Reconstructing the History of Polygenic Scores Using Coalescent Trees

Michael D. Edge[1] and Graham Coop[1]

Center for Population Biology, Department of Evolution and Ecology, University of California, Davis, California 95616

ORCID IDs: 0000-0001-8773-2906 (M.D.E.); 0000-0001-8431-0302 (G.C.)

ABSTRACT Genome-wide association studies (GWAS) have revealed that many traits are highly polygenic, in that their within-population variance is governed, in part, by small-effect variants at many genetic loci. Standard population-genetic methods for inferring evolutionary history are ill-suited for polygenic traits: when there are many variants of small effect, signatures of natural selection are spread across the genome and are subtle at any one locus. In the last several years, various methods have emerged for detecting the action of natural selection on polygenic scores, sums of genotypes weighted by GWAS effect sizes. However, most existing methods do not reveal the timing or strength of selection. Here, we present a set of methods for estimating the historical time course of a population-mean polygenic score using local coalescent trees at GWAS loci. These time courses are estimated by using coalescent theory to relate the branch lengths of trees to allele-frequency change. The resulting time course can be tested for evidence of natural selection. We present theory and simulations supporting our procedures, as well as estimated time courses of polygenic scores for human height. Because of its grounding in coalescent theory, the framework presented here can be extended to a variety of demographic scenarios, and its usefulness will increase as both GWAS and ancestral-recombination-graph inference continue to progress.

KEYWORDS polygenic score; coalescent; selection

SOME of the most compelling examples of phenotypic evolution come from time courses that reveal the pace of evolution, either through observations across generations (Cook *et al.* 1986; Grant and Grant 2002) or through changes in the fossil record (Gingerich 1983; MacFadden 2005; Bell *et al.* 2006). For many traits and species, it can be difficult to ascertain whether these changes reflect genetic change. For example, we have fairly detailed knowledge of human height through time, but some changes in height are likely driven by environmental and dietary changes (Stulp and Barrett 2016). Thanks to ancient DNA, we can now sometimes obtain a partial picture of long-term genetic changes involving relatively simple traits like pigmentation (Ludwig *et al.* 2009) or more complex traits (Mathieson *et al.* 2015). However, we are usually not fortunate enough to have access to

genotype data from across time, and even when ancient DNA are available, the resulting time courses will necessarily be incomplete.

One alternative to direct measurement of phenotypes through time is to reconstruct the history of a phenotype using contemporary genetic data. Positive selection on simple genetic traits drives large allele-frequency changes at the causal loci and linked neutral alleles (Smith and Haigh 1974). There are many procedures for detecting this kind of selection on individual alleles and for dating and modeling their spread through populations (Tajima 1989; Fay and Wu 2000; Sabeti *et al.* 2002; Voight *et al.* 2006; Ronen *et al.* 2013; Garud *et al.* 2015; Crawford *et al.* 2017).

One obstacle to understanding the evolutionary basis of phenotypes is the polygenic architecture of many traits. Complex traits (traits affected by many genetic loci and by environmental variation) are ill-suited to study by single-locus methods. In recent years, genome-wide association studies (GWAS) have made it possible to aggregate subtle evolutionary signals that are distributed across the many genetic loci that are associated with a trait of interest (Turchin *et al.* 2012; Berg and Coop 2014; Robinson *et al.* 2015; Field *et al.* 2016; Racimo *et al.* 2018; Uricchio *et al.* 2018). For example, Field

et al. (2016) developed the singleton density score (SDS) to infer recent selection on a variety of traits among the ancestors of people in the United Kingdom. [As discussed below, some empirical findings of these studies have not replicated using effect-size estimates from less-structured GWAS samples, raising the possibility that the selection tests are sensitive to residual population stratification (Berg *et al.* 2018; Sohail *et al.* 2018). Nonetheless, given correct effect-size estimates, these methods are useful.]

Field *et al.* relied on the fact that selection distorts the gene genealogies, or coalescent trees, at genetic loci under selection. In particular, loci under positive selection will have increased in frequency in the recent past, leading to relatively faster coalescence of lineages than if the allele frequency had been constant [the method of Palamara *et al.* (2018) also capitalizes on this idea]. The principle on which the SDS relies is quite general: selection, even when its effect is spread over many loci, leaves systematic signals in coalescent trees at loci underlying trait variation.

The ancestral recombination graph (ARG) (Griffiths and Marjoram 1997) collects coalescent trees at loci along a recombining sequence, encoding information about allele-frequency changes at each site, as well as recombination events between sites. Recently, computational methods for inferring ARGs have advanced considerably (Rasmussen *et al.* 2014; Mirzaei and Wu 2016), allowing a range of applications (Palacios *et al.* 2015).

In this work, we consider ways in which ARGs—and in particular, the coalescent trees of sites associated with a phenotype—might be used to reconstruct the history of the phenotype with which they are associated. The ARG-based approaches we consider are motivated by polygenic traits, and the population-mean level of a polygenic score (a prediction of phenotype from an individual's genotype) is the target of estimation. We present methods for estimating the time course of a population-mean polygenic score through the past, as well as a test for assessing whether an estimated time course is consistent with neutral evolution alone.

We begin by describing estimators and a hypothesis test for phenotypic time courses based on previous theory. Next, we apply these procedures to simulated data, using both true and reconstructed ARGs. Finally, we apply our methods to some human heights in the GBR (Great Britain) subset of the 1000 Genomes panel (1000 Genomes Project Consortium *et al.* 2012), using ARGs inferred by RENT+ (Mirzaei and Wu 2016).

Theory

Background and motivation

The ARG expresses the shared genealogical history of a sample of individuals at a set of genetic loci, accounting for correlations among neighboring loci that arise because of linkage. The ARG contains a coalescent tree for every site in the genome—these trees for specific sites are *marginal trees*. Our methods make use of the marginal trees at a set of sites that are associated with a phenotype. In particular, we

concentrate on the information revealed by the number of coalescent lineages that remain (*i.e.*, that have not yet coalesced) at a time t in the past. The number of lineages at a given time in the past is described by a stochastic process called the ancestral process (*e.g.*, Tavaré 1984).

The intuition behind the methods we present here is shown in Figure 1. If an allele has been selected upward in frequency in the recent past, then the number of chromosomes carrying the allele will likely have increased. Looking backward in time, the number of carriers in the recent past is less than in the present, which forces an excess of recent coalescence events. Similarly, if an allele has been selected downward recently, then there will tend to be fewer recent coalescent events compared with the neutral expectation. If the trait that has been under selection is polygenic, then the signal at each locus associated with the trait will be smaller, and its strength and direction will depend on the effect size at the locus. In Appendix A, we derive the relationship between the rate of coalescence and selection on the phenotype. We show that phenotypic selection acting to increase the population-mean trait value increases the rate of coalescence for alleles that increase the trait and lowers the coalescence rate for alleles that decrease the trait. In contrast, stabilizing selection acting on a trait for which the population mean is at the fitness optimum does not have a systematic directional effect on the coalescent rates.

Our target of estimation is the population-average polygenic score for a trait going backward through time. By polygenic score, we mean a weighted sum of an individual's genotypes, where the weights are the additive effect sizes of each allele. In our case, we are interested in the population-average polygenic score, so we take a weighted sum of the allele frequencies:

$$Z(t) = 2\sum_{i=1}^{k} \beta_i p_i(t). \qquad (1)$$

Here, $Z(t)$ is the population-average polygenic score at time t in the past; $p_i(t)$ is the population frequency of one of the two alleles at locus i at time t in the past, and β_i is the additive effect size of the allele whose frequency is $p_i(t)$, where the effect sizes have been scaled so that the other allele has an effect size of zero. (In practice, effect sizes will be estimated with error, but in this article we treat the effect sizes as known.) The 2 arises because of diploidy.

If the k loci included in the calculation of $Z(t)$ include all the causal loci, then changes in $Z(t)$ (the population-average polygenic score) reflect changes in the population-average phenotypic value in the absence of changes in the distribution of environmental effects on the trait, changes in the effect size, epistasis, and gene-by-environment interaction. Even if these strong assumptions are not met, rapid changes in $Z(t)$ could provide evidence that natural selection has acted on the trait.
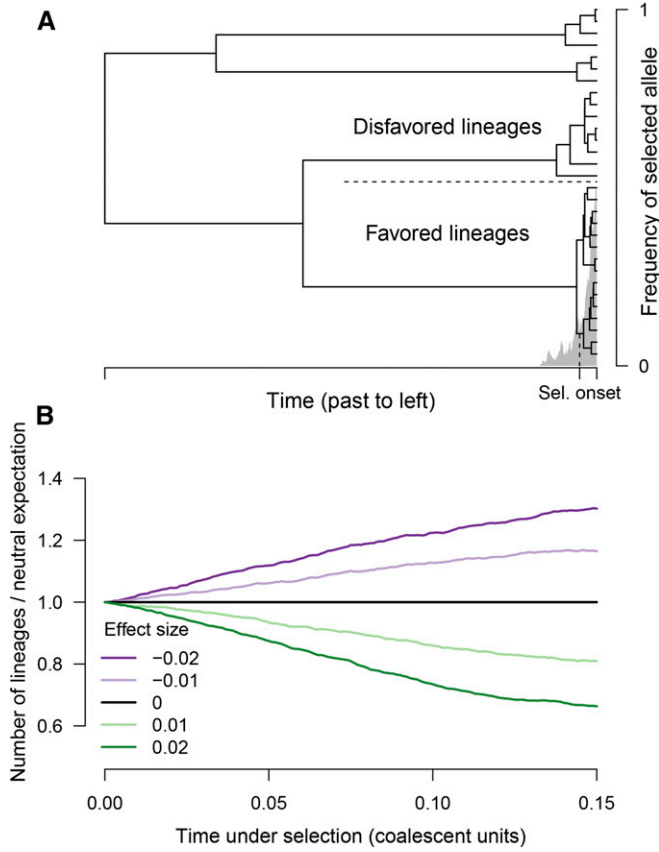
Figure 1 Selection distorts the coalescent tree at a locus. (A) The frequency time course for an allele that has been favored by recent selection is shown in gray. Superimposed is a coalescent tree relating a sample of chromosomes; the top half are of the disfavored type, and the bottom half are of the favored type. Because the favored alleles trace to a pool of ancestors that was small before selection, there is an excess of recent coalescence on the subtree of the favored allele. (B) If a polygenic trait is under directional selection, the rate of coalescence depends on effect size. The time under selection is shown on the horizontal axis, and the vertical axis shows the ratio of lineages remaining compared with the neutral expectation. Different lines show different effect sizes in units of trait standard deviation. Results are an average across 1000 simulations; selection coefficients are determined as if the trait experiences 1% directional truncation selection. Sel., selection.

Our strategy for estimating $Z(t)$ is to estimate the historical allele-frequency time courses, $p_i(t)$, for the loci associated with a trait. Given an estimator of the allele-frequency time courses, $\widehat{p_i(t)}$, we estimate polygenic scores as

$$\widehat{Z(t)} = 2 \sum_{i=1}^{k} \beta_i \widehat{p_i(t)}. \tag{2}$$

If the loci contributing to the polygenic score are independent, then the variance of the polygenic-score estimator is

$$\mathrm{Var}[\widehat{Z(t)}] = 4 \sum_{i=1}^{k} \beta_i^2 \mathrm{Var}[\widehat{p_i(t)}]. \tag{3}$$

We present three methods for estimating historical allele-frequency time courses. A number of authors have investigated

estimating allele-frequency time courses from coalescent genealogies (Slatkin 2001; Coop and Griffiths 2004; Chen and Slatkin 2013) or by applying Wright–Fisher diffusion theory to time-series data (Bollback *et al.* 2008; Schraiber *et al.* 2016). Our approaches are cruder than some of these but have the advantage of being fast enough to be applicable to thousands of GWAS loci.

Estimating the allele-frequency time course at a single locus

We present several methods for estimating the historical allele-frequency time course at a specific biallelic locus given a coalescent tree at the locus (Figure 2). (Our procedures could be generalized to loci with multiple alleles.) In each case, the goal is to estimate the frequency of an allele of interest (*e.g.*, the effect allele) at locus i at time t in the past, or $p_i(t)$.

Proportion-of-lineages estimator: The simplest way to estimate a historical allele frequency is to treat the lineages ancestral to the sample at time t as representative of the population at time t. If the locus has evolved neutrally between the present and time t in the past, then the lineages ancestral to the sample at time t are a random sample—with respect to allelic type—from the population at time t. If the lineages ancestral to the sample are a random sample from the population at time t, then a reasonable estimator of $p_i(t)$ is the proportion of lineages at time t that carry the allele of interest,

$$\widehat{p_i(t)} = \frac{j_i(t)}{r_i(t)}, \tag{4}$$

where $r_i(t)$ is the total number of lineages ancestral to the present-day sample at locus i at time t and $j_i(t)$ is the number of lineages at time t that carry the allele of interest. Assuming that the mutation distinguishing the alleles has only appeared once in the history of the sample, the lineages that carry the derived allele are those ancestral to contemporary copies of the derived allele, which must coalesce to one lineage before coalescing with the rest of the tree. If the tree for the locus is known, then the branch on which the mutation must have appeared is known, but the exact time of the mutation is not. In practice, we assume that the mutation occurred in the middle of the branch connecting the derived subtree to the rest of the tree. (We make this assumption when implementing all estimators.)

If the population size at time t is large compared with the number of ancestral lineages at time t, then conditional on $r_i(t)$, the number of ancestral lineages carrying the allele of interest, $j_i(t)$, is distributed approximately as a binomial$[r_i(t), p_i(t)]$ random variable. Thus, Equation 4 is the maximum-likelihood estimator of $p_i(t)$ and, conditional on $r_i(t)$, its sampling variance can be estimated as (dropping the subscript i's and parenthetical t's for compactness)

$$\mathrm{Var}(\hat{p}) = \frac{\frac{j}{r}(1 - \frac{j}{r})}{r} = \frac{j(r-j)}{r^3}. \tag{5}$$
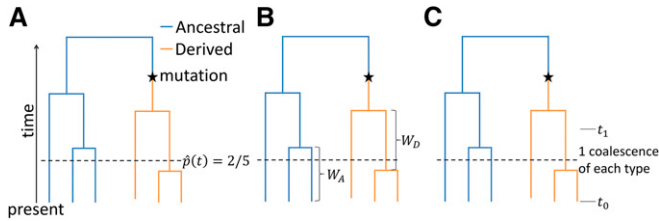
Figure 2 Schematics of three estimators of historical allele frequencies using coalescent trees. In each tree, we show the information used to estimate allele frequency in the population at a time represented by the horizontal dashed line. (A) The proportion-of-lineages estimator. The estimate of the derived-allele frequency in the population is the derived-allele frequency among the ancestors of the sample, or 2/5 in the picture. (B) The waiting-time estimator. Relative "sizes" of the subpopulations of ancestral-allele and derived-allele carriers are estimated as a function of intercoalescence times among the ancestors of the present-day ancestral- and derived-allele carriers. (C) The lineages-remaining estimator. The relative sizes of the subpopulations of ancestral-allele and derived-allele carriers are estimated by examining the number of coalescence events that occur between prespecified time points, here called t_0 and t_1.

In Appendix B, we give a Bayesian interpretation of the proportion-of-lineages estimator that relies on connections between neutral diffusion and the ancestral process (Tavaré 1984).

If chromosomes carrying the two alleles have differed in fitness between the present and time t, then Equation 4 will in general be a biased and inconsistent estimator of $p_i(t)$. Chromosomes carrying alleles that have been favored by selection will be more likely to leave offspring in the present-day population than will chromosomes carrying unfavored alleles. Favored alleles will thus be overrepresented among the lineages ancestral to the sample compared with their actual frequency at time t. However, even if Equation 4 is an inconsistent estimator of the population's allele frequency, it retains the interpretation of being the allele frequency among lineages ancestral to the sample. Thus, when Equation 2 is applied to allele frequencies estimated by Equation 4, the result is the mean polygenic score among chromosomes ancestral to the present-day sample at some time in the past.

Waiting-time estimator: The proportion-of-lineages estimator proposed in Equation 4 works well under neutrality, but under selection it tends to underestimate the degree of allele-frequency change experienced by selected alleles. One potential solution is to consider the chromosomes in the population as two separate subpopulations (Hudson and Kaplan 1988)—one for the carriers of each of the two alleles at the locus—and to estimate the sizes of those two populations over time. At locus i, denote the sizes of these two subpopulations at time t as $N_i(t)$ (for the allele of interest) and $M_i(t)$ (for the other allele). The frequency of the allele of interest at time t in the past, or $p_i(t)$, is

$$p_i(t) = \frac{N_i(t)}{N_i(t) + M_i(t)}. \tag{6}$$

Here, we propose to estimate $N_i(t)$ and $M_i(t)$ on the basis of properties of the coalescent trees for the two alleles. We then estimate by plugging these estimates, $\widehat{N_i(t)}$ and $\widehat{M_i(t)}$, into Equation 6, giving

$$\widehat{p_i(t)} = \frac{\widehat{N_i(t)}}{\widehat{N_i(t)} + \widehat{M_i(t)}}. \tag{7}$$

The estimator in Equation 7 is not unbiased in general, even if the estimates of $N_i(t)$ and $M_i(t)$ are unbiased. But separating the problems of estimating the two subpopulations confers an advantage: this estimator does not assume, as the proportion-of-lineages estimator does, that the two allelic types have had equal fitness between t and the present.

Assuming that the mutation that distinguished the two alleles occurred only once in the history of the sample, the chromosomes carrying the two alleles can be treated as two distinct subsamples between the time of the mutation and the present. Among the ancestors of the allele subsample carrying the allele of interest, coalescent time τ accrues according to

$$\tau_i(t) = \int_0^t 1/N_i(z)dz,$$

where t is measured in generations. It follows that

$$\frac{d\tau_i(t)}{dt} = 1/N_i(t),$$

suggesting that $N_i(t)$ can be estimated by assessing the rate at which coalescent time accrues. $M_i(t)$ can then be estimated analogously and $p_i(t)$ can be estimated by Equation 7. We present two related approaches to estimating the rate of accrual of coalescent time in each subsample: one approach in which estimates are made with respect to waiting times between coalescent events, and another in which estimates are made with respect to the number of lineages ancestral to the subsample at a specified time in the past. In both approaches, we assume that $N_i(t)$ and $M_i(t)$ are piecewise constant, but it is possible to modify these estimators under other assumptions about how $N_i(t)$ and $M_i(t)$ change between time points.

The number of coalescence events in a time interval depends on the amount of coalescent time passed. Our approaches amount to inverting this relationship to estimate the amount of coalescent time passed on each lineage (and thus their relative population sizes) in a method-of-moments framework. In this subsection, we assess $N_i(t)$ and $M_i(t)$ according to the time passed between fixed numbers of coalescence events.

Suppose that $N_i(t) = N$, assumed to be constant from a time point of interest (defined here to be $t = 0$) until ℓ

coalescences have occurred within the subsample, which at $t = 0$ consists of n_i lineages. Define Y_k as the waiting time to the kth coalescence, starting from the $k - 1$th coalescence event (or from $t = 0$ if $k = 1$). Define the total waiting time from $t = 0$ to the ℓth coalescence as $Y = \sum_{k=1}^{\ell} Y_k$. Then, in units of generations, each Y_k is an independent, exponentially distributed random variable with rate $(n_i - k + 1)(n_i - k)/(2N)$. Thus,

$$
\mathrm{E}(Y) = \sum_{k=1}^{\ell} \frac{2N}{(n_i - k + 1)(n_i - k)} = \sum_{w=n_i-\ell+1}^{n_i} \frac{2N}{w(w-1)}
$$
$$
= 2N\left(\frac{1}{n_i - \ell} - \frac{1}{n_i}\right), \tag{8}
$$

and

$$
\mathrm{Var}(Y) = 4N^2 \sum_{w=n_i-\ell+1}^{n_i} \frac{1}{w^2(w-1)^2}. \tag{9}
$$

For large n_i and intermediate values of ℓ, Y is approximately normally distributed with expectation and variance given by Equation 8 and Equation 9 (Chen and Chen 2013). One estimator of $N_i(t) = N$, which is both a method-of-moments estimator and the maximum-likelihood estimator under the asymptotically normal distribution of Y, is

$$
\widehat{N_i(t)} = \frac{Y}{2[1/(n_i - \ell) - 1/n_i]}. \tag{10}
$$

With $\ell = 1$, this is the "skyline" estimator of Pybus *et al.* (2000), and with $\ell \geq 1$, it is the generalized skyline estimator of Strimmer and Pybus (2001). It is unbiased under the assumption of constant $N_i(t)$ between coalescence events. In our implementation, after the tree has coalesced down to one lineage, we assume that $N_i(t)$ remains constant into the past if its allele is ancestral and that $N_i(t)$ remains constant before dropping to zero in the middle of the branch on which the mutation arose if its allele is derived. (In principle, derived *vs.* ancestral status may be determined by the tree topology or "forced" on the basis of prior knowledge if the tree topology may be in error. We use the tree topology.)

One may estimate $p_i(t)$ by

$$
\widehat{p_i(t)} = \frac{\widehat{N_i(t)}}{\widehat{N_i(t)} + \widehat{M_i(t)}}, \tag{11}
$$

where $M_i(t)$ is estimated analogously to $N_i(t)$. In principle, the estimates of $N_i(t)$ and $M_i(t)$ may be based on waiting times for different numbers of coalescences (ℓ), and they will likely be constant for different intervals of time. By a first-order Taylor approximation argument described in Appendix C,

$$
\mathrm{E}[\widehat{p_i(t)}] \approx p_i(t),
$$

but in practice, $\widehat{p_i(t)}$ can be substantially biased. Its approximate variance is given in Appendix C.

One important decision in using this estimator is the choice of ℓ—i.e., the number of coalescence events to wait for—for each allele. Small values of ℓ lead to variable estimates. On the other hand, the estimator assumes that the size of the subpopulation is constant until ℓ coalescent events have occurred, which may be increasingly unrealistic for large ℓ. In this article, we use $\ell = 1$ and defer investigation of different choices of ℓ for future work.

Lineages-remaining estimator: The next estimator, which we term the "lineages-remaining" estimator, operates on a principle similar to the waiting-time estimator. The coalescent time passed on each allele's background is evaluated by looking at the local rate of coalescence, and the relative numbers of carriers of each allele are estimated to form an allele-frequency estimate. The difference is that whereas the waiting-time estimator estimates population sizes for each allele between coalescent events, the lineages-remaining estimator estimates the population sizes between prespecified times by comparing the number of lineages of each type that remain (*i.e.*, have not coalesced) at the more ancient end of a time interval with the number present at the more recent end of the interval.

Suppose that $N_i(t) = N$ during an interval of Δt generations. At the end of the interval closer to the present there are $n_i^{(0)}$ lineages, and at the end of the interval further into the past there are $n_i^{(t)}$ lineages. The expected number of lineages remaining at the end of the interval can be approximated as

$$
\mathrm{E}\left[n_i^{(t)}\right] \approx \frac{n_i^{(0)}}{n_i^{(0)} + \left[1 - n_i^{(0)}\right]e^{-\tau/2}}, \tag{12}
$$

where τ is the amount of coalescent-scaled time elapsed during the interval (Griffiths 1984; Maruvka *et al.* 2011; Chen and Chen 2013; Jewett and Rosenberg 2014). Here, $N_i(t)$ is a haploid population size, so $\tau = \Delta t/N_i(t)$. For large $n_i^{(0)}$ and intermediate lengths of time, $n_i^{(t)}$ is approximately normally distributed (Griffiths 1984; Chen and Chen 2013). An estimator for $N_i(t) = N$, which is both a method-of-moments estimator—one based on the approximate value of $\mathrm{E}[n_i^{(t)}]$—and the maximum-likelihood estimator under the limiting normal distribution, is

$$
\widehat{N_i(t)} = \frac{\Delta t}{2\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)} - 1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)} - 1}\right]\right\}}. \tag{13}
$$

As the tree coalesces to few lineages, there are several edge cases in which the estimator is undefined. Our methods for handling these cases are discussed in Appendix C.

It may be impractical to estimate $N_i(t)$ because Δt in generations may be unknown. However, Δt cancels in the estimator of the allele frequency $p_i(t)$,

normality fails over longer timescales, in part because allele frequencies are bounded by 0 and 1.) Here, f is the coalescent time that has passed between time points and p_i

$$\frac{\widehat{N_i(t)}}{\widehat{N_i(t)} + \widehat{M_i(t)}} = \frac{\log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right] - \log\left[\frac{m_i^{(0)}}{m_i^{(0)}-1}\right]}{\log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right] - \log\left[\frac{m_i^{(0)}}{m_i^{(0)}-1}\right] + \log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)}-1}\right]}, \tag{14}$$

where $M_i(t)$ is the number of carriers of the reference allele in the population at locus i at time t, and $m_i(0)$ and $m_i(t)$ are the numbers of lineages carrying the reference allele at the ends of the time interval closer to and more distant from the present, respectively. Approximate expressions for the variance of the estimator in Equation 14 are in Appendix C.

In practice, the estimator in Equation 14 will be evaluated at a set of times. Here, we evaluate the estimator every 0.001 coalescent units, starting at the present and extending back four coalescent units. The grid of times at which changes in the number of lineages of each type are considered will influence the estimate. Finer grids will lead to estimates that are more variable but less biased by the assumption of constant $N_i(t)$ and $M_i(t)$ between time points.

Testing time courses for selection

Once a polygenic-score time course has been constructed, it can be tested for selection. Whereas the relevance for trait evolution of the estimators proposed in the previous sections depends on the proportion of trait variance accounted for by the polygenic score, polygenic-score time courses can be tested for selection even if they account for a small proportion of the trait variance. Further, although we apply the test in this section to estimated polygenic-score time courses, it is also applicable to measured time-series data on allele frequencies when available.

We propose a test for selection that amounts to a modification of the Q_X framework of Berg and Coop (2014), an analog of Q_{ST}–F_{ST} tests for phenotypic selection (Whitlock 2008). Berg and Coop proposed Q_X to test for overdispersion of polygenic scores among population samples, relative to neutral expectations. Here, we check for overdispersion among a set of time points along one population branch. We denote our time-based modification of Q_X as T_X.

Suppose that for each time t_j in a sequence of times, we observe a population-level polygenic score, $Z(t_j) = 2\sum_{i=1}^{k}\beta_i p_i(t_j)$, as in Equation 1. Assume that changes across time at distinct loci are independent. Following the Q_X framework, we model polygenic scores using a multivariate normal distribution. Specifically, we posit that at each locus, over short timescales, allele-frequency changes between time points follow a normal$[0, f p_i(1-p_i)]$ distribution (Cavalli-Sforza *et al.* 1964; Nicholson *et al.* 2002). (Approximate

is the allele frequency at locus i at one end of the interval. (In practice, we choose p_i to be the allele frequency at the end of the interval closer to the present.) The parameter f is constant across loci.

If allele-frequency changes at each locus are independent and normally distributed, then changes between time points in the polygenic scores are normal with expectation 0 and variance $f 4 \sum_{i=1}^{k} \beta_i^2 p_i(1-p_i)$. The variance of polygenic-score changes can also be written as $f 2 V_A$, where $V_A = 2\sum_{i=1}^{k} \beta_i^2 p_i(1-p_i)$ is the additive genetic variance of the trait.

Imagine we have a time course of polygenic scores $Z(t_0), Z(t_1), Z(t_2), \ldots, Z(t_w)$, with $t_0 < t_1 < \ldots < t_w$. Under neutrality, for each time point $j \in 1, \ldots, w$, the statistic

$$X_j = \frac{Z(t_j) - Z(t_{j-1})}{\sqrt{2V_A(t_j - t_{j-1})}} \tag{15}$$

has a normal$(0, 1)$ distribution. (V_A can be recomputed for each value of j using allele frequencies at t_{j-1}. In practice, we use the same V_A—the one computed from allele-frequency estimates closest to the present—for each time interval.) Moreover, under neutrality, allele-frequency changes in distinct time intervals are independent, so values of X_j are independent for distinct j. Thus, the sum across time points

$$T_X = \sum_{j=1}^{w} X_j^2 \tag{16}$$

has a $\chi^2(w)$ distribution. [For details on the equivalence of the T_X statistic in Equation 16 to the Q_X statistic of Berg and Coop (2014) in this scenario, see Appendix D.] In contrast, under directional selection, changes in allele frequency across time depend on the effect size of the locus, leading to large changes in the polygenic score and T_X values larger than predicted by the $\chi^2(w)$ distribution. This test is an analog of time-course tests for phenotypic selection based on neutral Brownian motion (Lande 1976; Turelli *et al.* 1988), but with the advantage that we know V_A.

T_X can be compared with a $\chi^2(w)$ distribution to test for significance, or it can be compared with a distribution obtained by permuting either effect sizes or their signs. If the $\chi^2(w)$ distribution is used, then coalescent times elapsed

between polygenic scores can be estimated by assessing estimated allele-frequency changes between time points, either at putatively neutral loci or trait-associated loci. Using estimated times produces type-I error rates closer to the nominal value because the estimators we use tend to change at a systematically different rate than the actual allele frequencies: the proportion-of-lineages estimator changes more slowly than do population allele frequencies, and the other estimators change more quickly than population allele frequencies. In practice, we use the sample variance of $\Delta(\hat{p})/[\hat{p}(1-\hat{p})]$ as an estimate of coalescent time passed between time points, where $\Delta(\hat{p})$ is an estimated allele-frequency change at a variable locus and $\hat{p}$ is the estimated allele frequency at the more recent end of the time interval. Assessing elapsed time on the basis of trait-associated loci may lead to a power decrement, but it will not be large—most of the T_X signal comes from coordination of small shifts across loci and not from larger-than-expected allele-frequency changes (Berg and Coop 2014).

Data availability

Supplemental material, including code for the article, available at Figshare: https://doi.org/10.25386/genetics.6955367. Additionally, code used for running the simulations and implementing all data analyses presented here is available at http://github.com/mdedge/rhps_coalescent. The version of the code used here is permanently archived at doi: 10.5281/zenodo.1461077.

Simulation Results

We examined the performance of our methods in coalescent simulations. In particular, we simulated coalescent trees for unlinked loci associated with a phenotype. The simulated loci evolve neutrally or under directional selection. In particular, if an allele at locus i has effect size β_i on a trait, and the trait experiences a selection gradient $\alpha(t)$ at time t, then the selection coefficient on the allele—representing the fitness of the heterozygote minus the fitness of the ancestral homozygote—is $s(t) = \alpha(t)\beta_i$ (Charlesworth and Charlesworth 2010, equation B3.7.7). The coalescent simulations were run in mssel (Berg and Coop 2015), a version of ms (Hudson 2002) that takes allele-frequency time courses that may be produced by selection, and assumed a constant population size. Effect sizes for the derived allele are drawn from a normal distribution centered at zero. For details on the simulations, see Appendix E.

We consider the performance of the methods when (1) the true trees are provided as input, and (2) when the trees must be reconstructed from sequence data. We use the software RENT+ (Mirzaei and Wu 2016) for tree reconstruction.

Before considering systematic results over many simulations, we present estimated time courses for one representative simulation (Figure 3). In the simulation shown, the trait was under selection upward in the past but has evolved neutrally recently.
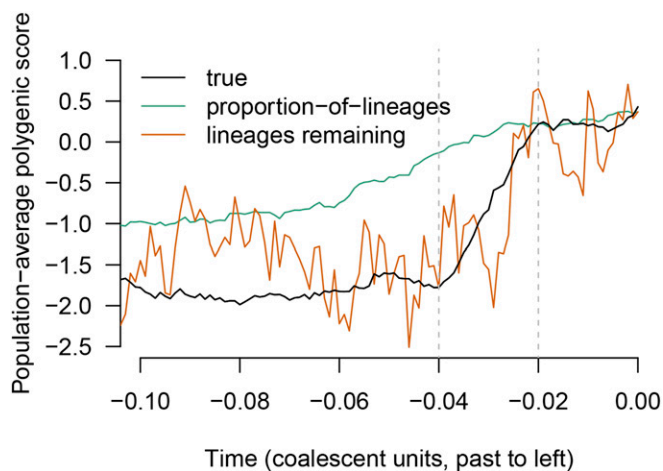

Figure 3 The behavior of two of the proposed estimators, the proportion-of-lineages estimator and the lineages-remaining estimator, in a single simulation. The true time course of the population-average polygenic score is shown in black. The trait is under selection upward from 0.04 to 0.02 coalescent units in the past, leading to a shift of approximately two standard deviations in present-day units. Between the present and the most recent period of selection, the proportion-of-lineages estimator performs well. During the period of selection (and before), the ancestors of the sample are no longer representative of the ancestral population and the estimator becomes biased. In contrast, the lineages-remaining estimator is variable at all times but is less affected by bias associated with the period of selection. (For readability, results for the waiting-time estimator are not shown, but they are similar to the lineages-remaining estimator.) The estimates shown are formed from simulated coalescent trees for a sample of 200 chromosomes.

The proportion-of-lineages estimator (Equation 4) estimates allele frequencies as the proportion of lineages ancestral to the sample carrying each allele. It is expected to perform well under neutrality, and it does here: in the neutral period between the offset of selection and the present, it tracks the true polygenic score closely. During the period of selection, looking backward in time, the proportion-of-lineages estimator strays off target, slowly recovering in the period before the onset of selection. Looking forward in time, it is as if the proportion-of-lineages estimator "anticipates" shifts due to selection. As mentioned in the *Proportion-of-lineages estimator* subsection, the apparent anticipation occurs because, if there has been selection between the present and time t in the past, then the ancestors of the present-day sample at time t are a biased sample from the population at time t. For example, if the trait has been selected upward, then the ancestors of a present-day sample will have had high trait values compared with their peers.

The waiting-time estimator (Equation 11) and the lineages-remaining estimator (Equation 14) do not rely on an explicit neutrality assumption. Instead, they track the relative passage of coalescent time—measured, roughly, in terms of coalescence events—for each allelic type. These estimators track the rapid change in the polygenic score during the period of selection much more closely than the proportion-of-lineages estimator. (Only the lineages-remaining estimator is
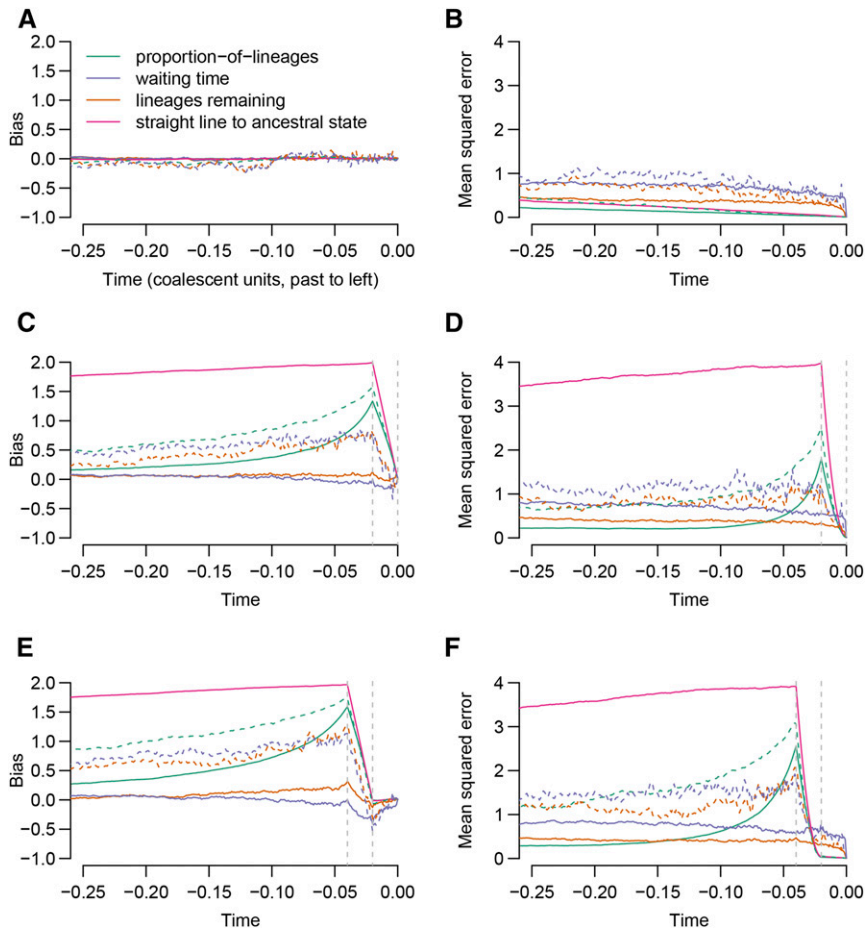
Figure 4 Bias and MSE of the proposed estimators. For the estimates computed from true coalescent trees (solid lines), 1000 simulations were performed; RENT+ was used to reconstruct trees for 100 of these trials (dashed lines). In each trial, the polygenic scores were formed from 100 loci. The polygenic scores either evolved neutrally (A and B) or with selection leading to a shift in mean polygenic score of approximately two standard deviations (in present-day units), either occurring from 0.02 coalescent units ago to the present (C and D), or from 0.04 to 0.02 coalescent units ago (E and F).

shown in Figure 3.) At the same time, these estimators rely on a highly stochastic signal—in the case of the waiting-time estimator (with $\ell = 1$), single coalescence events—and they are noisier than the proportion-of-lineages estimator as a result.

The patterns seen in Figure 3 reflect the performance of the methods over many simulations, as detailed in the next subsections.

Estimator performance: bias and mean squared error

Figure 4 shows bias and mean squared error (MSE) of our estimators of the historical polygenic score across three scenarios: one in which the trait has evolved neutrally, one in which there has been recent directional selection on the trait, and a third in which there has been directional selection on the trait in the past but the trait has recently evolved neutrally. These estimators are also compared with a "straight-line" estimator: a straight line that goes from the present value to the ancestral state (*i.e.*, all derived-allele frequencies zero) in two coalescent units. In the neutral case, none of the estimators show marked bias and the proportion-of-lineages estimator has the lowest variance (and thus lowest MSE). Estimators formed from trees reconstructed by RENT+ (dashed lines) rather than the true trees (solid lines) are noisier and they do not outperform the straight-line estimator under neutrality.

In the presence of selection, the proportion-of-lineages estimator is badly biased, and the severity of the bias increases during the interval of selection (looking backward in time). The waiting-time and lineages-remaining estimators are less strongly biased in the presence of selection and they achieve similar MSEs under selection and neutrality. Again, estimators formed from RENT+ trees perform worse than estimators formed from the true trees, but in the presence of selection, they outperform the straight-line estimator.

Interval estimation: coverage probabilities

Figure 5 shows the coverage probabilities of nominal 95% confidence intervals formed on the basis of the proportion-of-lineages, waiting-time, and lineages-remaining estimators and their (approximate) variances.

Under neutrality, and when using the true trees, all confidence intervals have approximately the correct coverage, although coverage for confidence intervals for both the lineages-remaining and waiting-time estimators decays further into the past. The decay of the coverage probability far in the past makes sense for the lineages-remaining and waiting-time estimators: both of these estimators implicitly assume that the number of carriers of each allele in the population remains constant between coalescent events. This assumption may be a reasonable approximation in the recent past, when
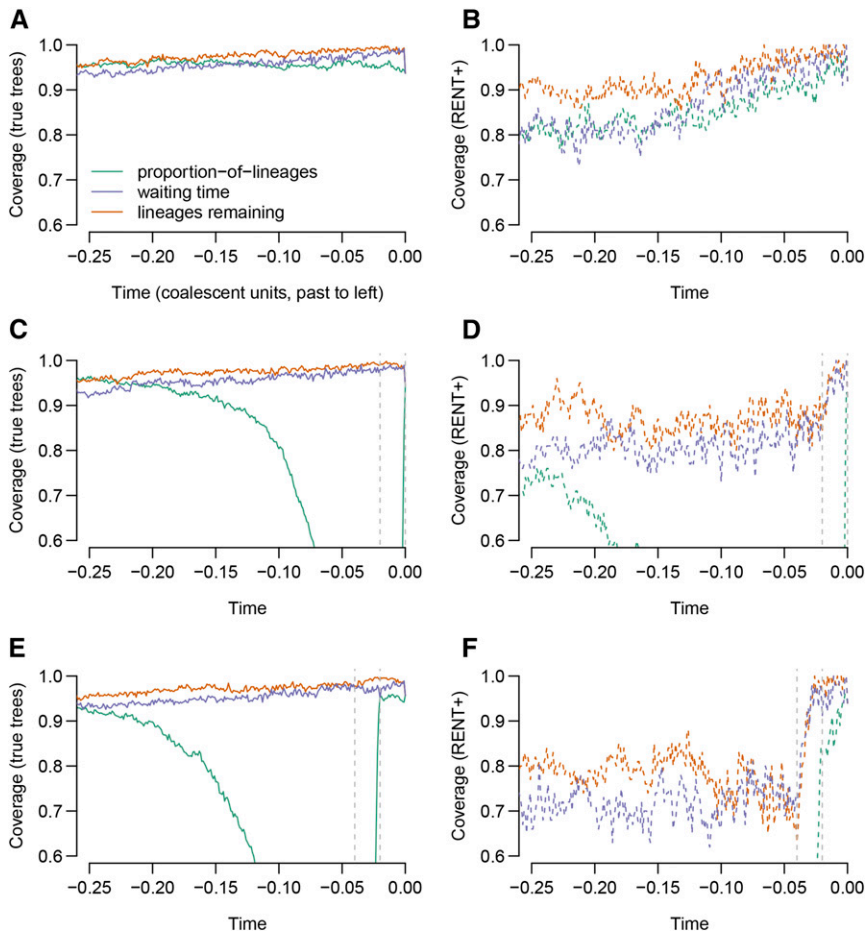
Figure 5 Confidence-interval coverage for nominal 95% confidence intervals based on the proposed estimators. All confidence intervals were formed assuming an approximately normal distribution for the estimator, adding ± 1.96 SE to the estimate. Standard errors were computed by taking the square root of the approximate variance of each estimator. Coverage probabilities are based on (A, C, and E) 1000 simulations for true trees or (B, D, and F) 100 simulations for RENT+ trees. Simulations were conducted under either (A and B) neutrality, (C and D) a shift of approximately two standard deviations over the last 0.02 coalescent units, or (E and F) a shift of two standard deviations from 0.04 to 0.02 coalescent units ago.

coalescent events are frequent, but become untenable in the distant past, when coalescence times are longer.

Confidence intervals computed on the basis of RENT+ trees only achieve the nominal coverage in the very recent past and become anticonservative further back in time. This behavior is expected; the variances we use incorporate stochasticity in the coalescent process but do not account for randomness arising from errors in tree estimation. [Stochasticity in the coalescent process has been called "coalescent error" and contrasted with randomness from errors in tree estimation, or "phylogenetic error" (Ho and Shapiro 2011)].

Under selection, confidence intervals from the proportion-of-lineages estimator have very low coverage—this arises from the bias documented in Figure 4. The coverage probabilities of the waiting-time and lineages-remaining estimators are less changed by selection.

Power of T_X

We assessed the performance of the T_X statistic (Equation 16) as a test statistic for detecting selection in the simulations shown in Figure 4 and Figure 5. We constructed the test statistic from the allele-frequency estimates produced by each of the proposed estimators and compared it against both the theoretical χ^2 distribution and against a permutation distribution. Table 1 shows the results. Under neutrality (first two rows), comparing T_X against a distribution formed by randomly permuting the effect sizes produces acceptable type-I error rates. (There are 100 simulations using RENT+ and the values in Table 1 do not differ significantly from 0.05.) When the theoretical χ^2 distribution is used, RENT+ type-I error rates are unacceptably high, but the type-I error rates produced from the true trees are acceptable.

Of the methods with acceptable type-I error, tests using the allele frequencies estimated by the proportion-of-lineages estimator have by far the highest power. It may seem paradoxical that the proportion-of-lineages estimator is the best of our estimators at detecting selection, given that the estimated time courses produced by the proportion-of-lineages estimator are biased in the presence of selection. However, in our simulations, the proportion-of-lineages estimator generally moves in the correct direction in the presence of selection, albeit more slowly than it should. In contrast, the other two allele-frequency estimators are highly variable, leading to wide null distributions and decreased power. The proportion-of-lineages estimator can also be thought of as the mean polygenic score among lineages ancestral to the sample, and the test for selection responds to changes in the mean polygenic score of the ancestors that are faster than would be expected under the null hypothesis of neutral evolution.

Polygenic-Score Time Courses **243**

Table 1 Power/type-I error of various implementations of the T_X statistic

	Input	Proportion-of-lineages estimator		Waiting-time estimator		Lineages-remaining estimator	
		χ^2 distribution	Permutation distribution	χ^2 distribution	Permutation distribution	χ^2 distribution	Permutation distribution
Neutral	True trees	0.053	0.054	0.030	0.060	0.061	0.064
	RENT+	0.17	0.08	0.31	0.04	0.11	0.09
Selection, 0.02 coalescent units ago–now	True trees	1	1	0.033	0.060	0.194	0.222
	RENT+	—	0.87	—	0.05	—	0.15
Selection, 0.04–0.02 coalescent units ago	True trees	0.969	0.954	0.035	0.065	0.058	0.084
	RENT+	—	0.37	—	0.08	—	0.02

Power/type-I error (at the 0.05 level) of various implementations of the T_X statistic in the simulations shown in Figure 4 and Figure 5, with 200 chromosomes sampled in the present. Power is not shown for methods with type-I error rates ≥ 0.1. T_X was computed using allele-frequency estimates $0, 0.01, 0.02, \ldots, 0.1$ coalescent units before the present.

With the proportion-of-lineages estimator, using true trees unsurprisingly gives better power than RENT+ trees, but RENT+ trees still have substantial power.

In Table 1, power is higher when selection occurs closer to the present. To explore the relationship between the timing of selection and present-day sample size, we conducted additional simulations. In these simulations, we assessed the power of the T_X test (using the proportion-of-lineages allele-frequency estimates from true trees, and comparing with a permutation distribution) to detect an approximate one-standard-deviation shift in the population-mean polygenic score. We varied the timing of the shift and the present-day sample size. Figure 6 shows the results. For detecting selection close to the present, power increases with sample size. However, for selection further in the past, power reduces to the type-I error rate, regardless of the present-day sample size. This is because power to detect selection depends on unusual coalescent times during the period of selection, and by 0.1 coalescent units in the past, most coalescent events have already occurred, even in large samples. For example, even extremely large present-day samples have, in expectation, ~ 200.5 ancestors tracing back 0.01 coalescent units in the past and ~ 20.5 ancestors 0.1 coalescent units in the past (Maruvka *et al.* 2011; Jewett and Rosenberg 2014). Thus, it will likely be impossible to detect all but the strongest selective events by their signatures in coalescent trees if they are over 0.1 coalescent units in the past. T_X's power to detect selection up to $\sim 0.02 - 0.04$ coalescent units into the past represents an extension of the SDS statistic (figure S6 in Field *et al.* 2016), which has excellent power in the very recent past but very little power beyond the expected length of a terminal branch $[2/n$ in coalescent units, where n is the present-day sample size (Fu and Li 1993)]. In Appendix F, we show empirical power for a test statistic analogous to SDS computed from the lengths of the terminal branches (Figure F1). This SDS analog has similar power to T_X near the present, but its power decays more rapidly for selection further in the past.

Simulations with larger numbers of loci

All of the above simulations are of polygenic scores that incorporate 100 loci each. Reassuringly, the performance is extremely similar for all estimators and tests if the number of loci is increased, using the estimators based on the true trees. In Supplemental Material, Figure S1, Figure S2, and Table S1, we show results analogous to Figure 4, Figure 5, and Table 1 with 1000 loci per polygenic score.

Empirical Application: Human Height

We applied our proposed estimators to human polygenic scores for height. Genetic variation within Europe related to human height has been studied by many investigators interested in polygenic selection (Turchin *et al.* 2012; Berg and Coop 2014; Robinson *et al.* 2015; Field *et al.* 2016; Berg *et al.* 2017; Racimo *et al.* 2018; Uricchio *et al.* 2018). A recent pair of articles compared the results produced by existing tests for polygenic selection when applied to human height using GWAS effect sizes from two different studies (Berg *et al.* 2018; Sohail *et al.* 2018). Most previous work has used GWAS effect sizes from the Genetic Investigation of Anthropometric Traits (GIANT) consortium (Wood *et al.* 2014), whereas the new work uses GWAS effect sizes from the larger and presumably less structured UK Biobank sample (Sudlow *et al.* 2015). Tests for polygenic selection on height provide much less evidence for selection when UK Biobank effect sizes are used than when effect sizes from GIANT are used. One possible explanation is that GIANT effect sizes are contaminated by some degree of population stratification.

In Figure 7, we show estimated population-mean polygenic-score time courses among populations ancestral to the GBR (British in England and Scotland) subsample of the 1000 Genomes Project (1000 Genomes Project Consortium *et al.* 2012). In the top panel GIANT effect sizes are used, and in the bottom panel UK Biobank effect sizes are used. Polygenic scores were constructed by taking the top locus in each of ~ 1700 approximately independent genetic regions. [These polygenic scores are identical to those used in Berg *et al.* (2018); our "UK Biobank" is their "UKB-GB."] Coalescent trees for these loci were estimated in RENT+. (Details in Appendix G.)

When the time courses are constructed using GIANT, all three estimators suggest that the population-mean polygenic
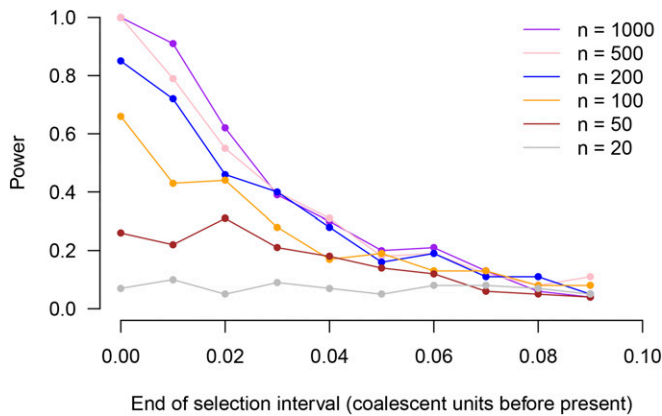
Figure 6 Power of the T_X test as a function of timing of selection and present-day sample size. Power is shown on the vertical axis. The horizontal axis shows the timing of a pulse of selection lasting 0.005 units and resulting in a shift in the population-average polygenic score of approximately one standard deviation (in present-day standard deviations). Each point is based on power from 100 simulations, using the proportion-of-lineages estimator on the true trees and using a null distribution obtained by randomly permuting locus effect sizes.

Figure 7 Estimated population-mean polygenic scores for height in the population ancestral to the GBR subset of the 1000 Genomes Project (91 genomes). Time is displayed in approximate coalescent units, defined such that the mean across loci of the time (measured in mutations) to the most recent common ancestor is 2. To construct the polygenic scores, the top locus in each of ~ 1700 approximately unlinked genomic regions was chosen. (See Appendix G for more information.) The top panel shows the results when effect sizes estimated by GIANT are used; the bottom panel shows analogous results when UK Biobank effect-size estimates are used. In both panels, the effect sizes are scaled so that the standard deviation of the polygenic score is 1 in the present, assuming linkage equilibrium among SNPs. 1000G, 1000 Genomes.

score for height has increased in the recent past. Using the proportion-of-lineages approach, an increase of approximately three present-day polygenic-score standard deviations is estimated. In contrast, time courses estimated using the UK Biobank effect sizes show little apparent change in the recent past: the proportion-of-lineages estimator suggests a recent decrease of ~ 0.34 standard deviations.

Further, when the change from the present to the most recent time point (0.001 units in RENT+) is assessed by the T_X test, both sets of effect sizes yield some evidence of selection, but the evidence is stronger with GIANT effect sizes than with UK Biobank effect sizes. Specifically, with GIANT, $T_X(1) = 14.9$, $P = 0.0005$ from 10,000 permutations, and with UK Biobank, $T_X(1) = 7.0$, $P = 0.0112$. [We do not claim that the difference between the T_X values across the data sets is itself significant (Gelman and Stern 2006), merely that the pattern of weakened evidence in the UK Biobank matches that observed by recent work (Berg *et al.* 2018; Sohail *et al.* 2018).] However, using both data sets, the evidence for selection is limited to periods very close to the present. If the same set of times evaluated in the simulations is evaluated for the height polygenic scores (approximate coalescent times $0, 0.01, 0.02, \ldots, 0.1$ before the present), neither polygenic-score time course provides evidence for selection ($P \approx 0.2$ in both cases). GIANT effect sizes produced much lower P-values for recent selection on height in the UK in a recent article (Field *et al.* 2016), but that work used a sample of 3195 genomes, whereas the GBR subset of the 1000 Genomes sample contains only 91 genomes.

Thus, our estimators broadly recapitulate the pattern of other methods for detecting polygenic selection, finding evidence suggestive of selection when GIANT effect sizes are used but much weaker evidence when UK Biobank effect sizes are used.
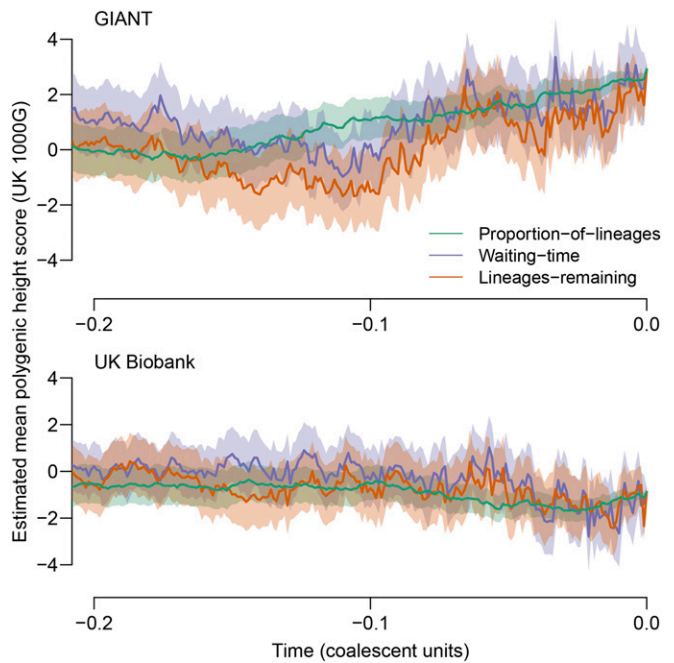
Discussion

We have proposed a set of estimators and tests for population-mean polygenic scores over time, given (additive) effect sizes for a trait at independent trait-associated loci and coalescent trees for the trait-associated loci. Estimation of the population-mean polygenic-score time course is most effective when the trait (and its associated loci) evolves neutrally and the ancestors of the sample are representative of the ancestral population. When the trait has been under selection, estimation is still possible, but the estimates obtained are noisier. Tests for polygenic selection that are based on coalescent trees have the potential to be powerful in the recent past.

In terms of practical applications, we have produced one estimator that produces good estimates of population-mean polygenic-score time courses under neutrality and that is also well powered to detect departures from neutrality (the proportion-of-lineages estimator). The other two estimators are less biased by selection, but they are variable and less useful for detecting selection. At this writing, one sensible procedure for fitting these methods to data would be to form initial estimates using the proportion-of-lineages estimator and test them for selection using the T_X statistic. If the test

suggests selection, then the ancestors of the sample may not be representative of the ancient population and polygenic-score time courses from the proportion-of-lineages estimator may be biased. In that case, the waiting-time or lineages-remaining estimators might be applied.

These methods add to a set of methods that use GWAS information to study the history of complex traits (Berg and Coop 2014; Field *et al.* 2016; Berg *et al.* 2017; Racimo *et al.* 2018; Uricchio *et al.* 2018). Many of these methods have been applied to human height, and our methods produce similar conclusions when applied to the same data (Berg *et al.* 2018; Racimo *et al.* 2018; Sohail *et al.* 2018; Uricchio *et al.* 2018). Our methods add to previous work by estimating the historical time courses of mean polygenic scores and by leveraging ARGs.

As with other population-genetic methods for studying polygenic traits, results from our methods are accompanied by many qualifiers to interpretation (Novembre and Barton 2018).

In general, estimates arising from the methods presented here should not be viewed as necessarily reflecting the historical time course of trait values within a population. In our simulations, the association between genotype and phenotype was assumed to remain constant over time. In contrast, polygenic scores estimated in practice are better thought of as functions that encode present-day associations between genotype and phenotype. The genotype–phenotype association captured by a polygenic score may be due to causal effects of the included genotypes, but it might also be due to linkage disequilibrium between tag SNPs and ungenotyped causal SNPs, to indirect genetic effects (Kong *et al.* 2018; Walsh and Lynch 2018, Chap. 22), or to environmental effects that covary with genotype for other reasons. Any of these sources of association between genotype and phenotype might change as the environment and genetic background of the population change over time, causing time courses estimated by our method to deviate from the history of average trait values in the population. For example, the genetic architecture may change across the time period over which estimates are made, for example because of changes in linkage disequilibrium between tag loci and causal loci (Martin *et al.* 2017), or because loci that explained trait variation in the past have since fixed or been lost. Further, real-world estimates of effect size will be subject to noise in estimation and possibly bias due to stratification. [Even small amounts of stratification can seriously mislead tests for selection (Berg *et al.* 2018; Sohail *et al.* 2018).] Particularly in case-control studies, ascertainment biases may also lead to confounding between the evolutionary status of an allele (*i.e.*, derived or ancestral status) and power to detect trait associations (Chan *et al.* 2014). Also, importantly, changes in the environment may drive changes in mean levels of the trait that either amplify or oppose changes in population-mean polygenic scores, either via their direct effects or via gene–environment interactions. Finally, for most human traits, current polygenic scores explain relatively small proportions of the trait variance. In addition to all the issues above that might lead to bias in an inferred time course, using a weakly predictive polygenic score adds measurement error to the inferred time course—much of the variance in the trait will not be reflected by the polygenic score. This added variance would not be expected to increase the type-I error rate of tests of neutrality, but it will make the inferred time courses less likely to reflect the history of the trait closely.

Beyond these general caveats, the methods we propose here have several limitations that suggest directions for future work. The first three limitations concern outstanding statistical issues. First, the polygenic scores we estimate here are weighted sums of effect sizes estimated under additive models. Our variance estimates also assume that the loci incorporated in the polygenic score are in linkage equilibrium. Because our estimators work by estimating historical allele frequencies at the loci contributing to the polygenic score, they can in principle be adapted to estimate any function of allele frequencies, including trait predictions that account for dominance, epistasis, and linkage among loci, and also to single-locus trajectories. However, the strategies we use here for variance estimation and hypothesis testing may need to be modified for more general functions of allele frequencies. Second, any applications of these methods to real data will entail noise that is not accounted for by the variance estimates we propose. In particular, effect sizes will be estimated with error and so will the coalescent trees for sites included in the polygenic score. It will be important to incorporate these sources of variance in future estimates of sampling variation. Third, as suggested in the *Theory* section, the waiting-time and lineages-remaining estimators implicitly contain smoothing parameters. Fully characterizing the effects of these smoothing parameters and of alternative smoothing strategies—such as those used to smooth coalescent-based estimates of population-size history (Drummond *et al.* 2005; Minin *et al.* 2008)—will reveal the potential of these estimators, which have high variance in the forms in which they are used here.

The next three possible extensions are suggested by biological applications and by the coalescent framework in which we work. First, the theory we develop here is for a single population, but our setting within a coalescent framework suggests the possibility for extension to multiple populations, perhaps by developing multivariate analogs of our statistics within a coalescent-with-migration framework (Kaplan *et al.* 1991). Similarly, whereas we work with polygenic scores for a single trait, our methods can be extended to consider polygenic scores for multiple, correlated traits. In a similar vein, Berg *et al.* (2017) have recently extended the Q_X statistic to multiple correlated traits, drawing inspiration from the framework of Lande and Arnold (1983). Working with multiple traits will allow us to distinguish hypotheses in which a trait is directly subject to selection from hypotheses in which a correlated trait is the target of selection. Finally, because the coalescent framework explicitly represents the evolution of the sample backward in time, it will be productive to incorporate ancient samples.

The methods we propose are promising in part because they capitalize on illuminating descriptions of genetic variation that will become increasingly widely available. ARGs encode all the coalescence and recombination events reflected in the present-day sample, and thus are richly informative about the history of the sample's ancestors. Statistics computed from these ARGs have the potential to capture all the information about an allele's frequency time course that is available in a present-day sample. Approaches to traits that are based on sample ARGs will improve with the development of our understanding of the architecture of complex traits and of our ability to reconstruct ARGs.

Acknowledgments

Literature Cited

1000 Genomes Project ConsortiumAbecasis, G. R., A. Auton, L. D. Brooks, M. DePristo *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. https://doi.org/10.1038/nature11632

Barton, N. H., 1986 The maintenance of polygenic variation through a balance between mutation and stabilizing selection. Genet. Res. 47: 209–216. https://doi.org/10.1017/S0016672300023156

Bell, M. A., M. P. Travis, and D. M. Blouw, 2006 Inferring natural selection in a fossil threespine stickleback. Paleobiology 32: 562–577. https://doi.org/10.1666/05026.1

Berg, J. J., and G. Coop, 2014 A population genetic signal of polygenic adaptation. PLoS Genet. 10: e1004412. https://doi.org/10.1371/journal.pgen.1004412

Berg, J. J., and G. Coop, 2015 A coalescent model for a sweep of a unique standing variant. Genetics 201: 707–725. https://doi.org/10.1534/genetics.115.178962

Berg, J. J., X. Zhang, and G. Coop, 2017 Polygenic adaptation has impacted multiple anthropometric traits. bioRxiv 167551. DOI: https://doi.org/10.1101/167551.

Berg, J. J., A. Harpak, N. Sinnott-Armstrong, A. M. Joergensen, H. Mostafavi *et al.*, 2018 Reduced signal for polygenic adaptation of height in uk biobank. bioRxiv 354951. DOI: https://doi.org/10.1101/354951.

Berisa, T., and J. K. Pickrell, 2016 Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32: 283–285.

Bollback, J. P., T. L. York, and R. Nielsen, 2008 Estimation of $2N_e s$ from temporal allele frequency data. Genetics 179: 497–502. https://doi.org/10.1534/genetics.107.085019

Bürger, R., 2000 *The Mathematical Theory of Selection, Recombination, and Mutation* (Wiley Series in Mathematical & Computational Biology). Wiley, New York.

Cavalli-Sforza, L. L., I. Barrai, and A. W. F. Edwards, 1964 Analysis of human evolution under random genetic drift. Cold Spring Harb. Symp. Quant. Biol. 29: 9–20. https://doi.org/10.1101/SQB.1964.029.01.006

Chan, Y., E. T. Lim, N. Sandholm, S. R. Wang, A. J. McKnight *et al.*, 2014 An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. Am. J. Hum. Genet. 94: 437–452. https://doi.org/10.1016/j.ajhg.2014.02.006

Charlesworth, B., and D. Charlesworth, 2010 *Elements of Evolutionary Genetics*, Vol. 42. Roberts and Company Publishers, Greenwood Village, CO.

Chen, H., and K. Chen, 2013 Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. Genetics 194: 721–736. https://doi.org/10.1534/genetics.113.151522

Chen, H., and M. Slatkin, 2013 Inferring selection intensity and allele age from multilocus haplotype structure. G3 (Bethesda) 3: 1429–1442. https://doi.org/10.1534/g3.113.006197

Cook, L. M., G. S. Mani, and M. E. Varley, 1986 Postindustrial melanism in the peppered moth. Science 231: 611–613. https://doi.org/10.1126/science.231.4738.611

Coop, G., and R. C. Griffiths, 2004 Ancestral inference on gene trees under selection. Theor. Popul. Biol. 66: 219–232. https://doi.org/10.1016/j.tpb.2004.06.006

Crawford, N. G., D. E. Kelly, M. E. B. Hansen, M. H. Beltrame, S. Fan *et al.*, 2017 Loci associated with skin pigmentation identified in African populations. Science 358: eaan8433. https://doi.org/10.1126/science.aan8433

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and vcftools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus, 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22: 1185–1192. https://doi.org/10.1093/molbev/msi103

Ewens, W. J., 2004 *Mathematical Population Genetics I. Theoretical Introduction*, Ed. 2. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-21822-9

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Field, Y., E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton *et al.*, 2016 Detection of human adaptation during the past 2000 years. Science 354: 760–764. https://doi.org/10.1126/science.aag0776

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent selective sweeps in north american Drosophila melanogaster show signatures of soft sweeps. PLoS Genet. 11: e1005004. https://doi.org/10.1371/journal.pgen.1005004

Gelman, A., and H. Stern, 2006 The difference between "significant" and "not significant" is not itself statistically significant. Am. Stat. 60: 328–331. https://doi.org/10.1198/000313006X152649

Gingerich, P. D., 1983 Rates of evolution: effects of time and temporal scaling. Science 222: 159–161. https://doi.org/10.1126/science.222.4620.159

Grant, P. R., and B. R. Grant, 2002 Unpredictable evolution in a 30-year study of Darwin's finches. Science 296: 707–711. https://doi.org/10.1126/science.1070315

Griffiths, R. C., 1984 Asymptotic line-of-descent distributions. J. Math. Biol. 21: 67–75. https://doi.org/10.1007/BF00275223

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution*. Springer, New York.

Ho, S. Y. W., and B. Shapiro, 2011 Skyline-plot methods for estimating demographic history from nucleotide sequences. Mol. Ecol. Resour. 11: 423–434. https://doi.org/10.1111/j.1755-0998.2011.02988.x

Hudson, R. R., 2002 Generating samples under a wright–fisher neutral model of genetic variation. Bioinformatics 18: 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Hudson, R. R., and N. L. Kaplan, 1988   The coalescent process in models with selection and recombination. Genetics 120: 831–840.

Jewett, E. M., and N. A. Rosenberg, 2014   Theory and applications of a deterministic approximation to the coalescent model. Theor. Popul. Biol. 93: 14–29. https://doi.org/10.1016/j.tpb.2013.12.007

Kaplan, N., R. R. Hudson, and M. Iizuka, 1991   The coalescent process in models with selection, recombination and geographic subdivision. Genet. Res. 57: 83–91. https://doi.org/10.1017/S0016672300029074

Kaplan, N. L., T. Darden, and R. R. Hudson, 1988   The coalescent process in models with selection. Genetics 120: 819–829.

Keinan, A., and A. G. Clark, 2012   Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740–743. https://doi.org/10.1126/science.1217283

Kong, A., G. Thorleifsson, M. L. Frigge, B. J. Vilhjalmsson, A. I. Young et al., 2018   The nature of nurture: effects of parental genotypes. Science 359: 424–428. https://doi.org/10.1126/science.aan6877

Lande, R., 1976   Natural selection and random genetic drift in phenotypic evolution. Evolution 30: 314–334. https://doi.org/10.1111/j.1558-5646.1976.tb00911.x

Lande, R., and S. J. Arnold, 1983   The measurement of selection on correlated characters. Evolution 37: 1210–1226. https://doi.org/10.1111/j.1558-5646.1983.tb00236.x

Lee, K. M., and G. Coop, 2017   Distinguishing among modes of convergent adaptation using population genomic data. Genetics 207: 1591–1619.

Li, H., 2011   Tabix: fast retrieval of sequence features from generic tab-delimited files. Bioinformatics 27: 718–719. https://doi.org/10.1093/bioinformatics/btq671

Ludwig, A., M. Pruvost, M. Reissmann, N. Benecke, G. A. Brockmann et al., 2009   Coat color variation at the beginning of horse domestication. Science 324: 485. https://doi.org/10.1126/science.1172750

MacFadden, B. J., 2005   Fossil horses–evidence for evolution. Science 307: 1728–1730. https://doi.org/10.1126/science.1105458

Martin, A. R., C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale et al., 2017   Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. 100: 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

Maruvka, Y. E., N. M. Shnerb, Y. Bar-Yam, and J. Wakeley, 2011   Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. Mol. Biol. Evol. 28: 1617–1631. https://doi.org/10.1093/molbev/msq331

Mathieson, I., I. Lazaridis, N. Rohland, S. Mallick, N. Patterson et al., 2015   Genome-wide patterns of selection in 230 ancient eurasians. Nature 528: 499–503. https://doi.org/10.1038/nature16152

Minin, V. N., E. W. Bloomquist, and M. A. Suchard, 2008   Smooth skyride through a rough skyline: bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25: 1459–1471. https://doi.org/10.1093/molbev/msn090

Mirzaei, S., and Y. Wu, 2016   Rent+: an improved method for inferring local genealogical trees from haplotypes with recombination. Bioinformatics 33: 1021–1030.

Neale Lab, 2017   Rapid gwas of thousands of phenotypes for 337,000 samples in the uk biobank. Available at: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank. Accessed: November 13, 2018.

Nicholson, G., A. V. Smith, F. Jónsson, O. Gústafsson, K. Stefánsson et al., 2002   Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Stat. Soc. B 64: 695–715. https://doi.org/10.1111/1467-9868.00357

Novembre, J., and N. H. Barton, 2018   Tread lightly interpreting polygenic tests of selection. Genetics 208: 1351–1355. https://doi.org/10.1534/genetics.118.300786

Palacios, J. A., J. Wakeley, and S. Ramachandran, 2015   Bayesian nonparametric inference of population size changes from sequential genealogies. Genetics 201: 281–304. https://doi.org/10.1534/genetics.115.177980

Palamara, P. F., J. Terhorst, Y. S. Song, and A. L. Price, 2018   High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nat. Genet. 50: 1311–1317. https://doi.org/10.1038/s41588-018-0177-x

Paradis, E., J. Claude, and K. Strimmer, 2004   APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290. https://doi.org/10.1093/bioinformatics/btg412

Przeworski, M., G. Coop, and J. D. Wall, 2005   The signature of positive selection on standing genetic variation. Evolution 59: 2312–2323. https://doi.org/10.1554/05-273.1

Pybus, O. G., A. Rambaut, and P. H. Harvey, 2000   An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155: 1429–1437.

Racimo, F., J. J. Berg, and J. K. Pickrell, 2018   Detecting polygenic adaptation in admixture graphs. Genetics 208: 1565–1584. https://doi.org/10.1534/genetics.117.300489

Rasmussen, M. D., M. J. Hubisz, I. Gronau, and A. Siepel, 2014 Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10: e1004342. https://doi.org/10.1371/journal.pgen.1004342

R Core Team, 2013   R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Robinson, M. R., G. Hemani, C. Medina-Gomez, M. Mezzavilla, T. Esko et al., 2015   Population genetic differentiation of height and body mass index across europe. Nat. Genet. 47: 1357–1362. https://doi.org/10.1038/ng.3401

Ronen, R., N. Udpa, E. Halperin, and V. Bafna, 2013   Learning natural selection from the site frequency spectrum. Genetics 195: 181–193. https://doi.org/10.1534/genetics.113.152587

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter et al., 2002   Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837. https://doi.org/10.1038/nature01140

Schraiber, J. G., S. N. Evans, and M. Slatkin, 2016   Bayesian inference of natural selection from allele frequency time series. Genetics 203: 493–511. https://doi.org/10.1534/genetics.116.187278

Simons, Y. B., K. Bullaughey, R. R. Hudson, and G. Sella, 2018   A population genetic interpretation of gwas findings for human quantitative traits. PLoS Biol. 16: e2002985. https://doi.org/10.1371/journal.pbio.2002985

Slatkin, M., 2001   Simulating genealogies of selected alleles in a population of variable size. Genet. Res. 78: 49–57. https://doi.org/10.1017/S0016672301005183

Smith, J. M., and J. Haigh, 1974   The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35. https://doi.org/10.1017/S0016672300014634

Sohail, M., R. M. Maier, A. Ganna, A. Bloemendal, A. R. Martin et al., 2018   Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. bioRxiv 355057. DOI: https://doi.org/10.1101/355057.

Strimmer, K., and O. G. Pybus, 2001   Exploring the demographic history of dna sequences using the generalized skyline plot. Mol. Biol. Evol. 18: 2298–2305. https://doi.org/10.1093/oxfordjournals.molbev.a003776

Stuart, A., and J. K. Ord, 1987   Kendall's Advanced Theory of Statistics, Distribution Theory, Vol. I. Oxford University Press, London.

Stulp, G., and L. Barrett, 2016   Evolutionary perspectives on human height variation. Biol. Rev. Camb. Philos. Soc. 91: 206–234. https://doi.org/10.1111/brv.12165

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton et al., 2015   Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12: e1001779. https://doi.org/10.1371/journal.pmed.1001779

Tajima, F., 1989   Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics 123: 585–595.

Tavaré, S., 1984   Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26: 119–164. https://doi.org/10.1016/0040-5809(84)90027-3

Turchin, M., C. W. K. Chiang, C. D. Palmer, S. Sankararaman, D. Reich et al., 2012   Evidence of widespread selection on standing variation in europe at height-associated SNPs. Nat. Genet. 44: 1015–1019. https://doi.org/10.1038/ng.2368

Turelli, M., J. H. Gillespie, and R. Lande, 1988   Rate tests for selection on quantitative characters during macroevolution and microevolution. Evolution 42: 1085–1089. https://doi.org/10.1111/j.1558-5646.1988.tb02526.x

Uricchio, L. H., H. C. Kitano, A. Gusev, and N. A. Zaitlen, 2018   An evolutionary compass for elucidating selection mechanisms shaping complex traits. bioRxiv 173815. DOI: https://doi.org/10.1101/173815.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006   A map of recent positive selection in the human genome. PLoS Biol. 4: e72 [erratum: PLoS Biol. 4: e154; corrigenda: PLoS Biol. 5: e147 (2007)].

Walsh, B., and M. Lynch, 2018   Evolution and Selection of Quantitative Traits. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198830870.001.0001

Watterson, G., 1975   On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Whitlock, M. C., 2008   Evolutionary inference from Q_{ST}. Mol. Ecol. 17: 1885–1896. https://doi.org/10.1111/j.1365-294X.2008.03712.x

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers et al., 2014   Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46: 1173–1186. https://doi.org/10.1038/ng.3097

Communicating editor: R. Nielsen

Appendix A: The Relationship Between Coalescent Rates and Phenotypic Selection

In this appendix, we describe the consequences of directional and stabilizing selection on a trait for coalescence rates at loci associated with the trait.

We begin by considering the general effects of selection on coalescence before considering the specific cases of directional and stabilizing selection. If the frequency of the allele of interest at locus i at time t is $p_i(t)$ in a population of $2N$ chromosomes, then the probability of any one of $\binom{n}{2}$ pairs coalescing in the preceding generation is

$$\lambda(t-1) = \binom{n}{2} \frac{1}{2Np_i(t-1)} \tag{A1}$$

(Hudson and Kaplan 1988). Writing the frequency in the preceding generation as $p_i(t-1) = p_i(t) - \Delta_i(t)$, then

$$\lambda(t-1) = \binom{n}{2} \frac{1}{2N[p_i(t) - \Delta_i(t)]} \approx \binom{n}{2} \frac{1}{2Np_i(t)} [1 + \Delta_i(t)/p_i(t)] \tag{A2}$$

assuming that $\Delta_i(t)$ is small so that higher-order terms in the Taylor expansion can be ignored. We can decompose the change in frequency from the preceding generation into $\Delta_i(t) = \Delta_i^{(N)}(t) + \Delta_i^{(S)}(t)$, the contributions of random genetic drift (N) and selection (S), respectively. Then, taking the expectation over the frequency in the preceding generation,

$$E[\lambda(t-1)] \approx \binom{n}{2} \frac{1}{2Np_i(t)} \{1 + E[\Delta_i(t)/p_i(t)]\}$$
$$= \binom{n}{2} \frac{1}{2Np_i(t)} \left[1 + \Delta_i^{(S)}(t) \Big/ p_i(t)\right] \tag{A3}$$

because $E[\Delta_i^{(N)}(t)] = 0$ and $\Delta_i^{(S)}(t)$ is the deterministic change due to selection.

If the allele of interest affects a trait that is under selection, then the form of $\Delta_i^{(S)}(t)$ depends on the allele's role in the trait's architecture and on the specifics of selection on the trait. The simplest case is one in which directional selection acts on a phenotype whose genetic architecture is purely additive. Denote the selection gradient—that is, the slope of fitness regressed on phenotype—as α. If the effect size of the ith locus on the phenotype is β_i, and the loci that affect the phenotype are unlinked, then

$$\Delta_i^{(S)}(t) = \alpha\beta_i p_i(1 - p_i) \tag{A4}$$

(equation 3.17 in Charlesworth and Charlesworth 2010). Therefore, the expected coalescent rate over the possible allele-frequency changes in the preceding generation is

$$E_\Delta[\lambda(t-1)] \approx \binom{n}{2} \frac{1}{2Np_i(t)} [1 + \alpha\beta_i(1 - p_i)]. \tag{A5}$$

Therefore, recent directional selection increases the coalescence rate among alleles that move the phenotype in the direction in which selection acts. For example, if larger trait values have been selected recently, the coalescence rate will be higher among alleles associated with larger values of the trait and lower among alleles associated with smaller values of the trait. This is true regardless of the frequency of the allele.

If the phenotype is subject to both directional selection and stabilizing selection, the allele frequency matters. For example, one common and reasonably general model incorporating directional and stabilizing selection is the quadratic optimum model, where the fitness of an individual with phenotype G is $W(G) = 1 - s(P_O - G)^2$. There is directional selection on the phenotype when the population mean ($\bar{G}$) deviates from the optimum $\delta = P_O - \bar{G} \neq 0$. (When $\delta > 0$, the population mean is below the

optimum, and selection will favor larger trait values.) Under this model,

$$\Delta_i^{(S)}(t) = \underbrace{2s\beta_i\delta p_i(1-p_i)}_{\text{Directional selection}} + \underbrace{s\beta_i^2 p_i(1-p_i)(2p_i-1)}_{\text{Stabilizing selection}}; \tag{A6}$$

see Barton (1986), Bürger (2000), and Simons *et al.* (2018) for a recent discussion. Note the similarity of the directional selection term—with the selection gradient now controlled by $s\delta$—to Equation 20, resulting in similar coalescent rates to Equation 21. The Δ term for stabilizing selection matches that of a disruptive-selection model with an unstable equilibrium at $1/2 - \delta/\beta_i$.

Incorporating both directional and stabilizing selection, the expected coalescent rate is approximately

$$E_\Delta[\lambda(t-1)] \approx \binom{n}{2}\frac{1}{2Np_i(t)}\left[1 + \underbrace{2s\beta_i\delta(1-p_i)}_{\text{Directional selection}} + \underbrace{s\beta_i^2(1-p_i)(2p_i-1)}_{\text{Stabilizing selection}}\right]. \tag{A7}$$

The stabilizing selection term decreases the coalescent rate of whichever allele has frequency $<1/2$. The average coalescent rate among "up" alleles ($\beta_i > 0$) depends on the value of $E_\Delta[\lambda(t-1)]$ averaged over the frequency distribution of p_i. Thus, when directional selection acts on the phenotype, we again expect an increased coalescence rate among alleles whose effects have been favored. If the population is at the optimum, *i.e.*, $\delta = 0$, then averaging across loci there is no net effect on the coalescent rates if there is a symmetric distribution of $p_i \sim 1/2$ (across all effect sizes). This symmetric distribution will result when the population is at equilibrium and there is no mutational bias (*i.e.*, there is no difference in the mutational input of up and down alleles).

Appendix B: A Bayesian View of the Proportion-of-Lineages Estimator

To arrive at a Bayesian view of the estimator proposed in Equation 4, define the frequency of an allele of interest at a biallelic locus at time t in the past as $p_i(t) = x$. If the allele of interest is equally likely to be derived or ancestral, a reasonable prior distribution would be proportional to the limit (as the mutation rates approach 0) of the stationary distribution of the neutral diffusion with mutation, or $x^{-1}(1-x)^{-1}$ (equation 5.70 in Ewens 2004). Assume that the two alleles at the locus have experienced equal reproductive success between the present and time t—that is, assume that the locus has evolved neutrally in the time interval. In that case, the lineages ancestral to a random sample of chromosomes in the present are themselves a random sample from the population at time t, and if there are r ancestral lineages at time t, then the number carrying the allele of interest is a binomial(r,x) random variable. If we observe that j out of r ancestral lineages at time t carry the allele of interest, then the posterior distribution for x is proportional to

$$f_x(x|r,j) \propto \frac{\binom{r}{j}x^j(1-x)^{r-j}}{x(1-x)} = \binom{r}{j}x^{j-1}(1-x)^{r-j-1}.$$

Noticing that $x^{j-1}(1-x)^{r-j-1}$ is the kernel of a $\beta(j, r-j)$ distribution gives the posterior,

$$f_x(x|r,j) = \frac{(r-1)!}{(j-1)!(r-j-1)!}x^{j-1}(1-x)^{r-j-1}. \tag{B1}$$

Equation B1 suggests the posterior mean of $p_i(t)$ as an estimate of the frequency (when $r \geq 2$ and $j \geq 1$), which is equal to Equation 4. The posterior variance is

$$\text{Var}\left[p_i(t)|p_i(0) = p, n_i^{(t)} = j, m_i^{(t)} = r-j\right] = \frac{j(r-j)}{r^2(r+1)}, \tag{B2}$$

which differs from the binomial sampling variance in Equation 5 by a factor of $r/(r+1)$.

It may be interesting to note that there is a connection between this Bayesian view of the estimator and the neutral Wright–Fisher diffusion process, reviewed by Tavaré (1984). Under a neutral Wright–Fisher diffusion model without mutation, the probability that an allele frequency changes forward in time from $p_i(t) = p$ to $p_i(0) = x$ in time t can be written as (equation 7.18 in Tavaré 1984)

$$P[p_i(0) = x | p_i(t) = p] = \sum_{r=2}^{\infty} g_r(t) \sum_{j=1}^{r-1} \binom{r}{j} p^j (1-p)^{r-j} \binom{r-1}{j-1} x^{j-1} (1-x)^{r-j-1}, \tag{B3}$$

here $g_r(t)$ is the probability that r lineages of an initially infinite number of lineages survive to time t. Because the neutral diffusion is time reversible (Ewens 2004),

$$P[p_i(0) = x | p_i(t) = p] = P[p_i(t) = x | p_i(0) = p],$$

and the expression on the right side of Equation B3 is thus also an expression for the allele frequency in the past given the allele frequency in the present, $P[p_i(t) = x | p_i(0) = p]$ One way to interpret Equation B3 is that if r lineages survive to time t, then the number j of lineages carrying a focal allele at time t is a binomial(r, p) random variable. Then, conditional on j, the frequency $p_i(t)$ has a $\beta(j, r-j)$ distribution.

Appendix C: Mathematical Details for the Waiting-Time and Lineages-Remaining Estimators

Taylor-Series Expansions for Expectations and Variances of Ratios

Both the waiting-time and lineages-remaining estimators take the form

$$\hat{p} = f(\hat{N}, \hat{D}) = \frac{\hat{N}}{\hat{D}}, \tag{C1}$$

where $\hat{D} = \hat{N} + \hat{M}$, and $\hat{N}$ and $\hat{M}$ are population-size estimators that are independent of each other. We assume that $P(\hat{D} \leq 0) = 0$. Equation C1 is equivalent to Equation 7 from the main text, with the subscripts and time notation removed for compactness.

In this subsection we present general Taylor-series approximations for the expectation and variance of the estimator in Equation C1 in terms of the expectation and variance of $\hat{N}$ and $\hat{M}$. Similar presentations are available in other references (sections 10.5–10.6 in Stuart and Ord 1987), but we present the argument here for completeness.

The first-order Taylor expansion of Equation C1 evaluated at $(\mu_{\hat{N}}, \mu_{\hat{D}})$ is

$$f(\hat{N}, \hat{D}) = f(\mu_{\hat{N}}, \mu_{\hat{D}}) + f'_{\hat{N}}(\mu_{\hat{N}})(\hat{N} - \mu_{\hat{N}}) + f'_{\hat{D}}(\mu_{\hat{D}})(\hat{D} - \mu_{\hat{D}}) + Z, \tag{C2}$$

where $f'_{\hat{N}}$ and $f'_{\hat{D}}$ are the partial derivatives of f, and Z represents the contribution of higher-order terms. The expectation of $f(\hat{N}, \hat{D})$ is therefore

$$\begin{aligned}
E[f(\hat{N}, \hat{D})] &= E\left[f(\mu_{\hat{N}}, \mu_{\hat{D}}) + f'_{\hat{N}}(\mu_{\hat{N}})(\hat{N} - \mu_{\hat{N}}) + f'_{\hat{D}}(\mu_{\hat{D}})(\hat{D} - \mu_{\hat{D}}) + Z \right] \\
&= f(\mu_{\hat{N}}, \mu_{\hat{D}}) + f'_{\hat{N}}(\mu_{\hat{N}}) E(\hat{N} - \mu_{\hat{N}}) + f'_{\hat{D}}(\mu_{\hat{D}}) E(\hat{D} - \mu_{\hat{D}}) + E(Z),
\end{aligned}$$

with the second step coming from the linearity of expectation and the fact that $f(\mu_{\hat{N}}, \mu_{\hat{D}})$ is not random. Choosing $\mu_{\hat{N}}$ and $\mu_{\hat{D}}$ to be equal to the expectations of $\hat{N}$ and $\hat{D}$ gives

$$\begin{aligned}
E[f(\hat{N}, \hat{D})] &= f(\mu_{\hat{N}}, \mu_{\hat{D}}) + f'_{\hat{N}}(\mu_{\hat{N}}) 0 + f'_{\hat{D}}(\mu_{\hat{D}}) 0 + E(Z) \\
&= f(\mu_{\hat{N}}, \mu_{\hat{D}}) + E(Z) \\
&\approx f(\mu_{\hat{N}}, \mu_{\hat{D}}) = \frac{\mu_{\hat{N}}}{\mu_{\hat{D}}},
\end{aligned} \tag{C3}$$

where the approximation depends on the higher-order terms represented by Z being small. Equation C3 underlies the claim in the main text that the waiting-time estimator is approximately, to first order, unbiased (under the unrealistic assumption that the number of carriers of each allele is constant between coalescent events).

To obtain the approximate variance of the estimator in Equation C1, we use Equation C3 and note that

$$
\begin{aligned}
\mathrm{Var}[f(\hat{N},\hat{D})] &= \mathrm{E}\Big(\{f(\hat{N},\hat{D})-\mathrm{E}[f(\hat{N},\hat{D})]\}^2\Big) \\
&\approx \mathrm{E}\Big\{[f(\hat{N},\hat{D})-f(\mu_{\hat{N}},\mu_{\hat{D}})]^2\Big\}.
\end{aligned}
\tag{C4}
$$

Substituting the first-order Taylor expansion in Equation C2 for $f(\hat{N},\hat{D})$, ignoring the higher-order terms represented by Z, gives

$$
\begin{aligned}
\mathrm{Var}[f(\hat{N},\hat{D})] &\approx \mathrm{E}\Big\{\Big[f(\mu_{\hat{N}},\mu_{\hat{D}})+f'_{\hat{N}}(\mu_{\hat{N}})(\hat{N}-\mu_{\hat{N}})+f'_{\hat{D}}(\mu_{\hat{D}})(\hat{D}-\mu_{\hat{D}})-f(\mu_{\hat{N}},\mu_{\hat{D}})\Big]^2\Big\} \\
&= \mathrm{E}\Big\{[f'_{\hat{N}}(\mu_{\hat{N}})(\hat{N}-\mu_{\hat{N}})+f'_{\hat{D}}(\mu_{\hat{D}})(\hat{D}-\mu_{\hat{D}})]^2\Big\} \\
&= \mathrm{E}\Big[f'_{\hat{N}}(\mu_{\hat{N}})^2(\hat{N}-\mu_{\hat{N}})^2 + 2f'_{\hat{N}}(\mu_{\hat{N}})f'_{\hat{D}}(\mu_{\hat{D}})(\hat{N}-\mu_{\hat{N}})(\hat{D}-\mu_{\hat{D}})+f'_{\hat{D}}(\mu_{\hat{D}})^2(\hat{D}-\mu_{\hat{D}})^2\Big] \\
&= f'_{\hat{N}}(\mu_{\hat{N}})^2\mathrm{E}\Big[(\hat{N}-\mu_{\hat{N}})^2\Big] + 2f'_{\hat{N}}(\mu_{\hat{N}})f'_{\hat{D}}(\mu_{\hat{D}})\mathrm{E}[(\hat{N}-\mu_{\hat{N}})(\hat{D}-\mu_{\hat{D}})]+f'_{\hat{D}}(\mu_{\hat{D}})^2\mathrm{E}\Big[(\hat{D}-\mu_{\hat{D}})^2\Big] \\
&= f'_{\hat{N}}(\mu_{\hat{N}})^2\mathrm{Var}(\hat{N}) + 2f'_{\hat{N}}(\mu_{\hat{N}})f'_{\hat{D}}(\mu_{\hat{D}})\mathrm{Cov}(\hat{N},\hat{D})+f'_{\hat{D}}(\mu_{\hat{D}})^2\mathrm{Var}(\hat{D}).
\end{aligned}
\tag{C5}
$$

The last step comes from the definitions of variance and covariance and the fact that we choose $\mu_{\hat{N}}$ and $\mu_{\hat{D}}$ to be equal to the expectations of $\hat{N}$ and $\hat{D}$. In our case, $f(\hat{N},\hat{D}) = \hat{N}/\hat{D}$, implying that $f'_{\hat{N}}(\mu_{\hat{N}}) = 1/\mu_{\hat{D}}$ and $f'_{\hat{D}}(\mu_{\hat{D}}) = -\mu_{\hat{N}}/\mu_{\hat{D}}^2$. Plugging these values into Equation C5 gives

$$
\mathrm{Var}[f(\hat{N},\hat{D})] \approx \frac{\mathrm{Var}(\hat{N})}{\mu_{\hat{D}}^2} - \frac{2\mu_{\hat{N}}\mathrm{Cov}(\hat{N},\hat{D})}{\mu_{\hat{D}}^3} + \frac{\mu_{\hat{N}}^2\mathrm{Var}(\hat{D})}{\mu_{\hat{D}}^4},
\tag{C6}
$$

with the useful alternative form

$$
\mathrm{Var}[f(\hat{N},\hat{D})] \approx \left(\frac{\mu_{\hat{N}}}{\mu_{\hat{D}}}\right)^2\left[\frac{\mathrm{Var}(\hat{N})}{\mu_{\hat{N}}^2} - \frac{2\mathrm{Cov}(\hat{N},\hat{D})}{\mu_{\hat{N}}\mu_{\hat{D}}} + \frac{\mathrm{Var}(\hat{D})}{\mu_{\hat{D}}^2}\right].
\tag{C7}
$$

Finally, in our setting, $\hat{D} = \hat{N} + \hat{M}$, and $\hat{N}$ and $\hat{M}$ are independent conditional on the true but unknown allele-frequency time course (Kaplan *et al.* 1988; Coop and Griffiths 2004). Conditional independence gives $\mu_{\hat{D}} = \mu_{\hat{N}} + \mu_{\hat{M}}$, $\mathrm{Var}(\hat{D}) = \mathrm{Var}(\hat{N}) + \mathrm{Var}(\hat{M})$, and $\mathrm{Cov}(\hat{N},\hat{D}) = \mathrm{Var}(\hat{N})$. Using these identities in Equation C7 gives

$$
\mathrm{Var}(\hat{p}) \approx \left(\frac{\mu_{\hat{N}}}{\mu_{\hat{N}}+\mu_{\hat{M}}}\right)^2\left[\frac{\mathrm{Var}(\hat{N})}{\mu_{\hat{N}}^2} - \frac{2\mathrm{Var}(\hat{N})}{\mu_{\hat{N}}(\mu_{\hat{N}}+\mu_{\hat{M}})} + \frac{\mathrm{Var}(\hat{N}) + \mathrm{Var}(\hat{M})}{(\mu_{\hat{N}}+\mu_{\hat{M}})^2}\right].
\tag{C8}
$$

Approximate Variance of the Waiting-Time Estimator

As noted in the main text, the waiting-time estimator has the form

$$
\widehat{p_i(t)} = \frac{\widehat{N_i(t)}}{\widehat{N_i(t)} + \widehat{M_i(t)}},
$$

where

$$
\widehat{N_i(t)} = \frac{Y}{2(1/[n_i-\ell]-1/n_i)}.
$$

Y is the waiting time associated with a set of ℓ coalescent events on the allele of interest's background that envelop time t, and n_i is the number of allele-of-interest lineages that exist before any of the ℓ coalescent events have occurred. $\widehat{M_i(t)}$ is analogous, substituting in a waiting time and number of lineages from the other allele's background, and possibly a different value of ℓ.

It is also noted in the main text (Equation 9) that the waiting time Y has variance

$$\text{Var}(Y) = 4N_i(t)^2 \sum_{w=n_i-\ell+1}^{n_i} \frac{1}{w^2(w-1)^2}.$$

Thus, the variance of $\widehat{N_i(t)}$ is

$$\text{Var}[\widehat{N_i(t)}] = N_i(t)^2 \frac{\sum_{w=n_i-\ell+1}^{n_i} \frac{1}{w^2(w-1)^2}}{[1/(n_i-\ell)-1/n_i]^2}. \tag{C9}$$

If $\ell = 1$, Equation C9 reduces to $\text{Var}[\widehat{N_i(t)}] = N_i(t)^2$. If $\ell \ll n_i$, then

$$\sum_{w=n_i-\ell+1}^{n_i} \frac{1}{w^2(w-1)^2} \approx \frac{\ell}{n_i^2(n_i-1)^2},$$

and $\text{Var}(\widehat{N_i(t)}) \approx N_i(t)^2/\ell$.

By Equation C8 and the fact that $\widehat{N_i(t)}$ and $\widehat{M_i(t)}$ are unbiased, the approximate variance of the waiting-time estimator is

$$\text{Var}[\widehat{p_i(t)}] \approx \frac{N_i(t)^2}{[N_i(t)+M_i(t)]^2} \left\{ \frac{\sigma_N^2}{N_i(t)^2} - \frac{2\sigma_N^2}{N_i(t)[N_i(t)+M_i(t)]} + \frac{\sigma_N^2+\sigma_M^2}{[N_i(t)+M_i(t)]^2} \right\}, \tag{C10}$$

where σ_N^2 and σ_M^2 are the variances of $N_i(t)$ and $M_i(t)$, respectively. If $\ell = 1$ is used to estimate both $N_i(t)$ and $M_i(t)$, then $\text{Var}[\widehat{N_i(t)}] = N_i(t)^2$, $\text{Var}[\widehat{M_i(t)}] = M_i(t)^2$, and Equation C10 reduces to

$$\text{Var}[\widehat{p_i(t)}] \approx \frac{N_i(t)^2}{[N_i(t)+M_i(t)]^2} \left\{ \frac{N_i(t)^2}{N_i(t)^2} - \frac{2N_i(t)^2}{N_i(t)[N_i(t)+M_i(t)]} + \frac{N_i(t)^2+M_i(t)^2}{[N_i(t)+M_i(t)]^2} \right\}$$

$$= \frac{N_i(t)^2}{[N_i(t)+M_i(t)]^2} \left\{ 1 - \frac{2N_i(t)}{N_i(t)+M_i(t)} + \frac{N_i(t)^2+M_i(t)^2}{[N_i(t)+M_i(t)]^2} \right\}$$

$$= \frac{N_i(t)^2}{[N_i(t)+M_i(t)]^2} \left\{ 1 + \frac{M_i(t)^2-N_i(t)^2-2N_i(t)M_i(t)}{[N_i(t)+M_i(t)]^2} \right\} \tag{C11}$$

$$= \frac{N_i(t)^2}{[N_i(t)+M_i(t)]^2} \left\{ 1 + \frac{2M_i(t)^2}{[N_i(t)+M_i(t)]^2} - 1 \right\}$$

$$= 2p_i(t)^2[1-p_i(t)]^2.$$

If $N_i(t)$ and $M_i(t)$ are estimated using the same value of ℓ, and ℓ is much smaller than the starting number of lineages for both loci, so that $\sigma_N^2 \approx N_i(t)^2/\ell$ and $\sigma_M^2 \approx M_i(t)^2/\ell$, then Equation C9 can be rewritten as

$$\text{Var}[\widehat{p_i(t)}] \approx \frac{2}{\ell} p_i(t)^2[1-p_i(t)]^2.$$

Approximate Variance of the Lineages-Remaining Estimator

In the main text, we propose to estimate $N_i(t)$ as

$$\widehat{N_i(t)} = \frac{\Delta t}{2\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)}-1}\right]\right\}},$$

(C12)

where Δt is the number of generations elapsed between two ends of a predefined time interval, $n_i(0)$ is the number of lineages of the allele of interest at the end of the time interval closer to the present, and $n_i(t)$ is the number of lineages of the allele of interest at the end of the interval further into the past. The only random component of $\widehat{N_i(t)}$ is $n_i(t)$, which depends on the number of coalescent events that occur during the interval. This estimator of $N_i(t)$ is leads to an estimator for $p_i(t)$, the allele frequency of interest,

$$\widehat{p_i(t)} = \frac{\log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right] - \log\left[\frac{m_i^{(0)}}{m_i^{(0)}-1}\right]}{\log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right] - \log\left[\frac{m_i^{(0)}}{m_i^{(0)}-1}\right] + \log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)}-1}\right]}.$$

(C13)

(The expression in Equation C13 is the same as in Equation 14.)

To obtain an approximate sampling variance for $\widehat{p_i(t)}$, we use Equation C8, making the substitutions

$$\hat{N} = \log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right] - \log\left[\frac{m_i^{(0)}}{m_i^{(0)}-1}\right]$$

and

$$\hat{M} = \log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)}-1}\right].$$

To use Equation C8, we need (at least approximate) expectations and variances for both these terms. To compute an approximate expectation, we replace $m_i^{(t)}$ and $n_i^{(t)}$ with their expectations. Taking the case of $n_i^{(t)}$, notice that by Equation 12,

$$\log\left\{\frac{\mathrm{E}\left[n_i^{(t)}\right]}{\mathrm{E}\left[n_i^{(t)}\right]-1}\right\} = \log\left\{\frac{\dfrac{n_i^{(0)}}{n_i^{(0)} + \left[1-n_i^{(0)}\right]e^{-\tau/2}}}{\dfrac{n_i^{(0)}}{n_i^{(0)} + \left[1-n_i^{(0)}\right]e^{-\tau/2}} - 1}\right\}$$

$$= \log\left\{\frac{\dfrac{n_i^{(0)}}{n_i^{(0)} + \left[1-n_i^{(0)}\right]e^{-\tau/2}}}{\dfrac{n_i^{(0)} - n_i^{(0)} - \left[1-n_i^{(0)}\right]e^{-\tau/2}}{n_i^{(0)} + \left[1-n_i^{(0)}\right]e^{-\tau/2}}}\right\}$$

$$= \log\left[\frac{n_i(0)e^{\tau/2}}{n_i(0)-1}\right] = \log\left[\frac{n_i(0)}{n_i(0)-1}\right] + \frac{\tau}{2}.$$

(C14)

In Equation C14, τ represents the coalescent time passed on the allele of interest's background, or $\Delta t / N_i(t)$, and so

$$\mathrm{E}\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)}-1}\right]\right\} \approx \frac{\Delta t}{2N_i(t)}, \tag{C15}$$

and analogously for $m_i^{(t)}$,

$$\mathrm{E}\left\{\log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right] - \log\left[\frac{m_i^{(0)}}{m_i^{(0)}-1}\right]\right\} \approx \frac{\Delta t}{2M_i(t)}. \tag{C16}$$

Next we compute the approximate variances of these terms. Because $n_i^{(0)}$ is a constant,

$$\mathrm{Var}\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right] - \log\left[\frac{n_i^{(0)}}{n_i^{(0)}-1}\right]\right\} = \mathrm{Var}\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right]\right\}. \tag{C17}$$

By a first-order Taylor approximation, the variance is approximately

$$\mathrm{Var}\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right]\right\} \approx \left(f'\left\{\mathrm{E}[n_i^{(t)}]\right\}\right)^2 \mathrm{Var}\left[n_i^{(t)}\right], \tag{C18}$$

where $f'\{\mathrm{E}[n_i^{(t)}]\}$ is the derivative of the right side of Equation C17 with respect to $n_i^{(t)}$, evaluated at $\mathrm{E}[n_i^{(t)}]$ (The argument for this approximation is entirely parallel to the one in Equation C5, but for a function of a single random variable.) The derivative of the right side of Equation C17 with respect to $n_i^{(t)}$, evaluated at $\mathrm{E}[n_i^{(t)}]$, is

$$f'\left\{\mathrm{E}\left[n_i^{(t)}\right]\right\} = \frac{1}{\mathrm{E}\left[n_i^{(t)}\right]\left\{\mathrm{E}\left[n_i^{(t)}\right]-1\right\}}. \tag{C19}$$

To write the approximate variance of $n_i^{(t)}$, we use the asymptotic variance derived by equation 5 in Griffiths (1984), using the version for $\beta = 0$ and $\alpha < \infty$ (see also Chen and Chen 2013). To change the expression into our notation, we make the replacements $\alpha = n_i^{(0)} \Delta t / [2N_i(t)]$, $\beta = -\Delta t / [2N_i(t)]$ (because we assume only a single mutation distinguishing the two allelic types, so $\theta = 0$ conditional on that mutation), and $\eta = \mathrm{E}[n_i^{(t)}] \Delta t / [2N_i(t)]$ (by Griffiths' equation 4). Doing so and simplifying yields

$$\mathrm{Var}\left[n_i^{(t)}\right] \approx \mathrm{E}\left[n_i^{(t)}\right]^2 \left\{\mathrm{E}\left[n_i^{(t)}\right]-1\right\}^2 \left\{\frac{1}{\mathrm{E}\left[n_i^{(t)}\right]} + \frac{1}{\mathrm{E}\left[n_i^{(t)}\right]-1} - \frac{1}{n_i^{(0)}} - \frac{1}{n_i^{(0)}-1} - \frac{\Delta t}{N_i(t)}\right\}. \tag{C20}$$

Plugging the expressions for $f'\{\mathrm{E}[n_i^{(t)}]\}$ and $\mathrm{Var}[n_i^{(t)}]$ into Equation C18 gives an approximate variance,

$$\mathrm{Var}\left\{\log\left[\frac{n_i^{(t)}}{n_i^{(t)}-1}\right]\right\} \approx \frac{1}{\mathrm{E}\left[n_i^{(t)}\right]} + \frac{1}{\mathrm{E}\left[n_i^{(t)}\right]-1} - \frac{1}{n_i^{(0)}} - \frac{1}{n_i^{(0)}-1} - \frac{\Delta t}{N_i(t)}. \tag{C21}$$

Analogously, for the $m_i^{(t)}$ term,

$$\mathrm{Var}\left\{\log\left[\frac{m_i^{(t)}}{m_i^{(t)}-1}\right]\right\} \approx \frac{1}{\mathrm{E}\left[m_i^{(t)}\right]} + \frac{1}{\mathrm{E}\left[m_i^{(t)}\right]-1} - \frac{1}{m_i^{(0)}} - \frac{1}{m_i^{(0)}-1} - \frac{\Delta t}{M_i(t)}. \tag{C22}$$

Now, to obtain an approximate variance for $\widehat{p_i(t)}$, we plug the expressions in Equation C21, Equation C16, Equation C21, and Equation C22 into Equation C8. Doing so gives the expression

$$\mathrm{Var}[\widehat{p_i(t)}] \approx \left[\frac{N_i(t)}{N_i(t)+M_i(t)}\right]^2$$

$$\left\{ \begin{array}{l} \left\{ \dfrac{1}{\mathrm{E}\left[m_i^{(t)}\right]} + \dfrac{1}{\mathrm{E}\left[m_i^{(t)}\right]-1} - \dfrac{1}{m_i^{(0)}} - \dfrac{1}{m_i^{(0)}-1} - \dfrac{\Delta t}{M_i(t)} \right\} \\[4ex] \left\{ \dfrac{1}{\left[\frac{\Delta t}{2M_i(t)}\right]^2} - \dfrac{2}{\frac{\Delta t}{2M_i(t)}\left[\frac{\Delta t}{2M_i(t)}+\frac{\Delta t}{2N_i(t)}\right]} + \dfrac{1}{\left[\frac{\Delta t}{2M_i(t)}+\frac{\Delta t}{2N_i(t)}\right]^2} \right\} + \\[6ex] \left\{ \dfrac{1}{\mathrm{E}\left[n_i^{(t)}\right]} + \dfrac{1}{\mathrm{E}\left[n_i^{(t)}-1\right]} - \dfrac{1}{n_i^{(0)}} - \dfrac{1}{n_i^{(0)}-1} - \dfrac{\Delta t}{N_i(t)} \right\} \\[4ex] \dfrac{1}{\left[\frac{\Delta t}{2M_i(t)}+\frac{\Delta t}{2N_i(t)}\right]^2} \end{array} \right\}.$$

Noticing that $p_i(t) = N_i(t)/[N_i(t)+M_i(t)]$, expanding the products in the denominators and factoring out common terms gives:

$$\mathrm{Var}\left[\widehat{p_i(t)}\right] \approx \frac{4p_i(t)^2 M_i(t)^2}{\Delta t^2}$$

$$\left\{ \begin{array}{l} \left\{ \dfrac{1}{\mathrm{E}\left[m_i^{(t)}\right]} + \dfrac{1}{\mathrm{E}\left[m_i^{(t)}-1\right]} - \dfrac{1}{m_i^{(0)}} - \dfrac{1}{m_i^{(0)}-1} - \dfrac{\Delta t}{M_i(t)} \right\}\left[1-p_i(t)\right]^2 + \\[5ex] \left\{ \dfrac{1}{\mathrm{E}\left[n_i^{(t)}\right]} + \dfrac{1}{\mathrm{E}\left[n_i^{(t)}-1\right]} - \dfrac{1}{n_i^{(0)}} - \dfrac{1}{n_i^{(0)}-1} - \dfrac{\Delta t}{N_i(t)} \right\}p_i(t)^2 \end{array} \right\}.$$

Finally, noticing that $N_i(t) = p_i(t)[N_i(t)+M_i(t)]$ and $M_i(t) = [1-p_i(t)][N_i(t)+M_i(t)]$ and defining the coalescent time passed with respect to the whole sample, $\tau_s = \Delta t[N_i(t)+M_i(t)]$, gives

$$\mathrm{Var}\left[\widehat{p_i(t)}\right] \approx \frac{4p_i(t)^2[1-p_i(t)]^2}{\tau_s^2}$$

$$\left\{ \begin{array}{l} \left\{ \dfrac{1}{\mathrm{E}\left[m_i^{(t)}\right]} + \dfrac{1}{\mathrm{E}\left[m_i^{(t)}\right]-1} - \dfrac{1}{m_i^{(0)}} - \dfrac{1}{m_i^{(0)}-1} - \dfrac{\tau_s}{1-p_i(t)} \right\}\left[1-p_i(t)\right]^2 + \\[5ex] \left\{ \dfrac{1}{\mathrm{E}\left[n_i^{(t)}\right]} + \dfrac{1}{\mathrm{E}\left[n_i^{(t)}\right]-1} - \dfrac{1}{n_i^{(0)}} - \dfrac{1}{n_i^{(0)}-1} - \dfrac{\tau_s}{p_i(t)} \right\}p_i(t)^2 \end{array} \right\}. \qquad \text{(C23)}$$

To estimate the approximate variance in Equation C23, we replace $p_i(t)$ with its estimated value from the lineages-remaining estimator, τ_s with the (known or estimated) coalescent time elapsed between time points, and the expectations of $n_i^{(t)}$ and $m_i^{(t)}$ with their realized values.

The approximate variance in Equation C23 relies on an asymptotic variance of the number of ancestral lineages from Griffiths (1984). In practice, Equation C23 gives variance estimates that are much too large when the number of lineages are small and the time passed is short. For example, with $n_i^{(0)} = m_i^{(0)} = 10$, $p_i(t) = 1/2$, and $\tau_s = 0.001$, Equation C23 produces ~ 1.4. Because $\widehat{p_i(t)} \in [0, 1]$, its maximum variance is 0.25. Our strategy is to use the minimum of Equation C23 and the estimated variance of the waiting-time estimator with $\ell = 1$, or $2\widehat{p_i(t)}^2[1-\widehat{p_i(t)}]^2$ (Equation C11). The rationale is that $2p_i(t)^2[1-p_i(t)]^2$ represents the variance obtained when one estimates $p_i(t)$ from only one coalescent event on each background—the lineages-remaining estimator is always based on at least one coalescent event on each background.

Computing the Lineages-Remaining Estimator for Edge Cases

If no coalescences have occurred in the interval and $n_i^{(0)} = n_i^{(t)}$, then the estimator in Equation 13 is undefined. In this case, we compute the estimator for a larger value of Δt in which at least one coalescence event has occurred, extending Δt to the next time on the predefined grid in which a coalescence has occurred. The estimator is also undefined if $n_i^{(t)} = 1$ because the expected number of lineages remaining approaches but never reaches 1. In this case, we assume the coalescent time passed in the interval is $2[1 - 1/n_i(0)]$, which is the expected amount of time to coalesce from $n_i(0)$ lineages to one lineage. Finally, once a subtree coalesces to one lineage, we assume that population size remains constant into the past for the ancestral allele and, for the derived allele, that it remains constant before dropping to zero at in the middle of the branch connecting the derived subtree to the rest of the tree.

The approximate variance in Equation C23 is undefined if the estimated allele frequency $\widehat{p_i(t)}$ is equal to 0 or 1, in which case we define the estimated variance of $\widehat{p_i(t)}$ to be 0. It is also undefined if only one lineage remains on either background at the more recent end of the time interval [$n_i(0) = 1$ or $m_i(0)=1$]. If one lineage remains on one of the backgrounds, then we estimate the variance as the variance of the waiting-time estimator with $\ell = 1$. If only one lineage remains on each of the two backgrounds, we define the estimated variance of $\widehat{p_i(t)}$ to be $1/12$, which is the variance of a uniform$(0, 1)$ random variable.

Appendix D: The Correspondence of T_X and Q_X

In this section, we explain the correspondence between the T_X statistic proposed in Equation 16 and the Q_X statistic of Berg and Coop (2014). If the population history is tree-like, then Q_X can be thought of as checking for overdispersion (relative to neutral expectations) of polygenic scores at the tips of the population tree. In contrast, T_X checks for overdispersion at a set of time points along one branch of a population tree (*i.e.*, a single population at different times).

Briefly, Berg and Coop (2014) start with a vector $\overrightarrow{Z}$ whose elements are population-mean polygenic scores for a set of populations minus an assumed value for the population-mean polygenic score in an ancestral population. Under neutrality, $\overrightarrow{Z}$ follows a multivariate normal (MVN)$(0, 2V_A F)$ distribution, where V_A is the additive genetic variance, and F is a matrix reflecting the shared history of the populations. In particular, if the population history is tree-like, then entry F_{ij} in the F matrix reflects the branch length shared by populations i and j since their descent from the ancestral population. If $\overrightarrow{Z} \sim \mathrm{MVN}(0, 2V_A F)$, then the transformed vector $\overrightarrow{X} = (1/\sqrt{2V_A})C^{-1}\overrightarrow{Z}$ (where C comes from the Cholesky decomposition of F, such that $F = CC^T$) obeys an $\mathrm{MVN}(0, I)$ distribution. Thus, the sum of the squared entries of the transformed vector X (the quantity labeled Q_X) obeys a $\chi^2(w)$ distribution under neutrality, where w is the length of $\overrightarrow{X}$.

In the setting of this article, we have population-mean polygenic scores $Z(t_0), Z(t_1), \ldots, Z(t_w)$ for the population ancestral to one present-day population, assessed at various times. Each t_j represents some amount of time before the present and $0 \le t_0 < t_1 < \ldots < t_w$. In practice, we set $t_0 = 0$ and treat $Z(0)$, the population-mean polygenic score in the present, as Berg and Coop treated the ancestral polygenic score, so

$$\overrightarrow{Z} = [\,Z(t_1) \quad Z(t_2) \quad \ldots \quad Z(t_w)\,] - Z(0).$$

In this setting, with the present-day population serving the role of "ancestor," the shared branch length for two time points along one branch of a population tree is their shared distance backward from the present day, which is to say the branch length from the present back to the more recent of the time points. Thus, for this instance of $\overrightarrow{Z}$, each entry F_{ij} in the F matrix is equal to $t_{\min(i,j)}$, meaning

$$F = \begin{pmatrix} t_1 & t_1 & t_1 & \ldots & t_1 \\ t_1 & t_2 & t_2 & \ldots & t_2 \\ t_1 & t_2 & t_3 & \ldots & t_3 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ t_1 & t_2 & t_3 & \ldots & t_w \end{pmatrix}.$$

The Cholesky decomposition $\boldsymbol{F} = \boldsymbol{C}\boldsymbol{C}^T$ then gives

$$\boldsymbol{C} = \begin{pmatrix} \sqrt{t_1} & 0 & 0 & \ldots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & 0 & \ldots & \ldots \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \sqrt{t_3 - t_2} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \sqrt{t_3 - t_2} & \ldots & \sqrt{t_w - t_{w-1}} \end{pmatrix}.$$

The inverse of this matrix is

$$\boldsymbol{C}^{-1} = \begin{pmatrix} 1/\sqrt{t_1} & 0 & 0 & \ldots & 0 \\ -1/\sqrt{t_2 - t_1} & 1/\sqrt{t_2 - t_1} & 0 & \ldots & \ldots \\ 0 & -1/\sqrt{t_3 - t_2} & 1/\sqrt{t_3 - t_2} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 1/\sqrt{t_w - t_{w-1}} \end{pmatrix}.$$

If V_A is a constant, then the entries of $\overrightarrow{X} = (1/\sqrt{2V_A})\boldsymbol{C}^{-1}\overrightarrow{Z'}$ are equal to the X_j values of Equation 15, and Q_X is equal to the quantity in Equation 16.

Appendix E: Simulation Details

We ran simulations of polygenic scores, as well as coalescent trees and haplotypes associated with them, with and without selection.

For each trait, we simulated the frequency of the derived allele at each of k independent, trait-associated loci. We ensured that at all loci the minor allele frequency was at least 0.01, both at the onset of selection (if applicable) and at the present. Allele-frequency time courses were simulated using the normal approximation to the diffusion.

In particular, for each locus, an effect size for the derived allele was drawn from the normal distribution with expectation zero and variance equal to $h^2 \sigma_Z^2 w/k$, where h^2 is the desired heritability (because we are focused on polygenic scores rather than traits themselves, $h^2 = 1$), σ_Z^2 is the desired variance of the trait (i.e., polygenic score; always 1 in our simulations), w is a modified version of Watterson's constant for the population $\sum_{i=\lceil 2Nc \rceil}^{\lfloor (1-c)2N \rfloor} 1/i$ with c equal to the minimum minor allele frequency being drawn (here, $c = 0.01$), and k is the number of independent loci affecting the trait. Drawing effect sizes from this distribution gives the property that, under neutrality, if the loci affecting the trait are independent and a trait is formed by adding the individual's polygenic score to an independent "environmental" term with variance $1 - h^2$, then the resulting trait has variance σ_Z^2 in expectation.

Once the effect size was specified, we selected a derived-allele frequency from the population-level neutral site frequency spectrum, conditional on the requirement that the minor allele frequency be at least 0.01. That is, we set the probability of drawing a derived-allele frequency $k/(2N)$ to be proportional to $1/k$ for all $k \in \{\lceil 0.01 \times 2N \rceil, \ldots, \lfloor 0.99 \times 2N \rfloor\}$ (and zero otherwise). Under neutrality, this randomly drawn allele frequency was treated as the derived-allele frequency at the present. In simulations including a period of selection before the present, this frequency was used as the allele frequency at the start of selection. In simulations with selection, the derived-allele frequency was simulated forward in time in steps of $1/(2N)$ coalescent units (i.e., one diploid generation). Conditional on frequency p_t, the frequency at the next time step, p_{t+1}, was drawn from a normal distribution with expectation $p_t + sp_t(1 - p_t)$ and variance $p_t(1 - p_t)/(2N)$. Here, s is the selection coefficient on the derived allele at time t, which is computed as $s = \alpha\beta$ (equation 3.17 in Charlesworth and Charlesworth 2010), where α is the selection gradient on the trait at time t and β is the effect size of the derived allele. Two clarifications about the selection coefficient: first, this is the selection coefficient for the heterozygote, meaning that if the fitness of the ancestral homozygote is proportional to 1 (marginalizing over the other loci that affect the trait), then the fitness of the heterozygote is $1 + s$ and the fitness of the derived homozygote is $1 + 2s$. Similarly, the effect size α is the expected trait value difference between the heterozygote and the ancestral homozygote, holding the genotype at all other loci constant. Second, the selection gradient is equal to $\alpha = S/\sigma_Z^2$, where S is the selection differential on the trait—that is, the mean difference in phenotype between the parents of the next generation and the overall population (section 3.3.iii in Charlesworth and Charlesworth 2010)—and σ_Z^2 is the trait variance. Here, we used the expected trait variance at the onset of selection (which was 1) to set the value of α, which was retained over the course of the period of selection, regardless of the realized value of the trait variance.

Once the allele-frequency time course was simulated forward to the present, it was retained if the minor allele frequency in the present was at least 0.01. (Otherwise, the time course was discarded and the process was repeated.) To simulate the period from the origin of the mutation up until the onset of selection, we simulated steps backward in time using the normal approximation to

the neutral diffusion, conditional on loss of the allele. In particular, given a derived-allele frequency p_t, the allele frequency $1/(2N)$ coalescent time units further in the past was drawn from a normal distribution with expectation $p_t[1 - 1/(2N)]$ and variance $p_t(1 - p_t)/(2N)$ (Przeworski *et al.* 2005; Berg and Coop 2015; Lee and Coop 2017).

In each set of simulations with selection, we retained polygenic-score time courses if their difference in value between the beginning and end of the period of selection was within 5% of the target value, which is given by $2N\delta tS$, where N is the effective population size, δt is the duration of selection in coalescent units, and S is the selection differential.

To simulate coalescent trees and haplotypes at each locus, these allele-frequency time courses were used as input to mssel (Berg and Coop 2015), a version of ms (Hudson 2002) that can incorporate by conditioning on user-specified allele-frequency time courses. In mssel, we set the sample size to 200 and the number of derived chromosomes to the value of a binomial$(200, p)$ random draw, where p is the present-day frequency of the derived allele. If the random draw was equal to 0 or 200, then it was set to 1 or 199, respectively, to force the locus to be segregating in the sample. We chose a population size of 10,000, a haplotype length of 200,000 (with the effect locus at position 100,000), a per-base-pair mutation rate of 2×10^{-8}, and a per-base-pair recombination rate of 2.5×10^{-8}. These values imply population-scaled ms inputs of -r 199.5 and -t 159.68. Here, $4NrL = 200$ and $4N\mu L = 160$; our slight differences from these values arise from the fact that we compute $4Nf(0)$, where f is the probability distribution function of either a binomial(L, r) random variable or a binomial(L, μ) random variable, respectively, with $L = 200,000$. That is, we define the locus-wide recombination (or mutation) rate as one minus the probability that no recombinations (or no mutations) occur anywhere in the locus. We used the flags -T and -L to include the true trees and times to the most recent common ancestor in the output.

The coalescent trees produced by mssel for the selected sites were used as the true trees, and the haplotypes produced were used as input to RENT+ to generate estimated trees (Mirzaei and Wu 2016). Both true and estimated trees were read into R (R Core Team 2013) and handled using the ape package (Paradis *et al.* 2004). All the estimators and tests used in this article were coded in R.

Appendix F: Power Comparison with an Analog of trait SDS

Field *et al.* (2016) proposed the SDS as a test statistic for detecting very recent selection. Because recent selection distorts the terminal branches of the coalescent at a selected locus, Field *et al.* reasoned that recent selection would alter the distribution of singleton mutations around selected sites. Specifically, favored alleles are expected to have short terminal coalescent branches and thus relatively few singletons nearby compared with disfavored alleles.

Field *et al.* use distances to the nearest singletons on each allele's background to estimate the difference in the mean length of the terminal branches for the two alleles. This estimated mean difference is then standardized by an empirical mean and standard deviation computed from neutral simulations at a similar derived-allele frequency. To test for selection on polygenic traits, they construct a trait SDS (tSDS) by summing each locus's standardized SDS, signed so that positive values suggest selection for higher trait values.

Here, we compute power for an analog of tSDS in the simulations shown in Figure 6, with the differences that in our case the true terminal branch lengths are known, and we use the true effect sizes in computing tSDS. Both of these changes should enhance the power of tSDS. Specifically, at each locus the SDS value is

$$S = \frac{\bar{t}_a - \bar{t}_d - \bar{D}_{\text{DAF}}}{\text{SD}(D_{\text{DAF}})}, \tag{F1}$$

where $\overline{t_a}$ is the mean terminal branch length for the ancestral allele, $\overline{t_d}$ is the mean terminal branch length for the derived allele, $\bar{D}_{\text{DAF}}$ is the sample mean of $\overline{t_a} - \overline{t_d}$ computed from 500 neutral simulations at the same sample size and derived-allele frequency (in bins of 0.005), and $\text{SD}(D_{\text{DAF}})$ is the sample standard deviation of $\overline{t_a} - \overline{t_d}$ from the same 500 neutral simulations. [It is also possible to replace $\bar{D}_{\text{DAF}}$ and $\text{SD}(D_{\text{DAF}})$ with analytical values from Fu and Li (1993) under the assumption that the allele frequency does not change. Doing so gives similar results to those we report here.] To compute the tSDS statistic for a polygenic score, we use

$$\text{tSDS} = \sum_i \beta_i S_i, \tag{F2}$$

where β_i is the derived-allele effect size at locus i and S_i is the value of Equation F1 for locus i. [In Field and colleagues' version, β_i is replaced with sign(β_i), which has lower power when the effect size is known.] A polygenic score will tend to have higher tSDS if higher polygenic-score values have been selected for recently. To test for significance, we form a permutation distribution by recomputing tSDS 10,000 times while randomly permuting the effect sizes.
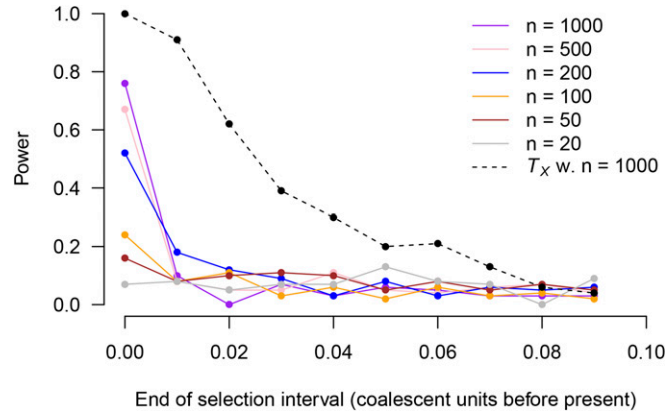
Figure F1 Power of the tSDS analog as a function of timing of selection and present-day sample size. In addition to results for the tSDS analog, we show power from the T_X test with a sample of 1000 chromosomes as a guide to the eye for comparisons with Figure 6 (dashed black line). As in Figure 6, power is shown on the vertical axis, and the horizontal axis shows the timing of a pulse of selection lasting 0.005 units and resulting in a shift in the population-average polygenic score of approximately one standard deviation (in present-day standard deviations). Each point is based on power from 100 simulations.

The power of the tSDS analog as a function of the timing of a selective pulse leading to a change in the polygenic score of approximately one standard deviation is shown in Figure F1. We also show the power of T_X with a sample size of 1000 chromosomes for comparison (from Figure 6). Compared with T_X, the tSDS has lower power, and its power drops off more rapidly as the timing of selection moves further into the past. This is not surprising because tSDS relies only on terminal branches, and T_X uses all the intercoalescent times.

At this writing, tSDS has the advantage of scaling readily to large samples. In contrast, T_X requires a reconstructed tree, which is currently error prone and computationally intensive. However, as tree reconstruction becomes faster and more accurate, the ability to use information from nonterminal branches will provide power benefits, especially beyond the very recent past.

Appendix G: Height Analysis Details

We estimated the time courses of two of the polygenic scores for height studied in Berg *et al.* (2018). Information about the loci included in each polygenic score (including reference SNP ID, effect size, chromosome, and position) was provided by Jeremy Berg. Each polygenic score was initially constructed by choosing the locus with the lowest *P*-value for a test of association with height, conditional on a minor allele frequency of at least 5%, within each of ∼1700 approximately independent genomic regions defined by Berisa and Pickrell (2016). (For GIANT, the polygenic score includes 1697 loci, and for UK Biobank, 1700 loci are included.) For the GIANT polygenic score, effect sizes and *P*-values were taken from Wood *et al.* (2014), and for the UK Biobank polygenic score, effect sizes and *P*-values were taken from Neale Lab (2017). [Berg *et al.* (2018) includes several polygenic scores constructed from UK Biobank effect sizes; the one we use corresponds to their UKB-GB.]

We used sequence information from the 1000 Genomes GBR subsample (1000 Genomes Project Consortium *et al.* 2012, release 20130502), which includes 91 genomes. For each locus included in each polygenic score, we extracted phased sequence information in a window extending 100,000 bases from the locus on each side using tabix (Li 2011). The resulting .vcf file was processed into a form acceptable by RENT+ using vcftools (Danecek *et al.* 2011) and R (R Core Team 2013). We then used RENT+ (Mirzaei and Wu 2016) to estimate an ARG for the region, including the "-t" flag to estimate branch lengths. Our version of RENT+ was modified slightly to print branch lengths to greater precision and to print RENT+'s internal estimate of θ that it uses to scale the branch lengths.

For both polygenic scores, we rescaled the effect sizes so that their standard deviation in the present-day sample is one, assuming linkage equilibrium among loci. In particular, if the sample frequencies of the effect alleles at the k loci are $p_1, p_2, \ldots, p_k$ and the effect sizes are $\beta_1, \beta_2, \ldots, \beta_k$, then if the loci are independent, the variance of the polygenic-score Z is

$$\text{Var}(Z) = 2 \sum_{i=1}^{k} \beta_i^2 p_i (1 - p_i).$$

If $\text{Var}(Z)$ is computed from the original effect sizes, then we rescale the effect sizes as $\beta_i' = \beta_i / \sqrt{\text{Var}(Z)}$ when computing the polygenic scores.

Finally, we rescaled the branch lengths estimated by RENT+. RENT+ estimates a per-nucleotide, population-scaled mutation rate $\theta = 4N\mu$ using Watterson's estimator (Watterson 1975). That is, for a sample of n haplotypes each covering a region of w base pairs, with S the number of segregating sites in the sample, θ is estimated as

$$\hat{\theta} = \frac{S/w}{\sum_{i=1}^{n-1} \frac{1}{i}}.$$

The time in coalescent units separating a pair of haplotypes is then estimated as $t = 2H/(L\hat{\theta})$, where H is the Hamming distance between a pair of haplotypes. For samples of size greater than two, RENT+ estimates branch lengths using a similar logic but using local distance matrices that take into account inferred recombinations.

We scaled RENT+'s estimated times because the Watterson estimator is biased downward if the population has been growing exponentially, and the human population has grown superexponentially (Keinan and Clark 2012). As a result, the inferred coalescent times from RENT+ were implausibly large. We first multiplied branch lengths at each locus by the θ estimate at each locus, giving branch lengths in units of twice the mutational distance per base pair at all sites. To convert these mutational distances into approximate coalescent units, we computed the time to the most recent common ancestor (tMRCA) at each locus, and then rescaled the branch lengths at all loci by a constant factor that set the mean tMRCA to two, the expectation in coalescent units under neutrality.