	2.9×10^{-6}*
Hb1A1c	T2D	−3.23	1.2×10^{-3}	−4.41	1.0×10^{-5}*	−3.17	1.6×10^{-3}*
	BP (diastolic)			−1.95	0.051	2.36	0.019
T2D	Hb1A1c	−0.32	0.75	2.75	6.0×10^{-3}*	4.34	1.5×10^{-5}*
	BP (diastolic)			0.28	0.78	2.10	0.036

Selection on the focal trait is estimated jointly with the conditional trait. We report the Z scores under both the marginal and joint tests, as well as the *R* statistic of the difference in joint versus marginal selection gradient estimates and their p values. Results for all trait pairs are available in Table S5. T2D, type 2 diabetes; HbA1c, glycated hemoglobin; BP, blood pressure. Asterisk (*) denotes significance at FDR = 0.05 ($n = 2 \times 137 = 274$ tests on 137 trait pairs with Bonferroni-significant $p_{r_g} < 0.005/(56 \cdot 55/2)$ and $|r_g > 0.20|$).

most likely due to indirect selection via correlated response rather than direct selection. However, we stress that these results do not provide evidence of direct selection on the conditional trait, here e.g., childhood sunburn occasions (sunburn) (see e.g., Figure 3E).

We also find significant attenuation of selection signals for pigmentation traits in our joint analyses (Table 3). In our joint analysis of hair color and tanning ($r_g = -0.17$, $p = 3.6 \times 10^{-3}$),[32] we found that after conditioning on tanning, there is no residual evidence for direct selection on hair color ($p_{marginal} = 2.2 \times 10^{-6}$; $p_{joint} = 0.056$; $P_R = 1.7 \times 10^{-4}$). (The same caveat above regarding the interpretation of correlated response applies here to tanning ability.)

We identified one case in which the joint analysis uncovers selection acting on a trait that did not show significant selection marginally; we found that type 2 diabetes (T2D), conditioning on HbA1c ($r_g = 0.69$),[38] shows significant selection to increase in prevalence($p_{marginal} = 0.75$; $p_{joint} = 0.0060$; $p_R = 1.5 \times 10^{-5}$; see Table 3). Estimates of negative selection on HbA1c are also enhanced after accounting for T2D ($p_{marginal} = 1.2 \times 10^{-3}$; $P_{joint} = 1.0 \times 10^{-5}$; $P_R = 0.0016$; see Table 3). This "cancelling-out" effect of opposing selection on T2D and HbA1c, two traits with strong (but not perfect) positive genetic correlation, is the second-strongest signal of correlated response in our joint analyses. We confirmed that the separability of these two phenotypes is not due to phenotype mis-specification; we confirmed T2D status by doctor's diagnosis strictly after 30 years of age in order to avoid the possibility that T1D individuals mistakenly self-report as T2D diagnosed.[35] The summary statistics also show close replication of T2D associations in a case-control study, thus further suggesting T2D is the predominant signal.[35]

We also illustrate our estimates of selection coefficients for ascertained T2D/HbA1c SNPs, found independently of one another, and their fit to our inferred model of antagonistic selection on T2D/HbA1c (Figure 6A). In general, T2D-increasing and/or HbA1c-decreasing SNPs are under positive selection, and vice versa. However, a subset of HbA1c-increasing SNPs show extremely strong signs of positive selection ($s > 0.03$); these SNPs tend to have visibly higher positive effects on T2D than other SNPs with comparable HbA1c effect. In a joint analysis of HbA1c and diastolic blood pressure (as a proxy for hypertension), our estimate of direct selection on HbA1c was significantly attenuated at a nominal level($p = 0.019$, Table 3), although it did not meet our FDR cutoff. We also did a joint analysis of T2D and diastolic blood pressure, finding a significant boost in the estimate of direct selection on T2D ($p = 0.036$, Table 3), although it did not meet our FDR cutoff.

Lastly, we tested our set of *R* statistics among the pairs of genetically correlated traits for enrichment in the tail over the null (Figure 6B). At the nominal 5% FPR level, we found significant (2.6-fold) enrichment for correlated response acting on these traits ($p = 1.5 \times 10^{-7}$, one-sided binomial test), suggesting that many additional traits in this analysis have evolved under indirect selection due to correlated response.

Discussion

We have presented a method, PALM, for estimating the directional selection gradient acting on a polygenic trait. Our method works by estimating likelihood functions for the selection coefficients of a set of GWAS SNPs and then aggregating these functions along with GWAS-estimated SNP effects to find the likelihood of the selection gradient. Through simulations, we showed that PALM offers improved power over current methods across a range of selection gradients($\omega = 0.025 - 0.10$) and polygenicities ($M = 10^2 - 10^4$) and is the first method to our knowledge that can estimate ω from nucleotide data. We showed that even for these extremely polygenic traits where power is attenuated, as methods to infer the tree become more accurate and scalable, and as GWAS sample sizes increase, so too will power to detect selection on extreme polygenic traits. We conducted robustness checks and showed that

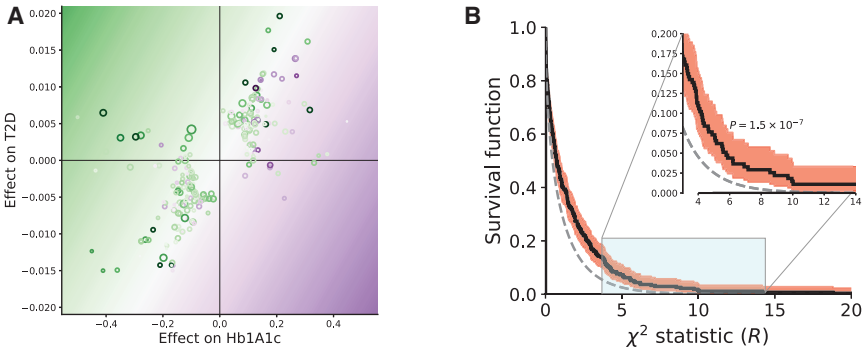

Figure 6. Correlated response in real traits

(A) Expanded view of antagonistic selection on glycated hemoglobin (HbA1c) versus type 2 diabetes (T2D). We estimate individual SNP selection coefficients by taking the maximum-likelihood estimate $\hat{s}$ for each SNP. We plot this value against the joint SNP effect estimates for HbA1c and T2D. Colored lines represent isocontours of $s(\beta) = \beta_{HbA1c}\hat{\omega}_{HbA1c} + \beta_{T2D}\hat{\omega}_{T2D}$, the estimate of the Lande transformation from SNP effects to selection coefficients, where $\hat{\omega}$ is inferred jointly for the two traits (Table 3). The purple-green color gradient illustrates expected selection coefficients under $\hat{\omega}$ (background) versus individual SNP selection coefficient estimates (rings). Ring diameter is proportional to SNP selection log-likelihood ratio.

(B) Enrichment of correlated response in analysis of genetically correlated traits. Enrichment in the tails of the distribution of our test statistic for correlated response R($p = 1.5 \times 10^{-7}$, binomial test), which had 2.6-fold enrichment at the nominal 5% level. We assessed $n = 2 \times 137 = 274$ estimates of correlated response on 137 trait pairs with Bonferroni-significant $p_{r_g} < 0.005/\frac{56}{2}$ and $|r_g > 0.20|$. Red area indicates pointwise 95% CI of the survival curve.

PALM is robust to typical sources of uncertainty and bias in GWAS summary statistics (e.g., sampling variation, ascertainment bias/Winner's Curse), allelic heterogeneity, purifying selection, and underpowered GWASs.

We also introduced a method, J-PALM, to jointly estimate the selection gradient on multiple traits in order to control for pleiotropy. We showed that, across a wide range of polygenic architectures ($M = 10^2 - 10^4$, $\varrho = 40\% - 100\%$), J-PALM can reliably detect and assign selection to the causal trait when it is considered in the analysis. Furthermore, PALM can be used to uncover genetically correlated traits under antagonistic selection where the marginal approach (e.g., PALM) is underpowered. We considered several additional sources of bias unique to multi-trait analyses (i.e., uneven GWAS sample sizes, correlation in trait environmental noise) and found J-PALM robust to these as well.

We note several areas in which the study of polygenic adaptation can be advanced. Our operative model of polygenic adaptation is based on the Lande approximation, which over long time-courses will overestimate the efficiency of adaptation under stabilizing selection with a shift in the optimum.[12,39] A model that incorporates these dynamics will potentially be better suited to detecting polygenic adaptation over longer time-courses, such as analyses of ancient DNA samples. Furthermore, under stabilizing selection, more SNP heritability is expected to be sequestered to low-frequency alleles and so common SNPs are expected to change less under adaptation than in our simulation model.[5,12]

Advances might also be made through more nuanced models that make fuller use of GWAS summary statistics and LD among GWAS markers. We showed our thresholding and pruning scheme for selecting sites did not substantially decrease our method's power. Pre-existing methods for fine-mapping or ascertaining pleiotropic loci might increase power even further.[40] It is also possible that for traits with extremely high polygenicity and/or low heritability, it will be necessary to utilize summary statistics that are sub-significant and account for uncertainty in the location of the causal site. While in this paper we explored a thresholding and pruning scheme, which previous work and our own simulations show to be robust for stringent thresholding,[24,25] we have not established how results would differ for an LD clumping approach or how misspecification of the LD reference panel (versus the GWAS and/or population genetic cohort) affects our results.

We showed that PALM is substantially less prone to bias due to uncorrected GWAS stratification than comparable methods, such as tSDS. However, we stress that PALM can nonetheless be biased under sufficiently strong uncorrected stratification. While we illustrated that residual stratification along major PCs is not present in any of the empirical GWASs we analyzed, the extent to which residual stratification along less significant axes of variation could cause bias is yet unclear. Forms of stratification that we did not explore, such as GxE interactions, may be more difficult to account for via standard kinship-based approaches; however, new methods have recently arisen to this particular end.[41] Our simulations based on empirical data suggest that cryptic population structure (e.g., within white British ancestry) does not confound PALM unless overlaid with fairly extreme stratification. Future work should consider how more pronounced structure or assortative mating distorts the ARG and/or GWAS effect size estimates, which may have downstream effects on estimates of polygenic adaptation. We also note that while PALM is robust to minor levels of mutational bias, the test can be confounded by more extreme mutational bias, i.e., attributing mutation-driven changes in the trait distribution to adaptation.

Another limitation of our model is the interpretation of the estimates of the selection gradient and correlated response. We showed through simulations that when a genetically correlated trait with causal fitness effect is excluded from the analysis, estimates of direct selection have no causal interpretation. To address this, we

introduced the notion of an effective selection gradient, which depends on which traits are modeled together. Estimates of the effective selection gradient allow us to determine whether a focal trait has evolved under correlated response another trait; however, this does not have the causal interpretation that the focal trait is under correlation response to a particular conditional trait. Furthermore, even in a simple model with a single trait considered and under selection, nonzero selection gradients may reflect environmental changes rather than a change in the trait's fitness optimum.

Applying PALM to study evolution of 56 human traits in British ancestry, we found eight traits under significant directional selection, recovering several previously reported targets, such as pigmentation traits, educational attainment, and glycated hemoglobin (HbA1c), in agreement with previous findings of selection on these traits in Europe.[15,16,42] We also report several novel targets of directional selection, such as increased bone mineral density and decreased neuroticism. Despite historical claims of selection to increase height in Europe,[22] we found no evidence for selection to increase height, consistent with recent analyses that showed that signals of directional selection on height have been drastically inflated by uncorrected population structure in GWAS summary statistics.[25,26]

We applied our joint test J-PALM to study 137 pairs of genetically correlated traits for signatures of correlated response. We found a highly significant enrichment of correlated response acting on these traits. Particularly, we found significant correlated response acting on pigmentation and life history traits (hair color, educational attainment). We showed that signal of selection on traits such as hair color and educational attainment, which have been widely reported to date,[15,16,42,43] is due in significant part to correlated response to selection on other traits versus direct selection acting on these traits.

One proposed theory for the diversification and increase of blonde hair color in Europe is sexual selection.[44,45] However, our results do not support this, as we show that evidence for selection on hair color is attributable mostly to correlated response, beyond which there is little evidence for direct selection on this trait. This echoes previous analysis showing selection at individual hair color loci may be indirect, via their pleiotropic effects (e.g., blonde hair gene *KITLG* responding to selection for tolerance to climate and UV radiation[46]), and conflicts with arguments that hair color has been under direct sexual selection.

Our marginal test for selection detects significant selection for increased sunburn ability. We caution that "childhood sunburn occasions" is a survey question and is most likely affected by many exogeneous factors beyond skin pigmentation (e.g., opportunity to visit the beach or use sunscreen). Furthermore, pigmentation traits are exceptionally prone to stratification. While we find no significant uncorrected stratification in the empirical GWAS summary statistics (for pigmentation or any other traits) and our simulations suggest that population structure (e.g., that between British and Italian individuals in Europe) absent a stratification effect does not incur a bias in PALM's inference of selection, we cannot exclude the possibility that residual stratification on less significant axes of genetic variation could play some role.

In our marginal test for selection, we detected significant selection to increase educational attainment, consistent with some previous work.[16] However, in a joint test with sunburn (i.e., "childhood sunburn occasions," the number of times the individual was sunburned as a child), strong signals of selection to increase educational attainment were significantly obviated. We conclude that signals of selection on educational attainment are driven significantly by correlated response. We propose that GxE interactions may be driving these signals of correlated response. Lewontin (1970), responding to Jensen (1968), pointed out that then-current estimates of intelligence quotient (IQ) heritability were inflated by GxE.[47,48] Indeed, in modern-day GWASs, we see that educational attainment polygenic scores in the UKBB are only 50% as predictive in adoptees as in non-adoptees, indicating a significant role of GxE in the expression of educational attainment, as well as estimates of its heritability and genetic correlations.[49] The role of GxE or indirect genetic effects has been further illustrated by the discrepancy of sibling-based versus standard GWAS estimates of SNP effects on educational attainment.[50] Hence, genetic correlation of sunburn and educational attainment may be overestimated (e.g., $\widehat{r_g} = 0.24$ with UKBB GWAS[32]). We do not have data to elucidate the mechanism of this proposed GxE interaction but hypothesize that educational opportunities could be affected by skin pigmentation via differential opportunities in education, such as racial discrimination. Additionally, it is notable that "childhood sunburn occasions" may not directly reflect skin color, as it may be related to other factors (e.g., number of visits to the beach), which in turn may be modulated by similar GxE mechanisms. Even in the absence of GxE, we stress that our results are not interpretable as evidence of direct selection on sunburn ability—let alone skin pigmentation—following from our simulation study. Also, the inferred correlation between the traits and/or the signals of selection could be affected by uncorrected GWAS stratification.[25,26] Lastly, it is worth noting that heritability and polygenic score-based phenotype prediction of educational attainment are largely driven by indirect effects (i.e., genetic nurture),[50,51] thus further complicating any interpretation of how effects on education attainment relate to effects on traits such as pigmentation. In conclusion, we caution that the observed genetic correlation may not arise from shared direct effects but rather some combination of GxE and/or indirect effects.

We found one case of significant antagonistic selection: T2D shows significant selection to increase, but this signal was initially occluded by the positive genetic correlation of T2D with negatively selected glycated hemoglobin (HbA1c). Our joint analysis with J-PALM disentangles this antagonism between T2D and HbA1c, revealing latent

adaptation of T2D. T2D is a complex disease with a complex etiology, involving obesity and various metabolic risk factors. Selection may have favored some of these factors under previous environmental conditions where both obesity and diets rich in simple sugars were uncommon (also known as the thrifty gene hypothesis).[52] HbA1c is a biomarker commonly used to not only diagnose pre-diabetes/diabetes but also to monitor chronic hyperglycemia as a risk factor for vascular damage.[53] T2D and HbA1c are strongly, although imperfectly, genetically correlated (r_g = 69%). Although this may seem peculiar because HbA1c is a diagnostic criterion for T2D, we speculate the distinction between these phenotypes could be driven by variation in HbA1c above and/or below diagnostic thresholds or variation of other molecular traits (e.g., fasting glucose) that are also used as diagnostic criteria. HbA1c is also associated with hypertension and other cardiovascular disease independently of T2D incidence.[38] It is therefore possible that selection might have favored some of the traits underlying increased T2D risk but acted against some of the more specific negative effects of T2D, which now are measured by HbA1c.[38,53,54] These results provide evidence in support of the thrifty gene hypothesis.[54]

Appendix A

Importance sampling estimation of the likelihood function of selection

Our likelihood model builds heavily on our previous work, which developed importance sampling approaches to estimating the likelihood function of the selection coefficient acting on a SNP, $L^{SNP}(s)$.[1] Here, we briefly explain the importance sampling method used to estimate $L(\omega)$, the likelihood of the multivariate selection gradient:

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}^{\top}\omega\right), \qquad \text{(Equation A1)}$$

where $\beta_{(i)}$ is the vector of trait effects for SNP i. In the following, we omit the subscript i for brevity. We can model the relationship between SNP s and the haplotype data D from a window around the SNP via the latent ancestral recombination graph (ARG) G,

$$L^{SNP}(s) = E_p[P(D|G,s)] = E_q\left[P(D|G,s)\frac{p(G|s)}{q(G)}\right] \qquad \text{(Equation A2)}$$

for any appropriate choice of q such that $p(s) > 0 \Rightarrow q(G) > 0$, which generally will hold in our case. Thus, we can approximate the SNP likelihood function as

$$\widehat{L}^{SNP}(s) := \frac{1}{m}\sum_{l=1}^{m} P\left(D\middle|G^{(l)},s\right)\frac{p\left(G^{(l)}\middle|s\right)}{q\left(G^{(l)}\right)} \to L_k^{SNP}(s), \qquad \text{(Equation A3)}$$

where the convergence is almost surely as $m \to \infty$. We are interested in the particular choice of $q(G) = p(G \mid D, s = 0)$, the posterior under selective neutrality, because programs such as ARGweaver[2] and Relate[3] can be used to approximately sample the posterior ARG or aspects of it (e.g., a local tree). We showed previously that the approximation

$$\widehat{LR}^{SNP}(s) = \frac{1}{m}\sum_{l=1}^{m}\frac{p\left(G_i^{(l)}\middle|s\right)}{p\left(G_i^{(l)}\middle|s=0\right)} \qquad \text{(Equation A4)}$$

is a tractable and accurate estimate of the likelihood ratio of s, where G_i denotes the local tree at SNP i, extracted from the ARG G. Here, we introduce and use a slightly different estimator,

$$\widehat{LR}^{SNP}(s) = \frac{\sum_{l=1}^{m}\frac{p\left(G_i^{(l)}\middle|s\right)}{\pi\left(G_i^{(l)}\right)}}{\sum_{l=1}^{m}\frac{p\left(G_i^{(l)}\middle|s=0\right)}{\pi\left(G_i^{(l)}\right)}}, \qquad \text{(Equation A5)}$$

where $\pi(\cdot)$ is a neutral prior on coalescence trees. While $p(\cdot)$ is calculated via the structured coalescent, with lineages subtending the same allele with frequency $X(t)$ coalescing at rate $\lambda(t) = N(0)/[N(t)X(t)]$, the prior $\pi(\cdot)$ is calculated via the unstructured coalescent with rate $\lambda(t) = N(0)/N(t)$. Note that we do not explicitly model population structure (e.g., gene flow).

We also note that we have made several additional modifications to the importance sampling approximation of the likelihood ratio: first, we assume that the allele frequency trajectory is a deterministic, logistic function of time, when previously we modeled stochasticity in the allele frequency trajectory (see the next section for more details). Because we focus on applying our method to detecting adaptation in the recent past, this approximation is appropriate when drift has had little opportunity to distort allele frequencies.

Second, we make a functional approximation to $\log\widehat{LR}^{SNP}(s)$. We do a grid search for the optimal value of s^*, and then we fit a quadratic function to points $\{(s, log\ \widehat{LR}^{SNP}(s)) : |s - s^*| < \delta\}$. Optimizing $\log\widehat{LR}(\omega)$ then becomes a simple process of solving a linear system of equations:

$$\log\widehat{LR}(\omega) = \sum_i\left(a_i\left(\beta_{(i)}^{\top}\omega\right)^2 + b_i\left(\beta_{(i)}^{\top}\omega\right) + c_i\right), \qquad \text{(Equation A6)}$$

where (a_i, b_i, c_i) are the fitting coefficients of the quadratic approximation for SNP i in descending order of degree. Thus,

$$\widehat{\omega} = \left[2\sum_i a_i\beta_{(i)}\beta_{(i)}^{\top}\right]^{-1}\left(\sum_i b_i\beta_{(i)}\right). \qquad \text{(Equation A7)}$$

This approximation has two benefits: (1) solving for the selection gradient estimate is extremely simple and fast and (2) it makes it feasible to calculate standard errors with resampling approaches.

Accounting for multiple SNPs in LD

In our analyses we assume independence of local LD blocks (see e.g., Berisa and Pickrell, 2016[36]). Generally, we choose to ascertain a single SNP for each LD block and include its SNP likelihood in the product (Equation A1). However, in joint analyses, it may be necessary to ascertain multiple SNPs per LD block, each corresponding to a GWAS hit for a different trait.

Let $B(i)$ denote the set of ascertained SNPs in the same LD block as i. If only one SNP from each LD block is included, then $B(i) = 1$for each ascertained SNP i. If multiple SNPs from the same LD block are included, we exponentiate each of these SNPs' likelihoods by a factor $1/|B_i|$:

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}^{\top}\omega\right)^{\frac{1}{|B_i|}}. \qquad \text{(Equation A8)}$$

This can be considered a conservative method for dealing with SNPs in LD. For example, let A be our set of ascertained SNPs. If two nearby SNPs i_1, i_2 are in perfect LD ($r^2 = 1$), then we expect $L_{i_1}^{SNP}(s) = L_{i_2}^{SNP}(s)$ and $\beta_{(i_1)} = \beta_{(i_2)}$. Suppose all other SNPs in A are independent (i.e., ascertained from distinct LD blocks). Then the exponentiation factor recovers the original likelihood

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}^{\top}\omega\right)^{\frac{1}{|B_i|}}$$

$$K(\omega)\cdot \prod_{i\in S: i\neq i_1, i_2} L_i^{SNP}\left(\beta_{(i)}^{\top}\omega\right), \qquad \text{(Equation A9)}$$

where $K(\omega) = L_{i_1}^{SNP}(\beta_{(i_1)}^{\top}\omega) = L_{i_2}^{SNP}(\beta_{(i_2)}^{\top}\omega)$. In the other limiting case $r^2 = 0$, this correction factor is conservative, as it discounts the contribution of i_1, i_2 to the log likelihood by a factor of $^1/_2$.

Selection gradient and correlated selection standard errors

We use a block bootstrap approach to calculating the standard errors of $\widehat{\omega}$. Specifically, we identify LD blocks and bootstrap loci ascertained in distinct blocks. Given the standard errors, we assess significance by using a Wald test on the Z statistic $\widehat{\omega}/\widehat{se_\omega}$.

We also compute a statistic we call R to assess whether a trait j has evolved under correlated response to selection on some disjoint set of traits T. To do this, we can estimate selection gradients for two sets of traits, T and $T\cup j$, and calculate

$$R = \omega^{(T\cup\{j\})} - \omega^{(\{j\})}, \qquad \text{(Equation A10)}$$

where $\omega^{(U)}$is the selection gradient of the trait estimated with respect to a set of traits U, calculate $\widehat{se_R}$ through block bootstrap, and assess significant via a Wald test on $R/\widehat{se_R}$.

Coalescent likelihood models: Relate prior

The prior $\pi(T)$ is the standard coalescent with changing effective population size. First, let U be the vector of $n-1$ coalescent times of T, ordered most to least recently. Due to exchangeability of lineages, the density only depends on T via these coalescent times U. Specifically,

$$\pi(T) = \prod_{i=1}^{n-1} p(U_i = u_i | U_{i-1} = u_{i-1}) \qquad \text{(Equation A11)}$$

$$p(U_i = u_i | U_{i-1} = u_{i-1}) = \frac{n-i+1}{2}\cdot N(0)\Big/N(u_i)\cdot exp\left(-\frac{n-i+1}{2}(\Lambda(u_i) - \Lambda(u_{i-1}))\right) \qquad \text{(Equation A12)}$$

$$\Lambda(u) = \int_0^u N(0)\Big/N(t)\cdot dt. \qquad \text{(Equation A13)}$$

We assume that $N(t)$ is piecewise constant and can be expressed with $\tau = (\tau_0, \tau_1, \ldots)$ and $N = (N_0, N_1, \ldots)$ such as the required models for ARGweaver and Relate; hence, finding $\Lambda(u)$ is a simple sum over integrals defined over constant functions:

$$\Lambda_i = \sum_{k=1}^{b(u_i)} N_0\tau_k\Big/N_k + N_0\left(u_i - \tau_{b(u_i)}\right)\Big/N_{b(u_i)}, \qquad \text{(Equation A14)}$$

where $b(u) := max\{k\in(0,1,2\ldots) : u > \tau_k\}$.

Coalescent selection likelihood under deterministic model

Unlike in our previous work,[1] in which we treated the allele frequency as a latent random variable, here, we use a deterministic approximation of the allele frequency trajectory. Under the standard "hard sweep" model, an appropriate approximation would be $X(t \mid s) = (1 + (1 - x_0)/x_0\cdot e^{st})^{-1}$. Technically, if we want to express the trajectory conditional on the present-day derived allele frequency (DAF) x_0, it would be more appropriate to use a closer approximation of the backward Wright-Fisher diffusion with selection (see e.g., Zeng et al., 2018[4]). However, since we are mostly interested in modeling the recent past for common alleles ascertained in a GWAS (usually DAF > 1%), this approximation is appropriate, especially in populations of large recent N_e such as humans, where drift is negligible on short timescales.

We assume a pulse of selection over some time interval (a, b), outside of which the allele is effectively neutral (and, we assume, at constant frequency):

$$X(t,s,x_0) = x_0, t < a \quad \text{(Equation A15)}$$

$$= \left(1 + (1-x_0)/x_0 \cdot e^{s(t-a)}\right)^{-1}, a \le t < b \quad \text{(Equation A16)}$$

$$= \left(1 + (1-x_0)/x_0 \cdot e^{sb}\right)^{-1}, t > b. \quad \text{(Equation A17)}$$

To calculate $p(T \mid s)$, we split the tree into two subtrees (imagine "deleting" the branch on which the mutant allele arose). Note that we implicitly assume the site is bi-allelic, such as under the infinite sites assumption. Let us label these alleles $A1$and $A2$; these labels must be consistent with the polarization of the GWAS summary statistics; we assume that those are polarized w.r.t the $A1$ allele. Within each of these subtrees, we find the coalescent times $U^{A1}$and U^{A2}. Then

$$p\left(T|s\right) = \prod_{i=1}^{n_1-1} p\left(U_{n-i} = u_i^{A1} \middle| U_{n-i+1} = u_{i-1}^{A1}, s, x_0\right) \times$$

$$\times \prod_{i=1}^{n_2-1} p\left(U_{n-i} = u_i^{A2} \middle| U_{n-i+1} = u_{i-1}^{A2}, -s, 1-x_0\right) \quad \text{(Equation A18)}$$

$$p(U_{k-1} = t | U_k = t', s, f) = \frac{k}{2} \cdot \frac{N(0)}{N(t)X(t)} \cdot exp\left(-\frac{k}{2}\left(\Lambda(t,s,f) - \Lambda(t',s,f)\right)\right) \quad \text{(A19)}$$

$$\Lambda(t,s,f) = \int_0^t N(0) \Big/ \left[N(\tau)X(\tau,s,f)\right] \cdot d\tau, \quad \text{(Equation A20)}$$

where $U^{A1}$and U^{A2} are measured in units of $2N(0)$ generations.

Data and code availability

Open-source code and documentation for PALM/J-PALM is available at https://www.github.com/35ajstern/palm. Formatted summary statistics/metadata and 1000 Genomes GBR selection likelihoods for ascertained SNPs are available for download on DataDryad: https://datadryad.org/stash/landing/show?id=doi%3A10.6078%2FD11M62.

Supplemental information

Supplemental Information can be found online at https://doi.org/10.1016/j.ajhg.2020.12.005.

Acknowledgments

We thank our anonymous reviewers, Doc Edge, Sasha Gusev, and Arbel Harpak for insightful comments and suggestions that have strengthened the manuscript and Jeremy Berg, Jennifer Blanc, Yun Deng, Arun Durvasula, Daniel Geschwind, Iain Mathieson, Priya Moorjani, Maxim Rabinovich, Monty Slatkin, and Lawrence Uricchio for helpful discussions. R.N. was supported by an NIH grant (R01GM138634). N.A.Z. was funded by NIH grants (R56MD013312, R01HG006399, R01CA227237, R01CA227466, R01ES029929, R01MH122688, and U01HG009080), the Chan Zuckerberg Initiative, and a DoD grant (W81XWH-16-2-0018).

Declaration of interests

The authors declare no competing interests.

Received: May 21, 2020
Accepted: December 7, 2020
Published: January 12, 2021

Web resources

1000 Genomes Phase 3 data, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/
Alzheimer disease GWAS summary statistics, https://ctg.cncr.nl/software/summary_statistics
BOLT-LMM summary statistics, https://alkesgroup.broadinstitute.org/UKBB/UKBB_409K/
GWAS Atlas, https://atlas.ctglab.nl/
LT-FH summary statistics, https://alkesgroup.broadinstitute.org/UKBB/LTFH/sumstats/
Neale Lab GWAS Round 2, https://docs.google.com/spreadsheets/d/1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU/edit?ts=5b5f17db#gid=227859291
PGC summary statistics, https://www.med.unc.edu/pgc/download-results/
Relate software, https://myersgroup.github.io/relate/
SDS scripts, https://github.com/yairf/SDS

References

1. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H.K., Bulik-Sullivan, B.K., Pollack, S.J., de Candia, T.R., Lee, S.H., Wray, N.R., Kendler, K.S., et al.; Schizophrenia Working Group of Psychiatric Genomics Consortium (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. *47*, 1385–1392.
2. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. *99*, 139–153.
3. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.
4. Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746–753.
5. Simons, Y.B., Bullaughey, K., Hudson, R.R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. *16*, e2002985.
6. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. Am. J. Hum. Genet. *105*, 456–476.

7. Schoech, A.P., Jordan, D.M., Loh, P.-R., Gazal, S., O'Connor, L.J., Balick, D.J., Palamara, P.F., Finucane, H.K., Sunyaev, S.R., and Price, A.L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nat. Commun. *10*, 790.
8. Sanjak, J.S., Sidorenko, J., Robinson, M.R., Thornton, K.R., and Visscher, P.M. (2018). Evidence of directional and stabilizing selection in contemporary humans. Proc. Natl. Acad. Sci. USA *115*, 151–156.
9. Walsh, B., and Lynch, M. (2018). Evolution and Selection of Quantitative Traits (Oxford University Press).
10. Stern, A.J., and Nielsen, R. (2019). Detecting Natural Selection. Handbook of Statistical Genomics: Two Volume Set, 397–340.
11. Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. *20*, R208–R215.
12. Hayward, L.K., and Sella, G. (2019). Polygenic adaptation after a sudden change in environment. bioRxiv. https://doi.org/10.1101/792952.
13. Berg, J.J., and Coop, G. (2014). A population genetic signal of polygenic adaptation. PLoS Genet. *10*, e1004412.
14. Racimo, F., Berg, J.J., and Pickrell, J.K. (2018). Detecting Polygenic Adaptation in Admixture Graphs. Genetics *208*, 1565–1584.
15. Field, Y., Boyle, E.A., Telis, N., Gao, Z., Gaulton, K.J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M.I., and Pritchard, J.K. (2016). Detection of human adaptation during the past 2000 years. Science *354*, 760–764.
16. Uricchio, L.H., Kitano, H.C., Gusev, A., and Zaitlen, N.A. (2019). An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evol Lett *3*, 69–79.
17. Griffiths, R.C., and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. *3*, 479–502.
18. Rasmussen, M.D., Hubisz, M.J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLoS Genet. *10*, e1004342.
19. Speidel, L., Forest, M., Shi, S., and Myers, S.R. (2019). A method for genome-wide genealogy estimation for thousands of samples. Nat. Genet. *51*, 1321–1329.
20. Edge, M.D., and Coop, G. (2019). Reconstructing the History of Polygenic Scores Using Coalescent Trees. Genetics *211*, 235–262.
21. Stern, A.J., Wilton, P.R., and Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. *15*, e1008384.
22. Turchin, M.C., Chiang, C.W.K., Palmer, C.D., Sankararaman, S., Reich, D., Hirschhorn, J.N.; and Genetic Investigation of ANthropometric Traits (GIANT) Consortium (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. *44*, 1015–1019.
23. Robinson, M.R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J.E., Vinkhuyzen, A., Berndt, S.I., Gustafsson, S., et al. (2015). Population genetic differentiation of height and body mass index across Europe. Nat. Genet. *47*, 1357–1362.
24. Berg, J.J., Zhang, X., and Coop, G. (2017). Polygenic adaptation has impacted multiple anthropometric traits. bioRxiv. https://doi.org/10.1101/167551.
25. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, 8.
26. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife *8*, 8.
27. Lande, R. (1975). The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet. Res. *26*, 221–235.
28. Chevin, L.-M., Billiard, S., and Hospital, F. (2008). Hitchhiking both ways: effect of two interfering selective sweeps on linked neutral variation. Genetics *180*, 301–316.
29. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015a). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
30. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015b). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
31. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol. Biol. Evol. *36*, 632–637.
32. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., de Leeuw, C., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. Nat. Genet. *51*, 1339–1348.
33. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
34. Churchhouse, C., Neale, B.M., Abbott, L., Anttila, V., Aragam, K., Baumann, A., Bloom, J., Bryant, S., Churchhouse, C., Cole, J., et al. (2017). Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK biobank (Neale Lab).
35. Hujoel, M.L.A., Gazal, S., Loh, P.-R., Patterson, N., and Price, A.L. (2019). Combining case-control status and family history of disease increases association power. Nat. Genet. *52*, 541–547.
36. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.
37. Siva, N. (2008). 1000 Genomes project. Nat. Biotechnol. *26*, 256.
38. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., and Mars, N.J. (2019). Genetics of 38 blood and urine biomarkers in the UK Biobank. bioRxiv. https://doi.org/10.1101/660506.
39. Thornton, K.R. (2019). Polygenic Adaptation to an Environmental Shift: Temporal Dynamics of Variation Under Gaussian Stabilizing Selection and Additive Effects on a Single Trait. Genetics *213*, 1513–1530.
40. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
41. Dahl, A., Nguyen, K., Cai, N., Gandal, M.J., Flint, J., and Zaitlen, N. (2020). A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits. Am. J. Hum. Genet. *106*, 71–91.

42. Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M., Unterländer, M., Hollfelder, N., Potekhina, I.D., Schier, W., Thomas, M.G., and Burger, J. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc. Natl. Acad. Sci. USA *111*, 4832–4837.
43. Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet. *3*, e90.
44. Cavalli-Sforza, L.L., Cavalli-Sforza, L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes (Princeton University Press).
45. Frost, P. (2006). European hair and eye color: A case of frequency-dependent sexual selection? Evol. Hum. Behav. *27*, 85–103.
46. Yang, Z., Shi, H., Ma, P., Zhao, S., Kong, Q., Bian, T., Gong, C., Zhao, Q., Liu, Y., Qi, X., et al. (2018). Darwinian Positive Selection on the Pleiotropic Effects of KITLG Explain Skin Pigmentation and Winter Temperature Adaptation in Eurasians. Mol. Biol. Evol. *35*, 2272–2283.
47. Jensen, A. (1969). How much can we boost IQ and scholastic achievement. Harv. Educ. Rev. *39*, 1–123.
48. Lewontin, R.C. (1970). Race and Intelligence. Bull. At. Sci. *26*, 2–8.
49. Cheesman, R., Hunjan, A., Coleman, J.R.I., Ahmadzadeh, Y., Plomin, R., McAdams, T.A., Eley, T.C., and Breen, G. (2020). Comparison of Adopted and Nonadopted Individuals Reveals Gene-Environment Interplay for Education in the UK Biobank. Psychol. Sci. *31*, 582–591.
50. Young, A.I., Frigge, M.L., Gudbjartsson, D.F., Thorleifsson, G., Bjornsdottir, G., Sulem, P., Masson, G., Thorsteinsdottir, U., Stefansson, K., and Kong, A. (2018). Relatedness disequilibrium regression estimates heritability without environmental bias. Nat. Genet. *50*, 1304–1310.
51. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. eLife *9*, 9.
52. Neel, J.V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am. J. Hum. Genet. *14*, 353–362.
53. Lyons, T.J., and Basu, A. (2012). Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers. Transl. Res. *159*, 303–312.
54. Bower, J.K., Appel, L.J., Matsushita, K., Young, J.H., Alonso, A., Brancati, F.L., and Selvin, E. (2012). Glycated hemoglobin and risk of hypertension in the atherosclerosis risk in communities study. Diabetes Care *35*, 1031–1037.