Title
Probabilistic models of DNA sequence and complex trait variation under adaptation

Author
Stern, Aaron J

Publication Date
2020

Peer reviewed|Thesis/dissertation

Probabilistic models of DNA sequence and complex trait variation under adaptation

by

Aaron J. Stern

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Professor Yun S. Song
Associate Professor Noah A. Zaitlen
Assistant Professor Priya Moorjani

Fall 2020

Probabilistic models of DNA sequence and complex trait variation under adaptation

Abstract

Probabilistic models of DNA sequence
and complex trait variation under adaptation

by

Aaron J. Stern

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Rasmus Nielsen, Chair

Adaptation is a fundamental process in evolution, which leads populations to better survive and reproduce in changing environments. A key insight of population genetics has been that present-day genetic variation is affected by past, and even ongoing, adaptations. Recent expansion of DNA sequencing has afforded us access to genetic variation from now up to nearly millions of individuals. In this dissertation, I develop modeling and inference for DNA sequence variation in order to identify the genetic bases of adaptations, with a focus on evolution of complex traits in humans.

First, I review population genetics approaches for detecting natural selection, and argue that these approaches have been hamstrung by the intractability of the so-called 'full likelihood' of selection (Chapter 1); I then develop a method to tractably compute this likelihood via importance sampling of the ancestral recombination graph (ARG), enabling us to find targets of selection too subtle to detect with previous methods (Chapter 2); I extend this likelihood method to jointly model DNA sequence variation *and complex trait variation* (via genome-wide association study [GWAS] summary statistics) to quantify the amount of selection acting on a complex trait, and to account for pleiotropy/correlated response in these estimates (Chapter 3); Finally, I present a method to detect polygenic adaptations in the presence of population structure, which explicit accounts for uncorrected stratification and other sources of error in the GWAS via an approach similar to LD score regression (Chapter 4).

To Mom, Dad, Hanna, and Bel, for their love and support.
And to Rasmus, from whom I've learned how to be a scientist and a scholar.

Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful for my advisor Rasmus. When I arrived at Berkeley, I was very green and very distractable. Most other professors would have been tempted to take on a more mature, more focused student. Despite this, Rasmus welcomed me into his lab and invested time and attention in me. Whenever I had a hair-brained research idea, I would sprint down the hall in VLSB to Rasmus's office and gently knock on his door. Almost always, Rasmus would usher me in, and we would discuss ideas or exciting new results. Rasmus, your enthusiasm and generosity of time has mattered more to me than you could imagine. I will always be thankful to have you as a mentor, an intellectual partner, and a source of support, even as I move on (perhaps temporarily...?) from Berkeley and academia.

To Noah Zaitlen: Thank you for teaching me complex trait genetics, and opening me up to new questions in biology that fascinate me to this day. One thing I love about Noah is that whenever I was down the 'rabbit hole', bemoaning an intractable problem or disappointing result, he would – in his contagiously optimistic way – step back and calmly suggest trying something simpler, or setting more realistic expectations. Noah, thank you for taking me in as one of your own and teaching me so much.

To my labmates of the last 5 years – Peter Wilton, Lenore Pipes, April Wei, Yun Deng, Hongru Wang, Debora Brandt, Tyler Linderoth, Vladimir Shchur, Emilia Huerta Sanchez, Joana Rocha, Diana Aguilar Gomez, Sandra Hui, Amy Ko, Greg Owens, Emma Steigerwald, Maya Lemmon-Kishi, Andrew Vaughn, Max Murphy, Geno Guerra, Zehui Chen, Mingpeng Zhang, Wanchang Zhang, Russ Corbett-Detig, Amy Goldberg, Becca Tarvin, Davide Marnetto – I've learned so much from all of you. I wish I were able to give you all a proper goodbye (I'm writing this in the Fall of 2020, during lockdown).

I want to give a special thank you to some of my co-authors, collaborators, and mentees: Peter Wilton, Leo Speidel, Fernando Racimo, Yulin Zhang, Courtney Rauchman, Zaid Ahmad, Evan Irving-Pease, and Jade Cheng. I have cherished working with and learning from you.

To the other friends I've made during my time at Berkeley: Max Rabinovich, you are a true mensch and meeting you at Berkeley has made this all worthwhile. Yun Deng, Alan Aw, Lawrence Uricchio, Hunter Nisonoff: I have always enjoyed our informal meetings & lunches to discuss project ideas and research. Nic Alexandre & Kirsten Verster: my dear VLSB buddies! Also, I'm grateful for all of the friends I have met through popgen, who have shown me immense generosity – especially Vince Buffalo, Doc Edge, Arun Durvasula, Joe Marcus, Josh Schraiber, and Shiya Song.

Chapter 1

Background: Detecting natural selection

This is work co-authored by Rasmus Nielsen. It is published in Handbook of Statistical Genetics, 4th Edition [1].

Abstract

Understanding natural selection is at the core of many evolutionary and population genetic investigations. However, it is typically difficult to directly detect natural selection. Instead, it has to be inferred from observations of DNA sequence data. In this chapter, we will briefly introduce some standard models of natural selection used in population genetics. We will then review some of the main signatures of selection that can be identified by analyses of DNA sequence data, and finally provide an overview of some of the many different statistical methods that have been developed to identify natural selection. We will argue that the lack of tractable likelihood approaches has spurred a large literature on more *ad hoc* statistical approaches based on summary statistics.

1.1 Introduction

Natural selection arises when individuals differ in fitness due to genetic factors — that is, have heritable differences in survival probability (viability) or reproductive success (fertility). Other factors, such as mutation and genetic drift, the random sampling of gametes (via which a new generation is formed by the previous generation), are also important evolutionary factors. However, selection plays a special role in driving adaptation, the evolutionary changes in response to environmental stimuli. Hence, understanding selection is the key to understanding how populations adapt to environments. Furthermore, at the molecular level, determining which genetic variants affect fitness provides information about which variants are important in interactions with the environment, including the response and susceptibility to disease.

In population genetics, the effects of natural selection are modeled as changes in allele frequencies. However, changes in allele frequencies can only rarely be directly observed, at least at this point in time (although with the increasing number of ancient DNA samples, this may become increasingly viable). Most studies aimed at detecting selection, therefore, focus on inferring selection indirectly from contemporary samples of DNA sequences. In this chapter we will discuss some of the statistical methods commonly used to infer selection (Section 4). Before doing so, we will briefly review some of the population genetic theory on natural selection (Section 2), and the general signatures associated with natural selection (Section 3). Although we will review some basics of negative and balancing selection, we will focus much of our review on methods for detecting positive selection.

1.2 Types of selection

We will begin by reviewing the most common models of selection in population genetics theory, acquainting the reader with terminology and properties of these models.

For the sake of simplicity, we will initially consider selection acting on a single di-allelic locus, with alleles A and a, in a diploid population. The changes in the frequency of A from generation to generation, the *trajectory* of the allele, is in part determined by the relative fitnesses of the three possible genotypes, w_{AA}, w_{Aa}, and w_{aa}. We can write these relative fitnesses in terms of the *selection coefficients*, s_{AA}, s_{Aa}, and s_{aa}[1], and if we assume that the frequency of A at generation t is $X_t \in [0, 1]$, we expect the frequency in the subsequent generation X_{t+1} to be

$$E[X_{t+1} \mid X_t = p] = p\frac{1 + s_{Aa}q + s_{AA}p}{1 + 2s_{Aa}pq + s_{AA}p^2}$$

where $q = 1 - p$. We can then use standard techniques for recurrence relations to describe the expected trajectory through time. However, since genetic drift is acting on the population at the same time, the recurrence relation based on the expectations will only be a rough approximation that tends to work well when the effect of selection is strong relative to genetic drift. Adding genetic drift to the model results in discrete-time Markov chain models, such as the familiar Wright-Fisher model, that describes the trajectory in discrete generations forward in time assuming binomial sampling of alleles between generations [2, 3].

These discrete generation models can be approximated in continuous time using diffusion equations [4, 5] which have revealed many interesting mathematical results regarding natural selection, such as the probability of fixation of an allele (the probability that it reaches a frequency of 100%) or the expected time it will take for the allele to reach fixation or loss [6, 7]. One important insight gained from theoretical population genetics is the fact

[1] There is a one-to-one mapping between relative fitnesses and selection coefficients. E.g., here we choose to define the selection coefficients by $w_{AA} = 1 + s_{AA}, w_{Aa} = 1 + s_{Aa}, w_{aa} = 1, s_{aa} = 0$. Then, given w_{AA} we can obtain s_{AA}, and given w_{Aa} we can obtain s_{Aa}, and vice versa.

that the effect of genetic drift is stronger in populations with smaller effective population sizes (N_e). Population geneticists, therefore, often see genetic drift and selection as two different forces acting at the same time, where N_e and the selection coefficients determine which of these two forces have the strongest effect on the dynamics of the allele frequency trajectory.

Directional selection

Models of selection on a di-allelic locus are sometimes classified into either directional positive selection, directional negative selection, or balancing selection based on the values of the selection coefficients. Directional positive selection is the case that $w_{AA} \geq w_{Aa} \geq w_{aa}$, excluding the case of $w_{AA} = w_{Aa} = w_{aa}$, if A is the derived (mutant) allele. In this case we expect the derived allele frequency to increase through time, i.e., $E[X_{t+1} \mid X_t = p] > p$ if $0 < p < 1$, and in the absence of genetic drift the selected allele will eventually go to fixation (reach a frequency of 100%).

Similarly, directional negative selection is the case that $w_{AA} \leq w_{Aa} \leq w_{aa}$, excluding the case that $w_{AA} = w_{Aa} = w_{aa}$. Here we expect the frequency of A to decrease on average, and in the absence of genetic drift A will approach a frequency of 0%. The special case of $w_{AA} = w_{Aa} = w_{aa}$ is the neutral case in which no selection is acting, and $E[X_{t+1}|X_t = p] = p$. In this case, fluctuations from p are only due to genetic drift, rather than both genetic drift and selection. Throughout this chapter, often we refer to a single selection coefficient s, rather than s_{AA} and s_{Aa}; unless stated otherwise, we assume that $s_{Aa} = s$ and $s_{AA} = 2s$, i.e., selection on A/a is additive, as well as positive directional.

Balancing selection

Unlike directional selection, in which alleles under selection tend to be lost or fixed in the long term, balancing selection refers to selection schemes that maintain multiple alleles in the population. One situation that produces this effect is heterozygote advantage (overdominance), where $w_{Aa} > w_{AA}$ and $w_{Aa} > w_{aa}$. In this case, if we ignore the effects of drift, X_t converges over time to an intermediate frequency; if $w_{Aa} = 1, w_{AA} = 1 - s_{AA}$, and $w_{aa} = 1 - s_{aa}$, then the equilibrium frequency $x^* = \lim_{t\to\infty} X_t = s_{aa}/(s_{AA} + s_{aa})$, rather than the boundaries at 0 or 1 . By contrast, under heterozygote disadvantage, where $w_{Aa} < w_{AA}$ and $w_{Aa} < w_{aa}$, there may exist an equilibrium frequency $x^* \in (0, 1)$ in the absence of genetic drift; however, if genetic drift causes the frequency to fluctuate away from x^*, then we expect X_t to be fixed or lost in the long term. For a deeper discussion of these models as well as the genomic signatures of balancing selection, we direct the reader to [8].

In addition to heterozygote advantage, selection schemes that can maintain variation include time- and space-varying selection. In these scenarios, directional selection can maintain variation, so long as the sign of the selection coefficient changes with sufficient frequency, either over time or space. Additionally, when the absolute fitness of an allele

depends negatively on its own frequency — so-called *negative frequency-dependent* selection — alleles can also stabilize at intermediate frequencies.

Polygenic selection

The simple di-allelic models described in the sections above are probably not realistic for much of the selection acting on the genomes of humans or most other organisms. Typically, a trait will be affected by multiple mutations and selection will, therefore, be polygenic (see e.g., [9]). While much of the early literature in population genetics focused on selection affecting a single locus for reasons of mathematical simplicity, there has recently been a resurgence of interest in polygenic selection. This interest stems in part from the realization due to genome-wide association studies (GWAS) that most humans traits of interest are highly polygenic (see e.g., [10]). However, methods for detecting polygenic selection from DNA sequence data are still in their infancy, and this chapter will focus primarily on methods aimed at detecting selection affecting a single locus. However, we note that a promising and very active research area is the development of methods for detecting and analyzing polygenic selection, and we later review several advances in this regard.

1.3 The signature of selection in the genome

In the previous sections, we reviewed the basic behavior of alleles under selection assuming a simple di-allelic model of selection acting on a single locus. The trajectory of allele frequency changes can be estimated directly from experiments of viral or bacterial evolution [11, 12, 13] or from analyses of ancient DNA (see e.g., [14, 15]), and can be used for quantifying and detecting selection [16, 17, 18]. However, most of the time, such direct inferences of the trajectory of allele frequency change is not possible. Instead, inference regarding past selection has to be made solely from observations of modern DNA. In the following sections we will discuss some of the patterns that can be observed in DNA sequences that have been subject to selection. Afterwards, we will review statistical methods for detecting and quantifying these patterns.

The signature of positive directional selection

We begin by reviewing the signatures that arise due to positive directional selection (see Section 2.1). In this section, we will review signatures such as increased rates of substitution, changes in the allele frequency distribution around selected alleles, and the hitchhiking effect.

Rates of substitution

An obvious consequence of positive selection is that favored alleles will have increased rates of substitution — i.e., the rate at which these alleles fix at a frequency of 100% is greater than the rate at which neutral alleles fix. Many methods for detecting selection take advantage of this insight. However, there are factors other than selection that can increase the rate of substitution, for example, increased mutation rate. Therefore, methods aimed at detecting positive selection by identifying increased rates of substitution must employ some standard of comparison to control for these factors. A common way of doing this is to compare mutations that *a priori* are, or are not, expected to be more likely to be under selection than other mutations. The most common comparison is of the number of non-synonymous and synonymous mutations that have fixed in protein coding regions (e.g., [19]). Due to the redundancy of the genetic code, mutations in protein coding regions come in two flavors: those that change the amino acid sequence (non-synonymous changes) and those that do not (synonymous changes). We expect that most selection in protein coding regions act at the amino acid level and that non-synonymous mutations, therefore, are more likely to experience selection than synonymous mutations. Hence, a signature of positive direction selection would be an increased number of fixed non-synonymous mutations compared to the number of fixed synonymous mutations. However, we note that selection may also act on synonymous mutations due factors such as codon usage preferences or maintenance of splice sites [20, 21, 22].

Frequencies of selected alleles

[23] showed that under equilibrium conditions, mutations are more likely to segregate at higher frequencies if they are under selection than if they are not. Thus, the frequency of an allele at a single time-point in itself provides information on whether a particular allele is under selection. However, if we make the assumption that the vast majority of mutations entering the population are more or less selectively neutral, then even high-frequency derived alleles must be primarily neutral [24]. Allele frequency alone is therefore not a reliable indicator of selection, and we must look to additional genomic signatures to improve our power to discriminate between selection and neutrality.

Nonetheless, on aggregate, selected mutations will tend to have different frequencies than neutral mutations. Comparisons of the distribution of allele frequencies in different categories of mutations, such as non-synonymous and synonymous mutations, can therefore be used to infer selection acting on sets of mutations. The distribution of allele frequencies — the so-called *site frequency spectrum* (SFS) — in models of selection is typically modeled using Poisson random field models pioneered by [25]. In these models, mutations enter the population according to a Poisson process, and selection and drift then act on the mutations to modify allele frequencies. Comparisons of the SFS stratified by different categories of mutations is an important tool in analyses of genomic data [26]. In particular, selection acting on a specific category of sites causes the SFS for that cate-

gory to differ from that of a category of sites assumed to be neutral, or the expected SFS under selective neutrality. The latter has a particularly simple expression: the expected proportion of mutations with allele frequency i, in a sample of size n, is given by $1/(ia_n)$, where $i \in \{1, 2, \ldots, n-1\}$ and $a_n = \sum_{j=1}^{n-1} j^{-1}$ is a normalizing factor [27].

Up to this point, we have mostly considered polymorphisms in a panmictic population. Another common signature of selection is differentiated allele frequencies among populations. Natural selection can increase the level of genetic differentiation among populations if selection acts differently in different populations or geographic regions due to differences in environmental factors [28]. Similarly, increased genetic differentiation among populations could also happen due to selection if selection acts on a recently arisen mutation that has not yet spread to other populations [29]. In fact, increased genetic differentiation among populations is one of the most characteristic signatures of natural selection. However, many highly differentiated allele frequencies may be driven by a combination of genetic drift and restricted gene flow, rather than selection.

Hitchhiking

We have so far discussed the direct effect of selection on the selected allele itself. However, selection also affects variation at linked neutral sites in the genome. When a selected allele increases in frequency, linked neutral alleles will also increase in frequency. This is the so-called "hitchhiking" effect [30, 31]. The consequence is a *selective sweep* (Fig. 1.1), which in the genomic region surrounding the favoured allele will lead to decreased variability (e.g., the number of segregating sites), increased identity by descent (IBD; i.e., DNA sequence identity due to recent common ancestry), and increased haplotype homozygosity [32]. Importantly, these patterns differ at different times during the sweep. During the earlier phase in which the favored allele has reached intermediate frequencies in the population (an incomplete sweep), haplotypes carrying the selected allele are highly uniform, since these haplotypes have increased in frequency so fast that recombination and mutation have not had much time to act. Consequently, haplotype homozygosity and IBD at this locus will be high [33, 34]. Furthermore, as we demonstrate in Fig. 1.2, neutral linked alleles will hitchhike along with the selected allele, resulting in an excess of alleles at frequency greater than or equal to that of the selected allele. As the sweep completes, haplotype homozygosity increases, and thee frequencies of linked alleles shift towards 100% along with the selected allele; thus, the SFS in a region that has recently underwent a recent sweep often is bimodal, with peaks at the frequencies 1 and $n-1$. However, this signature of bimodality is transient, as with time many of these high-frequency hitchhiking alleles will fix. Thus, a longer-term signature of a completed sweep is an SFS with an excess of low-frequency alleles due to recent mutation in the region around the selected allele. With passing time and accruing mutation and recombination, so too the patterns of increased IBD and haplotype homozygosity in this region become less pronounced.

Another effect of hitchhiking is a change in linkage disequilibrium (LD) around the selected site. [35] showed that while LD between neutral linked alleles on either side of the

selected allele increases on average, LD between the neutral linked alleles from opposite sides will be erased by the selective sweep.

So far, we have discussed models of the effect of a new, favored mutation rapidly increasing in frequency in the population. This model is known as a *hard sweep*. An alternative model involves *soft sweeps* [36], in which selection is acting on either recurrent mutations or on standing variation, i.e. on alleles that were already segregating in the population by the time that selection started to act. The signature of a soft sweep can differ substantially from that of a hard sweep. [37] showed that while a sweep from standing variation (SSV) from a low frequency $< (2N_e s)^{-1}$ results in the same reduction in diversity as a hard sweep, SSVs affecting alleles of sufficiently high initial frequency $> (2N_e s)^{-1}$ do not result in this decrease in diversity. The increased diversity resulting from SSVs relative to hard sweeps is due to accumulation of recombination near the selected site during the allele's neutral phase, before selection started acting.

Soft sweeps do share some signatures with hard sweeps, including decreased variability and increased haplotype homozygosity as well as increased IBD [38, 34]. However, these signatures are less pronounced for soft sweeps, because in this type of sweep several distinct haplotypes will increase in frequency, rather than a single haplotype under a hard sweep.

Balancing selection

The effect of balancing selection is in many ways opposite to that of a selective sweep. Mutations are maintained in the population for a prolonged period of time, leading to an increase in variability in the region around the selected variant. The SFS in regions surrounding the selected allele will contain many more alleles of intermediate frequency than expected for neutral regions. For a hard selective sweep, as selection becomes stronger, the width of the genomic region affected by the sweep becomes larger. However, for balancing selection, the width of the region in which linked neutral variants are affected by selection is narrow and on the order of $(2N_e r)^{-1}$, where N is the population size and r is the recombination rate per site [39].

Polygenic selection

Like hard sweeps, polygenic selection has been proposed as a mechanism for rapid adaptation. But polygenic selection produces a very different genomic signature than classical sweeps. Polygenic adaptations can occur without any particular selected allele fixing or rising in frequency as quickly as a hard sweep. When many polymorphisms control fitness, it is possible for a population to adapt with subtle allele frequency changes spread across many sites, rather than a classical sweep at any one of these sites; these interactions across loci create a different genomic signature in surrounding regions, which we illustrate using IBD tracts and allele frequency trajectories (Fig. 1.3). Polygenic selection may act on *de novo* mutations, standing variants, or recurrent mutation. While some younger alleles

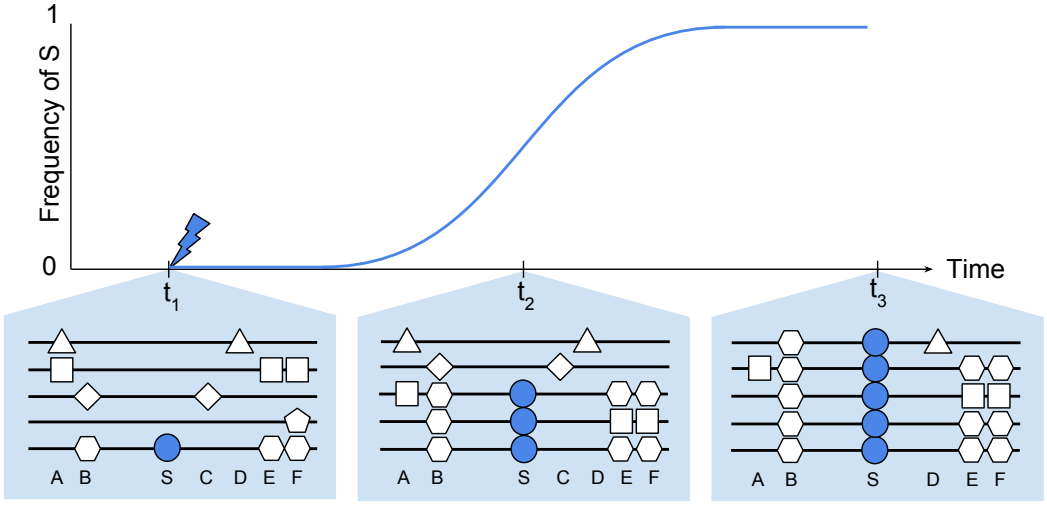


Figure 1.1: Genetic variation changes as a sweep progresses due to hitchhiking. Individual haplotypes ($n = 5$) are denoted using a different shape for each haplotype in the sample at t_1, to keep track of recombination events during the sweep. At t_1, the selected allele S mutates into the population on the ⬡ background. At this stage, there are six neutral polymorphisms (A–F) in the sample. At t_2, the sweep is ongoing and the frequency of the selected allele is intermediate (incomplete selective sweep). Relative to a neutral allele, fewer recombination events have occurred around the selected allele by the time it reaches this frequency, due to its rapid increase in frequency driven by selection. As a result, diversity within haplotypes carrying S is much reduced compared to haplotypes carrying the disfavored ancestral allele, i.e. there has been an increase in haplotype homozygosity within the allelic class carying S. Note that at this stage, two recombinations have occurred, both between □ and ⬡. At t_3, S has swept to fixation, along with the B allele. Another recombination event has occurred between △ and ⬡. Note the increase in high-frequency derived alleles and overall reduced levels of variability at t_3 relative to t_1 and t_2, exemplified by the loss of diversity at three sites (B, C, S). Furthermore, the sample is IBD for the entire tract $B - S$.

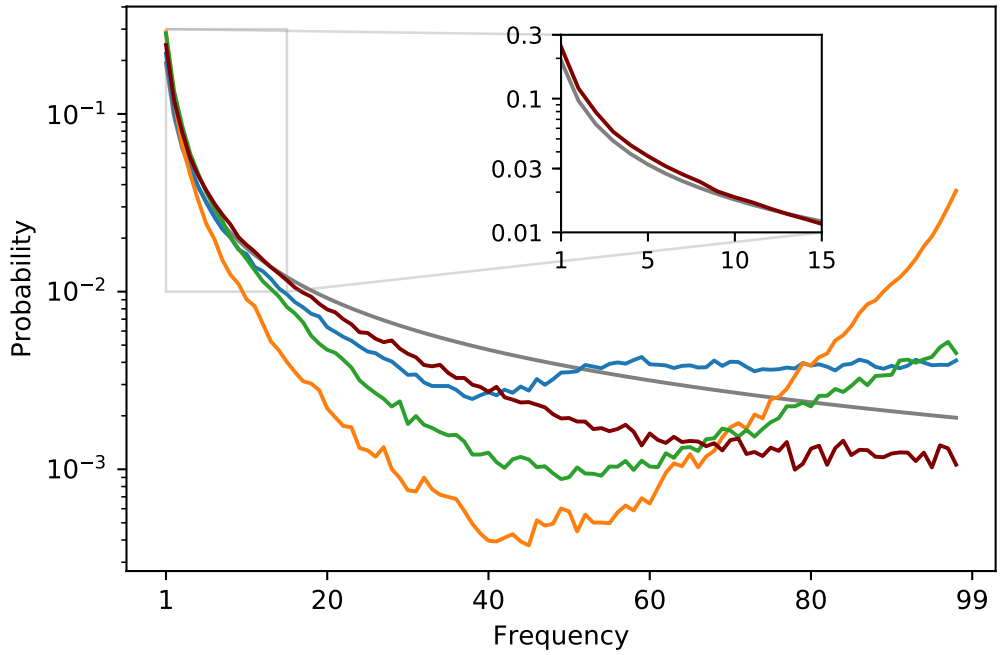


Figure 1.2: The SFS ($n = 100$) of a genomic region undergoing a selective sweep, in equilibrium population of $N = 10000$ sampled during different timepoints in a selective sweep with $s = 0.1$. We let $\theta = \rho = 100$ (where θ and ρ are the population-scaled mutation and recombination rates, respectively) and average the SFS over 100 independent simulations. Gray: the null expected SFS; blue: the favored allele is at 50% in the population (incomplete sweep); orange: the favored allele just fixed; green: 4000 generations after fixation; maroon: 12000 generations after fixation. Inset: Depicting the excess of low-frequency alleles lasting long after fixation.

under selection are likely to carry some classical signatures of selection, such as elevated IBD, standing variants under selection are less likely to possess such a drastic excess of IBD. Hence, all in all the signatures of polygenic selection can be very subtle and therefore difficult to detect.

[40] argued that in humans, polygenic selection has served as the major mode of recent adaptation. This is because human population-specific traits such as height and skin color exhibit high heritability and correlation to environment, and yet we observe a relative dearth of large allele frequency differences between human populations. These putative adaptations could be explained more feasibly as polygenic adaptations, rather than classical sweeps.

Confounders

A number of extraneous factors are frequently confounded with selection because they create a similar genomic signature to that left behind by selection. Several of these factors, such as genetic drift and increased mutation rate, have already been mentioned. However, there are a few more confounders of which to be wary. Most notably, selection scans are frequently confounded by unspecified non-equilibrium demography [41, 42]. For example, a hard sweep will in the long-term cause an excess of low-frequency derived alleles, just as an expansion in population size will cause the same signature, even in the absence of selection (Fig. 1.4). [43] showed that population size bottlenecks have a local effect on variation that is indistinguishable from that of a selective sweep. Similarly, balancing selection and recent selective sweeps can result in an excess of intermediate- and high-frequency derived alleles, respectively; these signatures can be mimicked under selective neutrality when sampled individuals hail from two different populations, unknown to the geneticist [44]. A common strategy for dealing with these demographic confounders is to control tests for selection using empirical distributions calculated genome-wide; this approach takes advantage of the tendency for demography to effect the entire genome, whereas signatures of selection tend to effect smaller genomic regions [43, 41].

Different modes of selection can also produce similar genomic signatures. [45] demonstrated that soft sweeps can be erroneously detected at the "shoulders" of hard sweeps, far enough from the selected mutation to host increased diversity, yet close enough to have an aberrant level of diversity relative to the background. They also show that these shoulders can be erroneously identified as ongoing sweeps [45].

1.4 Methods for detecting selection

In this section, we give an overview of statistical methods that make use of the signatures of selection that we discussed in the previous section. We begin with a discussion of methods to detect selection based on summary statistics, such as substitution rates, the site frequency spectrum, and haplotype homozygosity. Common to most of the described

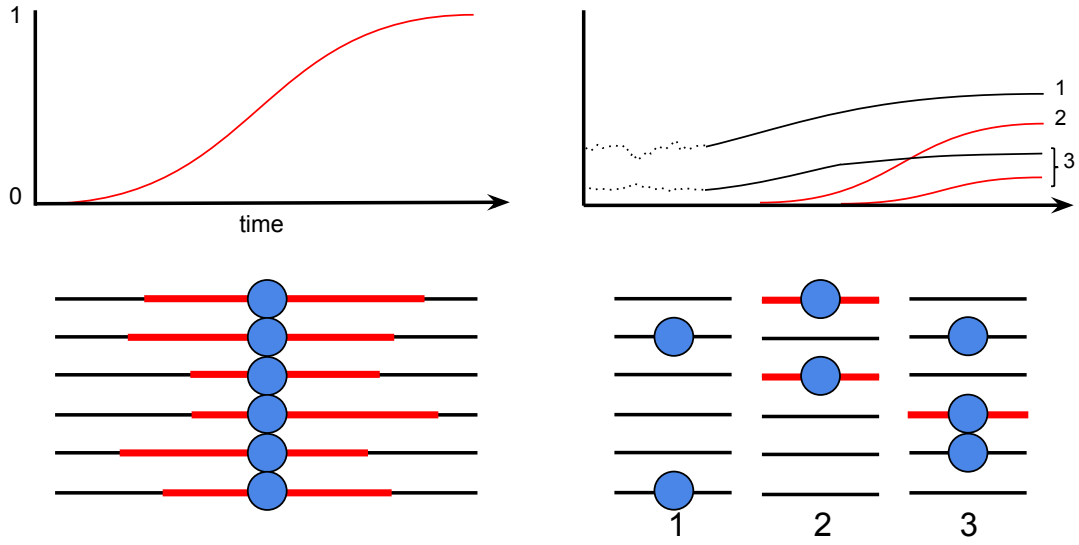


Figure 1.3: The genomic signatures of hard sweeps *vs* polygenic adaptations. In the bottom section of the figure, each row represents the chromosome/genome of an individual ($n = 6$), and each column represents a genomic region surrounding a favored allele, colored in blue. In the top section of the figure, allele frequency trajectories of selected alleles are shown, with *de novo* variants colored in red and standing variants in black. Left: In a hard sweep, selected alleles rise to high frequency or fixation with an excess of identity-by-descent (IBD) surrounding the driving allele. The red tracts around the selected allele represent IBD to the ancestral haplotype carrying the selected allele. Right: Polygenic adaptation acting on both standing and *de novo* variation at three different loci. Here, fitness is determined by two standing variants (loci 1 and 3) and two *de novo* variants that arise after selection begins (loci 2 and 3). At locus three we demonstrate the interference of a recent recurrent mutation (red tract signifies IBD between the present-day sample and the original haplotype carrying the mutation). Trajectories of alleles undergoing a sweep from the time of mutation are drawn in red, to signify the increased levels of IBD that tend to surround these alleles. This figure was adapted from [40].

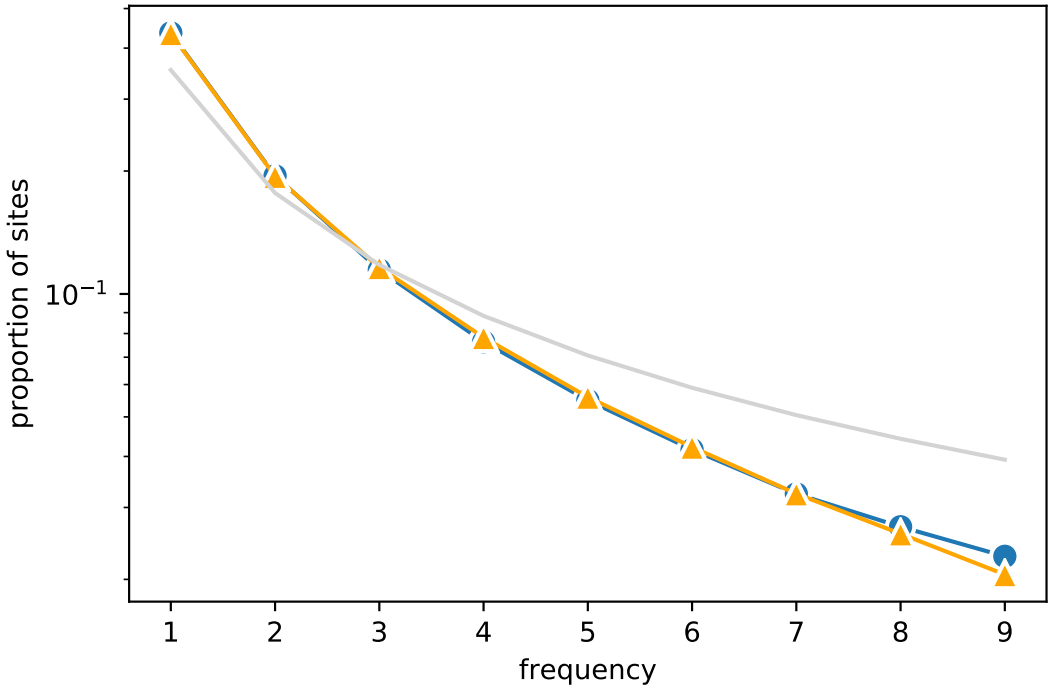


Figure 1.4: The expected site frequency spectra of $n = 10$ individuals under equilibrium demography (i.e., constant N_e in a panmictic population) with a selective sweep (◯), fluctuating N_e with no selection (△), and equilibrium demography with no selection (straight line). We obtain these SFSs by simulating 5000 unlinked loci with $\theta = \rho = 100$ (these parameters are the population-scaled mutation and recombination rates, respectively). The selective sweep has $N_e = 10^4$ and $s = 0.1$, conditioned on fixing 1.6×10^4 generations ago. The fluctuating N_e model has $N_e = 10^4$ for $0 \leq t < 1.2 \times 10^4$, $N_e = 8 \times 10^3$ for $1.2 \times 10^4 \leq t < 2 \times 10^4$, and $N_e = 2 \times 10^3$ for $t \geq 2 \times 10^4$.

methods is that they aim to detect selection by looking for deviations from the genetic patterns expected under neutrality (and compatible with selection). Thus, these methods are often called neutrality tests. Another commonality is that often these methods are applied in the context of a whole-genome scans, i.e., the summary statistic is calculated at a set of sites throughout the genome to identify regions with extreme deviations. Later, we discuss how to combine information across these various statistics; we discuss the challenges associated with likelihood-based inference, and review alternative approaches in this regard, such as composite likelihood, approximate Bayesian computation, and machine learning techniques.

Substitution-based methods

The pattern of an increased rate of substitution in a locus under positive selection (Section 3.1.1) has been extensively exploited to identify natural selection. Perhaps most famous are the tests based on the d_N/d_S ratio comparing two or more DNA sequences, typically from different species. The d_N/d_S ratio is the ratio of non-synonymous mutations per non-synonymous site to the number of synonymous mutations per synonymous sites. When comparing mutations among different species, the mutations largely reflect fixations between species, i.e. substitutions. The basic idea is that if no selection is acting on the mutations, then $d_N/d_S = 1$ in expectation. However, if positive selection is acting, then $d_N/d_S > 1$ in expectation. In its original formulation [46, 47], non-synonymous and synonymous sites were considered to be physical entities. However, because of the structure of the genetic code, both synonymous and non-synonymous mutations can occur in the same physical sites. A solution to this problem was proposed by [48] and [49] who developed Markov chain models of molecular evolution with a state space on the set of the 61 possible sense codons. Using such models, parameterized in terms of the rate of non-synonymous and synonymous rates of evolution, likelihood ratio tests of $H_0 : d_N/d_S = 1$ against alternatives of $H_A : d_N/d_S > 1$ can be established. Furthermore, these processes can be superimposed along the edges of a phylogeny to allow joint analysis of d_N/d_S ratios in multiple species. Popular computer programs implementing such tests include PAML [50] and HyPhy [51]. These tests have since been extended in various ways to detect selection acting along the edges of a phylogeny [52], acting in subsets of sites [53], or a combination of both [54]. Methods for detecting $H_A : d_N/d_S > 1$ allowing d_N/d_S to vary among sites according to some distribution have, in particular, been useful, as the intensity of selection is likely to vary greatly among sites in real proteins. Furthermore, even in proteins experiencing substantial amounts of positive selection, we would expect $d_N/d_S < 1$ for most sites, as selection acts mostly to preserve function on most protein coding genes.

While d_N/d_S tests originally were intended mostly for comparative data (data from different species), they can also be similarly applied to data from within a species, although recombination then poses a challenge [55]. Analyses of d_N/d_S ratios are usually carried out assuming a fixed gene tree topology, which mostly is not a problem when only one

sequence is included from each of a set of divergent species. However, when multiple sequences from the same species are included, different recombining sites will have different gene tree topologies and the assumption of a single shared gene tree topology is not longer satisfied by the data.

Methods comparing substitutions and diversity

Some of the most popular methods for detecting selection compare divergence between species with the amount of diversity within species. For example, the famous McDonald-Kreitman (MK) test establishes a 2×2 contingency table of the number of non-synonymous and synonymous mutations — NS and S, respectively — within and between species, estimated from multiple aligned sequences [56]. The MK test is then performed as a simple test of homogeneity. As the same underlying (set of) gene-trees are shared by synonymous and non-synonymous mutations, $NS_{within} \sim \text{Bin}(\lambda, NS_{within} + S_{within})$ and $NS_{between} \sim \text{Bin}(\lambda, NS_{between} + S_{between})$, where $\text{Bin}(\cdot, \cdot)$ indicates the binomial distribution and λ is the ratio of the rate of new neutral non-synonymous to synonymous mutations. If no selection is acting, except to immediately eliminate strongly deleterious mutations, λ should be the same within and between species. Significant deviations from the null hypothesis can be caused by either positive selection, resulting in an decrease in

$$NI = (NS_{within}/NS_{between})/(S_{within}/S_{between})$$

or negative selection causing a similar increase in NI, where NI is the so-called "neutrality index" [57]. However, some models of negative selection combined with fluctuations in population size may also cause decreases in the neutrality index [58, 59] A related test, and the first test for detecting selection aimed at DNA sequencing data, is the HKA test [60]. It is similar to the MK test in that it establishes a 2×2 contingency table comparing data within and between species. However, instead of comparing non-synonymous mutations and synonymous mutations, it compares variability in different regions of the genome. The test can, therefore, be extended to an arbitrary number of loci, k, in a $2 \times k$ table. Unfortunately, the rationale for the use of a simple test of homogeneity used for the MK test does not hold for the HKA test and simulations are needed to test significance. As for many of the tests discussed in this review, these simulations must necessarily assume a specific demographic model and there is no reason to assume that the results are particularly robust to the assumptions regarding demography.

Methods using the frequency spectrum

As previously mentioned in Section 3.1.2, the site frequency spectrum (SFS) describes the distribution of allele frequencies in multiple sites. Tests for selection acting on some category of mutations, such as non-synonymous substitutions, relative to synonymous substitutions, can also be carried out at the level of allele frequencies. A particular advantage of this approach is that parametric models of selection can be used to estimate

distributions of selection coefficients, using the comparisons of the SFS in different categories of sites, if one of the categories can be assumed to be neutral [61, 24, 26]. Some of our best estimates of the distributions of selection coefficients in humans and other organisms come from such comparisons [24, 26].

However, some of the most popular methods for detecting selection are based on a comparison of the SFS, not to a presumed neutral category of mutations, but rather to the expected SFS under models of a standard neutrally evolving population. As we discussed in Section 3.1.3, certain deviations from this null expected SFS can be indicative of selection. The most commonly used methods in this regard are based on simple summary statistics of the frequency spectrum, the most famous of which is Tajima's D [62]:

$$D = \frac{\widehat{\theta}_\pi - \widehat{\theta}_W}{z(S)}$$

where S is the number of segregating sites in the sample, $\widehat{\theta}_\pi$ is the average number of pairwise differences between individuals, and $\widehat{\theta}_W$ is Watterson's estimator of θ, i.e. $\widehat{\theta}_W = S/a_n$ where $a_n = \sum_{i=1}^{n-1} i^{-1}$. (The term $z(S)$ standardizes the variance of D.)

Under the null model of a population with constant effective size and selective neutrality, $E_0[\widehat{\theta}_\pi] = E_0[\widehat{\theta}_W] = \theta \equiv 4N_e\mu$, where μ is the per-generation mutation rate of the locus. Thus, $E_0[D] = 0$, and deviations from the underlying neutral model can be tested as deviations from $D = 0$. When applying Tajima's D to genome-wide data, D is typically calculated in sliding windows, or non-overlapping windows, to obtain a genome-wide distribution to which each local value can be compared, which makes it possible to control for confounding factors like non-equilibrium demography (see Section 3.4). Tajima's D detects selection primarily because the estimator $\hat{\theta}_\pi$ places a heavy weight on intermediate-frequency alleles, which are depleted after a selective sweep. Under these conditions, $E[D] < 0$, whereas under balancing selection, $E[D] > 0$.

Other SFS-based statistics related to Tajima's D have been proposed. One choice is Fu and Li's D (which we call D_{FL} to avoid ambiguity), defined as

$$D_{FL} = \frac{\widehat{\theta}_W - \xi_1}{z(\xi)}$$

where ξ is the SFS, ξ_1 denotes the number of sites at which only one individual carries the derived allele (i.e., the number of singletons), and z is again a scaling factor to standardize the variance of D_{FL} [63]. Similarly to Tajima's D, D_{FL} has the property that $E_0[D_{FL}] = 0$. Additionally, as previously mentioned, selective sweeps cause an excess of singletons after fixation; thus, like Tajima's D, we expect D_{FL} to take negative values after a sweep.

Another SFS-based statistic of this type is Fay and Wu's H, defined as

$$H = \frac{\hat{\theta}_\pi - \hat{\theta}_H}{z(\xi)}$$

where

$$\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

and ξ_i is the number of i-tons in the sample [64]. A key property of this statistic is that it places high weight on high-frequency derived alleles, and is thus sensitive to very recent selection rather than old sweeps; this is because the excess of high-frequency derived alleles immediately following a sweep is extremely transient. Thus, like D and D_{FL} we expect H to take on negative values if a selective sweep occurred recently.

This class of methods has also been adapted specifically for detecting balancing selection by [65], who introduced a statistic β that detects balancing selection by weighting alleles at intermediate frequency. As the reader might have noticed, these classical methods share a simple mathematical quality: they are all linear combinations of the SFS and some weight vector ω (see e.g. [66], which shows how to choose ω to detect specific violations of the neutral model with optimal power. As we discussed regarding Tajima's D, we can assess significance by calculating the genome-wide statistics under a particular choice of ω, and picking the most deviant regions). Other methods that use information from the SFS to detect selection include machine learning methods and composite likelihood methods, which we will discuss in later sections.

The performance of the methods based on summary statistics of the SFS varies wildly depending on whether a sweep has gone to fixation or is currently segregating (a so-called *incomplete* sweep); furthermore, they are easily confounded by demography such as population size bottlenecks [43, 42]. Power is also lessened when the sweep fixed long time ago after mutation and recombination have diminished the characteristic patterns of a sweep [42].

Methods using genetic differentiation

The very first test of neutrality [67] was based on detecting patterns of increased genetic differentiation among populations (see Section 3.1.2). While such methods were perhaps, for some years, viewed with skepticism by many researchers due to the strong assumptions they have to make regarding the history of the populations, they have had a resurgence after the emergence of genomic data as a convenient and simple tool to scan the genome for evidence of local selection. The most common measure of genetic differentiation among populations is F_{ST}, which can be defined in various ways, and confusingly both can take on the properties of a statistic and of a parameter (see [68] for a discussion). A common definition of F_{ST} for two populations, in a single di-allelic locus is [69]:

$$F_{ST} = \frac{c_1 p_1 (1-p_1) + c_2 p_2 (1-p_2)}{\bar{p}(1-\bar{p})}$$

where $c_1, c_2 > 0$, $c_1 + c_2 = 1$ (i.e., c_1, c_2 are the proportion of samples from each population), p_1 and p_2 are the sample allele frequencies in the first and second population,

respectively, and $\bar{p} = c_1 p_1 + c_2 p_2$ is the mean allele frequency (i.e., the frequency in the combined sample). Notice that the value of F_{ST} falls in $[0, 1]$; for highly differentiated allele frequencies we expect a value closer to 1, and for roughly undifferentiated frequencies we expect a value close to 0.

In genomic data, so-called F_{ST} scans are often used to detect selection, as elevated among-population genetic differentiation (high F_{ST}) may be a consequence of selection (see Section 3.1.2 and [28]). Whether an F_{ST} value is significantly elevated can be tested using parametric models or simulations. For example, [70] developed a hierarchical-Bayesian method for identifying outlier loci assuming a multinomial-Dirichlet likelihood function. However, often significance is not tested directly, but rather a list of the most extreme loci are presented without claims regarding significance.

Methods based on F_{ST} can also be extended to identify selection in an individual population by comparisons to multiple other populations. One particularly simple method for doing this is the so-called population branch statistic (PBS), which is based on transforming F_{ST} estimates between pairs of populations to an approximately linear distance and then inferring the amount of genetic drift distance on each branch of a tree with three populations [71]. Extreme drift on a population's branch at a particular locus is compatible with selection specific to that population acting on that locus.

More parametric methods will likely provide more power than simple methods based on F_{ST}. Furthermore, they can be used to test more specific hypotheses about the factors driving selection. Of particular interest in this regard are methods such as the one by [72] which uses a Bayesian model based on a Gaussian likelihood function for allele frequencies, to identify correlations between allele frequency changes across populations and specific environmental variables.

Methods based on genetic differentiation are also used to study polygenic traits. For example, the measure Q_{ST} is used to quantify differentiation of quantitative traits among populations [73, 74]. This quantity is calculated analogously to F_{ST}, but for a phenotypic trait instead of allele frequencies, and is often directly compared with F_{ST} to infer selection acting on traits. Indeed, under simple neutral conditions we expect genome-wide $F_{ST} = Q_{ST}$. Under negative selection for the same phenotype value across populations, we expect $F_{ST} > Q_{ST}$, and under directional selection on differing phenotype values among populations, we expect $F_{ST} < Q_{ST}$.

Unfortunately, for many realistic models of population structure, such as hierarchical population models, the assumptions of many of the standard tests are violated. [75] recently developed a method to test for selection on polygenic traits while controlling for hierarchical population structure, applying the same principles as in the aforementioned Gaussian approximation deployed by [72]. For a particular trait, they consider the quantity

$$\vec{z} = 2\mathbf{A}\vec{p}$$

where $\mathbf{A}$ is an $M \times L$ matrix of population-specific additive effect sizes of alleles on trait values (typically an estimate obtained from GWAS), $\vec{p}$ is a vector of allele frequencies, M

is the number of sub-populations and L is the number of alleles genotyped. Correcting for hierarchical population structure using the inverse sample covariance matrix $\mathbf{F}^{-1}$, they obtain a measure of the deviance of $\vec{z}$, called Q_X. Under the selectively neutral model with hierarchical structure, $Q_X \sim \chi^2_{M-1}$, and thus the significance of Q_X can be evaluated using the χ^2 distribution. A significant value of Q_X suggests the trait of interest has undergone recent selection. More recently, [76] developed an even more generalized approach for inferring polygenic selection in the presence of population structure, allowing selection to act along specific edges of an admixture graph.

Methods using haplotype structure

The methods discussed previously focused on allele frequencies and allele frequency changes. However, other features of the data can be leveraged for detecting selection, in particular, haplotype structure, a signature we introduced in Section 3.1.3. Many methods aimed at detecting ongoing selection (incomplete selective sweeps, a concept we introduced in Section 3.1.3) focus solely on haplotypes — specifically, the pattern of increased haplotype homozygosity on chromosomes carrying the advantageous mutation.

[33] developed a statistic called extended haplotype homozygosity (EHH) that estimates the probability that two randomly chosen haplotypes are identical up to a distance x around a particular candidate SNP called the core SNP. More precisely, we can define EHH as the number of pairs of identical haplotypes in a window of length x divided by total number of pairs:

$$\text{EHH}(x) = \sum_{h \in \mathcal{H}(x)} \frac{\binom{n_h}{2}}{\binom{n}{2}}$$

where $\mathcal{H}(x)$ is the set of distinct haplotypes in the sample only considering sites within a distance x from the core SNP, n_h is the number of type-h chromosomes, and n is sample size[2]. Notice that if we calculate $\text{EHH}(x)$ right around a particular site so that the window size is 0, then $\text{EHH}(0) \approx p^2 + q^2$, where p is the frequency of the core SNP and $q = 1 - p$. For increasingly large window sizes, $\text{EHH}(x)$ converges to zero because all n haplotypes become distinct when considering a sufficiently large region. However, the *rate* at which EHH decays to 0 with respect to x reflects the age of the core SNP, which depends on the strength of selection. A slow decay of EHH is compatible with recent selection; as discussed in the section on hitchhiking (Section 3.1.3), the region surrounding the selected allele tends to be depleted of variation and recombination, and thus EHH decays more slowly in this case than under selective neutrality. Thus, an elevated value of EHH around a core SNP serves as a convenient feature for detecting loci under positive selection. However, one general challenge, in addition to the reliance on phased data and a genetic map, is that other processes, such as a local reduction in mutation rate or an increase

[2]Note that $\sum_h n_h = n$; thus, another approximately equivalent calculation of EHH (assuming n is large) is $\text{EHH}(x) = \sum_h p_h^2$ where $p_h = n_h/n$.

in negative selection, also can lead to increased haplotype homozygosity. Therefore, the relative EHH (rEHH) between different classes of haplotypes (i.e., haplotypes grouped by the allelic state of the core SNP) was proposed as a more robust method for detecting selection [77].

A further development of the EHH family of statistics was proposed by [78], who developed the Integrated Haplotype Score (iHS), a statistic designed to detect ongoing selection. iHS partitions haplotypes based on the ancestral/derived states of the core SNP, and is based on integrating EHH from the core SNP until EHH reaches a certain fixed value (typically 0.05). These integrated EHH values are called iHH_A and iHH_D, and the statistic $\log(\text{iHH}_A/\text{iHH}_D)$ is then calculated genome-wide and standardized by the empirical mean and variance of this statistic using other genomic SNPs at the same frequency. Since selected alleles tend to carry longer surrounding IBD tracts than neutral alleles at the same frequency as the selected allele, we expect the most negative iHS values to indicate strongly selected derived alleles. Importantly, iHS is standardized by allele frequency because low-frequency alleles tend to be younger and thus carry high amounts of IBD, even if they are selectively neutral. Notice that to calculate iHS, it is assumed that a genetic map is known to integrate EHH with respect to distance. [79] proposed an alternative to iHS called nSL (number of segregating sites by length) that avoids relying on a genetic map by, for each pair of haplotypes, using the number of mutations within the other $n-2$ haplotypes to measure a mutational distance, leading to increased robustness against recombination and mutation rate variation. One important note is that the expectation of iHH is infinite under a standard neutral model, making statistics based on the iHH statistic highly sensitive to the choice of maximal window size for calculations of iHH [79].

Importantly, the aforementioned measures of haplotype homozygosity are underpowered to detect soft sweeps [38, 80]. [81] developed an alternative haplotype-based statistic specifically designed to detect both hard and soft sweeps. They defined

$$H_{12} = (p_{h_{(1)}} + p_{h_{(2)}})^2 + \sum_{j>2} p_{h_{(j)}}^2$$

where $h_{(1)}$ and $h_{(2)}$ are the first- and second-most frequent haplotypes in the set of distinct haplotypes $\mathcal{H}$, and $p_h = n_h/n$. Under a hard sweep, we expect $p_{h_{(1)}} \gg p_{h_{(2)}}$, whereas under a soft sweep, the discrepancy tends to be less severe; nonetheless, under a soft sweep we expect the several most frequent haplotypes to still dominate the haplotype distribution, and thus H_{12} is sensitive to both cases. To distinguish between hard and soft sweeps, they propose

$$H_1/H_2 = \frac{p_{h_{(1)}}}{\sum_{h \neq h_{(1)}} p_h}$$

By the same intuition for defining H_{12}, here a high value of H_1/H_2 implies a hard sweep, whereas a low value implies a soft sweep.

Recently, [80] developed a haplotype score called the Singleton Density Score (SDS) designed to have especially high sensitivity to detect extremely recent signatures of selection, relative to comparable methods such as iHS. Their approach is based on the intuition that for ongoing or recent sweeps, the haplotypes carrying the favored allele have a dearth of singletons. To compute the SDS at a particular site, SDS iterates through each diploid individual. For each individual, the distance between the nearest singletons up- and downstream of the core allele is computed, and these n distances are binned based on whether the individual is homozygous for the derived allele, homozygous for the ancestral allele, or heterozygous. A likelihood model is used to infer the "mean tip length" of ancestral and derived lineages; essentially, long singleton distances imply a short mean tip length. The inferred ancestral and derived mean tip lengths, called $\hat{t}_A$ and $\hat{t}_D$, respectively, are standardized similarly to iHS. The SDS exploits the fact that we expect $\hat{t}_A > \hat{t}_D$ when the derived allele has risen sharply in frequency in the immediate past. Thus, the haplotypes surrounding a positively selected allele are expected to be depleted of singletons relative to a neutral allele segregating at the same frequency. The authors also designed a score called the trait SDS (tSDS), where the sign of the SDS is flipped in the case that the ancestral allele is associated with increasing the value of the trait (e.g., associated with a positive change in height). This measure can be used to demonstrate polygenic selection on a trait by showing an excess in tSDS across associated sites.

While haplotype-based methods are mostly designed for detecting ongoing sweeps, rather than completed sweeps, there is also a distinct pattern of linkage disequilibrium arising after a sweep that can be exploited for detecting sweeps. As discussed in Section 3.1.3, right after a completed sweep there will be increased LD to either side of the selected sweep, but no LD between SNPs from opposite sides of the selected sites [35]. [35] proposed using a statistic, ω, to detect this pattern.

Why full-likelihood methods are intractable for population samples

So far we have discussed various statistics used in tests aimed at detecting natural selection. The statistically minded reader might appropriately at this point wonder why there exists such a plethora of more or less *ad hoc* statistics, and why there are no methods for detecting selection based on likelihood functions that incorporate all information regarding the selection, including allele frequencies and haplotypes. Unfortunately, full likelihood methods that incorporate selection are considered computationally intractable. Some progress was done on models of weak selection without recombination [82]. However, these methods never scaled up to genomic data. Several methods have been developed that use simulations to approximate the likelihood for a single non-recombining locus under various assumptions [83, 84]. In particular, [84] developed a likelihood method for detecting and estimating the strength of selection by first simulating an ancestral allele frequency trajectory and then simulating a coalescence tree conditionally on the allele frequency trajectory. Like [82], it is computationally intensive and the assumed absence of recombination makes it inapplicable to most data, such as human nuclear DNA.

Because full-likelihood inference is not viable for even small sample sizes, most methods for detecting selection rely on summary statistics that capture particular signatures of selection. The major challenge is that calculating the likelihood requires integrating out many sources of stochasticity, including allele frequency trajectories, the latent ancestral recombination graph (ARG)[3] conditional on this trajectory, and neutral mutations superimposed on the ARG. The vast combinatorial space of ARGs makes analytical calculations impracticable. However, in lieu of tractable full likelihood methods, a number of methods have been developed that attempt to detect and/or quantify selection using functions that approximate the likelihood function.

Composite likelihood methods

Using diffusion theory, [86] and [87] developed expressions for the distribution of sample allele frequencies (i.e., the SFS) as a function of the genetic distance from a recently completed sweep. Based on these calculations, [87] could define a composite likelihood formed as the product of the individual likelihood functions calculated for each site along the length of the sequence, as a function of the sites recombination distance to the selected SNP and the selection coefficient. They then proposed to use a likelihood ratio to test the null hypothesis of no selection ($s = 0$) and to estimate the strength and location of the sweep. The advantage of this method over previous methods was multiple: First, it employed all of the information from the allele frequency by using a full-likelihood approach to the allele frequencies. Secondly, it used the spatial distribution of SNPs and their allele frequencies to gain power and to locate the most likely selected SNPs. [41] extended the method using an approximation by [88] which considered the probability that a particular lineage in the genealogy "escaped" a sweep, i.e., the probability that a neutral allele linked to the non-beneficial allele recombined onto a beneficial background prior to loss of the non-beneficial allele. Using this result, the composite likelihood function could be calculated faster and could incorporate any SFS as the 'background' neutral allele frequency distribution to be tested against. This method has since been modified in multiple ways, including extensions to incomplete sweeps [89], modeling of population structure [90], and incorporation of negative selection in the genomic background [91].

Approximate Bayesian computation

The previous section on composite likelihood methods introduced approaches for estimating selection coefficient and the location of the selective sweep. However, there are other approaches for addressing this problem—in particular, methods based on approximate Bayesian computation (ABC) [92]. ABC works by repeatedly sampling parameters (such as s) from a prior distribution and then subsequently simulating a genomic data set for

[3] A complex graphical structure that represents all of the genealogical and recombinant events occurring in a sample; see e.g. [85]

each sampled parameter value. Without loss of generality, assuming we wish to approximate the posterior of s, we use an acceptance/rejection scheme where sampled values of s are rejected if the distance between the resulting simulated data and the observed data is sufficiently large. To determine distance between the simulated and observed data, many approaches use classical SFS- or haplotype-based summary statistics and calculate a distance (e.g., Euclidean distance) between the two summary vectors calculated for the observed and simulated data. After sampling and simulations are completed, the estimate of the posterior can be used to derive a maximum *a posteriori* (MAP) estimate of s, as well as its Bayesian credible interval. Unlike other methods based on individual summary statistics, ABC takes full advantage of the informative correlation structure between different types of summary statistics.

[93] developed an ABC method to jointly estimate selection coefficients, the time at which selection started, and the frequency of the selected mutation at the time selection started. The method could also perform model selection, distinguishing soft sweeps from hard sweeps. Similar methods have been used by [94] and [95].

A common pitfall of ABC methods is that they are computationally intensive, and can suffer from the Curse of Dimensionality. That is, when the sample space of the summary statistics increases in dimensionality, so does the instability of the likelihood estimate [96]. Recently, [97] showed that using an average one-dependence estimate (AODE) assumption of the structure of the likelihood can ameliorate this instability problem, while retaining some of the informative correlation structure of the likelihood.

Machine learning methods

An alternative to composite likelihood and approximate Bayesian methods, for incorporating more information from the data, is to take advantage of standard *supervised* machine learning methods. Broadly speaking, supervised methods aim to train some model to use statistics extracted from the data, often called *features*, with the goal of making accurate predictions based on such features. By contrast, so-called *unsupervised* methods are used to find structure in data, rather than generate predictions. Principal components analysis (PCA) is one example of an unsupervised method, often applied in statistical genetics to illustrate and control for population structure [98, 99].

Mathematically, we can summarize the supervised learning problem as follows: assume the availability of a training set of pairs $\{(\vec{x}_i, y_i)\}_{i=1}^n$, where each pair consists of a feature vector $\vec{x}_i$ and its *label* y_i for each sample $i = 1, 2, \ldots, n$. The objective is then to select a function f^* that maps feature vectors to labels such that

$$f^* = \text{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(f(\vec{x}_i), y_i) \right\}$$

where $l(\hat{y}, y)$ is a *loss function* that is minimized when $\hat{y} = y$ (i.e., when the estimated labels $\hat{y}$ perfectly match the true labels), and $\mathcal{F}$ is a pre-specified set of prediction functions. This

framework, can be used to develop methods for inferring selection: for example, a so-called *classifier* can be trained to detect selection, where the labels y are binary variables such that $y = 0$ signifies "neutrality" and $y = 1$ signifies "sweep". Additionally, techniques such as linear regression can be used to estimate the value of the selection coefficient s. Notice however, that a training set needs to be available for which the correct labels are known. In population genetics, such training sets are rarely available. Instead, simulated data are used to train the classifiers. Also, the feature vector has to be chosen, and is usually based on the same type of summary statistics as used in other simulation based methods. [100] reviewed supervised machine learning in the context of population genetics and detecting selection, defining the problem and illustrating practical concerns in greater detail.

In the previously mentioned classical SFS-based methods such as Tajima's D, the SFS is summarized as a linear combination that has expectation 0 under neutral equilibrium conditions. By contrast, a method by [101] (SFSelect) uses the full SFS as the high-dimensional analog to these classical methods . They simulate data under specific sweep and neutral conditions and train a machine learning model called a support vector machine (SVM; see [102]) to classify SFS vectors. Similar approaches have been taken to integrate various haplotype statistics along with the SFS; [103] trained an SVM using a combination of LD- and SFS-based statistics, and the method EvolBoosting [104] integrates both SFS- and haplotype-based statistics to detect selection using boosting [105]. SHiC [106] is a method that extracts a number of haplotype- and SFS-based statistics from simulated data to train an Extra-Trees classifier [107]. Notably, the latter method has been shown to retain good power to detect strong selection despite model misspecification during training (i.e., trained under equilibrium demography and tested under fluctuating N_e) [106].

1.5 Discussion

As evident from the previous sections, there is a very large set of different methods for detecting selection. In fact, in this review we have only covered a subset of the most commonly used methods. For example, recent methods for detecting adaptive introgression are not covered (see e.g., [108]). A common theme for many of the methods is that there typically is a trade-off between power and robustness. Since the emergence of the first neutrality tests [28], there has been an awareness that many demographic models can mimic the signature of selection. With the emergence of genomic data, it was generally hoped that this problem would vanish as the signature of demographic processes affect the entire genome, while selection may only affect one or a few loci. While genomic data certainly has helped identify signatures of selection, we are still facing challenges when assigning p-values, or other methods of statistical confidence for inferences of selection. In the end, almost all methods rely to some degree on the assumption of a demographic model. By the very nature of the data, the null hypothesis considered will always be a composite hypothesis that also includes features of the demography. To address this problem, most studies rely on one of two possible strategies: (1) They may give up on

including measures of statistical confidence and instead simply produce a list of the best candidates for targets of selection. One variant of this approach is the use of so-called "empirical p-values" (e.g., [78]), which in this context simply are quantiles of the empirical distribution of the test statistic. They are, therefore, not p-values in the classical sense and should probably more appropriately simply be reported as quantiles. (2) The alternative approach is to make specific assumptions about the demography, typically based on estimates of demography obtained from the same or other data. Simulations are then used in one form or another to generate the distribution of the test statistic under the null hypothesis. Variants of this approach includes the machine learning and ABC methods which include simulations as an integrated part of the inference framework.

As previously mentioned, another major challenge of inferences of selection is that full likelihood methods are not available. However, they may eventually be practicable, or at least closely approximated, by building on advances in inferring ARGs. ARG inference methods have historically been impractical for even modest sample size and locus length [109, 110]. To ease computational costs, [111] calculated a heuristic for the conditional sampling distribution (CSD) of the nth sequence given $n-1$ other sequences, which allowed them to conduct approximate maximum-likelihood inference of recombination rates without explicitly sampling the ARG. [112] showed that the so-called sequentially Markov coalescent (SMC) is remarkably consistent with the coalescent with recombination; this approximation, along with the SMC', a similar approximation due to [113], allow extremely efficient approximate simulation of the ARG and maximum-likelihood inference of population size history under the pairwise sequentially Markov coalescent [114]. Recently, [115] developed an probabilistic method approximate the posterior distribution on ARGs, based on both the CSD and SMC/SMC' approximations. Their method ARGweaver efficiently samples posterior ARGs, and scales well with genome length and sample size. Based on these advances, it may be possible in the future to develop methods for detecting selection that more closely approximate the full likelihood function.

1.6 Overview of dissertation

In this dissertation, I place a special emphasis on methods I developed which address the aforementioned issue of full-likelihood inference by leveraging recent advances in ARG inference (see Chapters 2 and 3). Another focus of this dissertation is in developing methods that scale to highly polygenic traits; I approach this problem from two angles, one using a full-likelihood approach (Chapter 3) aimed at examining within-population variation, and a simpler regression approach based on random-effects models (Chapter 4) aimed at examining variation along axes of population variation (e.g., principal components).

Chapter 2

Approximate full-likelihood of selection via importance sampling of ancestral recombination graphs

This is work co-authored by Peter Wilton and Rasmus Nielsen. It is published in PLOS Genetics [116].

Abstract

Most current methods for detecting natural selection from DNA sequence data are limited in that they are either based on summary statistics or a composite likelihood, and as a consequence, do not make full use of the information available in DNA sequence data. We here present a new importance sampling approach for approximating the full likelihood function for the selection coefficient. The method treats the ancestral recombination graph (ARG) as a latent variable that is integrated out using previously published Markov Chain Monte Carlo (MCMC) methods. The method can be used for detecting selection, estimating selection coefficients, testing models of changes in the strength of selection, estimating the time of the start of a selective sweep, and for inferring the allele frequency trajectory of a selected or neutral allele. We perform extensive simulations to evaluate the method and show that it uniformly improves power to detect selection compared to current popular methods such as nSL and SDS, under various demographic models and can provide reliable inferences of allele frequency trajectories under many conditions. We also explore the potential of our method to detect extremely recent changes in the strength of selection. We use the method to infer the past allele frequency trajectory for a lactase persistence SNP (*MCM6*) in Europeans. We also study a set of 11 pigmentation-associated variants. Several genes show evidence of strong selection particularly within the last 5,000 years, including *ASIP*, *KITLG*, and *TYR*. However, selection on *OCA2/HERC2* seems to be much older and, in contrast to previous claims, we find no evidence of selection on

TYRP1.

2.1 Introduction

Direct observation of the change in allele frequency over time (the allele frequency trajectory) allows one to make powerful inferences regarding whether selection acted on the allele [117, 118]. However, outside of certain contexts such as experimental evolution of viruses or bacteria [119, 11, 12, 13] or analyses of ancient DNA samples [14, 15], in most cases such direct observations of allele frequencies at multiple points in the history of a population are unavailable. Instead, selection must be inferred from contemporary, modern data. A wide variety of methods have been developed to detect selection based on patterns observed from modern DNA sequences (e.g. [33, 41, 78]).

The hitch-hiking effect provides a key signature of selection in modern datasets [30]. [30, 31]. Hitch-hiking causes aberrations in the spatial pattern of genetic diversity, including the site frequency spectrum (SFS) [62, 86] and the pattern of haplotype homozygosity [33]. Methods designed to detect these aberrations are particularly useful in the setting where a single population is surveyed, and the only information available is variation within this single population.

The most familiar methods for detecting selection are based on linear functionals of the SFS, such as Tajima's D, Fu and Li's D, or Fay and Wu's H [62, 63, 64]. An advantage of SFS-based methods is that they do not require the data to be phased. However, these methods have several limitations: they tend to confound selection with other non-equilibrium conditions, such as a fluctuating population size [41, 42]; they are not suitable for estimating parameters such as the value of the selection coefficient s; significance can usually only be established using an empirical null distribution; and crucially, these methods do not incorporate any features of the haplotype structure.

To make fuller use of information provided by phased sequence data, a number of methods have incorporated summary statistics based on haplotype structure. In a broad sense, these methods are based on calculations of haplotype similarity in a window around some core site of interest [33]. Several methods have adapted this general concept to specifically detect ongoing selection [78, 79, 38]. More recently, [80] showed that the density of singletons surrounding a focal SNP can be a powerful signal of extremely recent selection in large cohorts. In addition to recent and ongoing selection, it has been demonstrated that these methods have compelling advantages to detecting selection from standing variation [38, 106, 80]. However, these methods share the major limitation of SFS-based method in that they are not suitable for parametric inference and it is unclear how to establish significance without use of an empirical null model.

Recently, supervised machine learning methods have been proposed as an alternative to traditional summary-statistic based methods (see e.g., [100]). Standard machine learning techniques applied to population genomic data afford some major advantages over methods based on summary-statistics: standard techniques can produce accurate classi-

fiers based on summaries of the data that live in much higher-dimensional space than the aforementioned summary statistics, and these techniques often encompass a wide space of classification functions that are often non-linear (see e.g. [104, 101]). Some studies have demonstrated these methods can have improved robustness to demographic model misspecification [96, 106]. Although these methods can potentially detect complex patterns left by selection, they accordingly demand a great deal more training data or otherwise risk overfitting.

In contrast to the aforementioned methods, one might aim to develop a full likelihood methods which would take into account the full data set, rather than merely summary statistics. A common strategy for obtaining the full likelihood has been to find the distribution of the genealogy under selection. For example, Krone and Neuhauser described the distribution of the coalescence tree of a locus under weak selection and no recombination [82]. Alternatively, one can describe how the genealogy depends on the trajectory of the selected allele (first described by [120]), and in turn how the trajectory depends on selection. To this end, Coop and Griffiths [84] developed a sampling method for approximating the full likelihood of the selection coefficient. Their method uses sampling to marginalize out two layers of latent variables: the allele frequency trajectory and the genealogy of the locus. To estimate the likelihood function, they perform random sampling of both the trajectory, and the genealogy conditioned on the trajectory. Unfortunately, methods that consider the both coalescence and recombination are generally considered computationally intractable.

Composite likelihood methods (see e.g. [41, 89]) are able to approximate the likelihood function using tractable expressions for the frequency distribution of a neutral site linked to the selected site [86, 87]. These methods approximate the joint distribution of frequencies observed at linked sites as the product of their marginals. These approaches can be applied to test for selection, and estimate the strength of selection. The approximations made by composite likelihood methods are more accurate under strong selection (arguably beyond the strength of most recent selection in humans), and thus have less power to detect weak selection — although to some extent low power to detect weak selection is a natural outcome of any selection method.

Approximate Bayesian computation (ABC) and rejection sampling methods approximate the likelihood function by simulation. One advantage over the composite likelihood approach is that ABC can capture dependencies between linked neutral sites. For example, methods have been used to jointly infer the strength and timing of selection acting on a locus and determine whether a sweep occurred from a *de novo vs* standing variant [93, 94, 121, 122]. However, a major disadvantage of such approaches is that the amount of simulation necessary to obtain an accurate estimate grows dramatically with the dimensionality of the observed data (for a discussion, see e.g. [97]); similar issues arise in the process of training machine learning methods (e.g. [106]), requiring considerations to prevent overfitting and avoid excessive simulation.

The method we present in this paper draws inspiration from the Coop & Griffiths method [84], and has several key similarities: our method produces a likelihood and

involves integrating out the trajectory and genealogy, i.e., the aforementioned two hidden layers. However, there are several key differences between this method and our approach: while Coop and Griffiths assume no recombination of the locus, our method is based on the coalescent with recombination (i.e. the ancestral recombination graph or ARG) [115]. Also, whereas Coop & Griffiths simulate random trajectories, we use a hidden Markov model (HMM) to completely marginalize the latent trajectory. Lastly, our method uses a novel importance sampling scheme that allows us to sample ARGs assuming a neutral prior, and find the likelihood function at arbitrary values of s; this drastically reduces the amount of ARG sampling necessary.

Furthermore, the new method is, to our knowledge, the first that is capable of inferring the allele frequency trajectories for models with recombination and selection using only modern data. We are able to accomplish this task using the aforementioned Markovian structure of both coalescence and the trajectory, forming a HMM over these two hidden states and solving for the posterior marginals of each hidden allele frequency state over time. Recently, Edge & Coop proposed a method to reconstruct changes to polygenic scores over time via such estimates of the local trees, but their method is not suitable for estimating allele frequency changes or selection at individual loci [123].

2.2 Materials and methods

Overview

We begin with an overview of our method for jointly inferring selection and the allele frequency trajectory, which we summarize in Fig. 2.1. Our method begins with input in the form of haplotype data (Fig. 2.1A), although technically, it is also possible to use unphased data, and sample possible phasings.

Next, we sample the posterior distribution on the genealogy at the selected site (Fig. 2.1B); in other words, we marginalize out the hidden coalescence events, the first of two latent variables or "hidden layers" in our model. Specifically, we sample the full ancestral recombination graph (ARG) of the input haplotypes. The ARG is a graph that summarizes all of the common ancestry and recombination events that have occurred within the sample. We sample ARGs rather than gene trees in order to account for recombination, and to incorporate information from sites in long-range linkage disequilibrium with the selected site. Then we extract the genealogy at the site of interest (the "local tree") and from here on, this is the only component of the ARG that goes into our subsequent calculations. To perform ARG sampling, we choose to use ARGweaver [115], which is the only currently available method to sample the posterior ARG. In practice, it is possible and straightforward to adapt this method to other ARG inference methods designed for larger samples, but sampling the posterior yields beneficial statistical properties (see "Importance Sampling" under Materials and Methods).

Then, for each local tree we have sampled, we form a hidden Markov model (HMM,

Fig. 2.1C) where observed states are coalescences in this local tree, and hidden states are the selected allele's frequency trajectory over time (i.e., the second hidden layer of our overall model). We use a discrete-time model of the coalescent process to match the model used by ARGweaver, so that the length of the HMM is of manageable, finite length. Emission probabilities (i.e., coalescence probabilities) depend both on the frequency and the most recent prior emission, whereas transition probabilities depend on the selection coefficient s, the parameter we are ultimately interested in estimating. Solving the HMM yields the probability of the sample local tree as a function of s. To obtain the likelihood function of s, we perform importance sampling over all of sample trees, reweighting their coalescent probabilities and summing them up. This approach allows us to use trees sampled exclusively under a prior of selective neutrality ($s = 0$) to calculate the likelihood function at arbitrary values of s. In other words, this approach allows us to minimize the amount of ARG sampling necessary to estimate the likelihood function, which is notable because ARG sampling is generally the most computationally intensive step of our method.

Finally, we can analyze the results to test for selection or estimate the selection coefficient (Fig. 2.1D). Additionally, we show that we can decode the HMMs depicted in Fig. 2.1C and use them to obtain a posterior estimate of the allele frequency trajectory (Fig. 2.1E).

Coalescent model for a site under selection

First, let us consider how the distribution of the local tree T at a site under selection depends on the frequency trajectory of an allele at that site. We assume that the tree is labeled, i.e. we know which branches subtend each allele. We also assume the tree to be compatible with the infinite sites assumption, i.e. that there is at most one mutation event that has occurred at the focal site, and thus the site is bi-allelic. We model the likelihood of the tree using a structured coalescent; moving backwards in time from the time of sampling until the time of the mutation, lineages can only coalesce with other lineages that subtend the same allele, and the coalescence rate within the derived and ancestral classes depends on both the derived allele frequency $X(t)$ and the effective population size $N(t)$, both indexed by the time $t \geq 0$ in coalescent units before the present day. Proceeding back in time, lineages coalesce freely after the time of mutation, and the coalescence rate depends only on $N(t)$. In the rest of this section we treat the trajectory $X(t)$ as known, but in practice the trajectory is hidden and highly stochastic; in a later section we develop a hidden Markov model to efficiently integrate out $X(t)$.

We use a discrete-time model of the coalescent employed also by ARGweaver [115]. That is, we only observe the coalescent process at a discrete set of timepoints $\{t_1, \ldots, t_K\}$, and also make the additional assumption that all lineages must coalesce by t_K. (Typically t_K is set to $\sim 100 \times N_e$, implying coalescence would be extremely unlikely to occur after t_K, and hence this assumption is very reasonable.) Henceforth, using this discretization

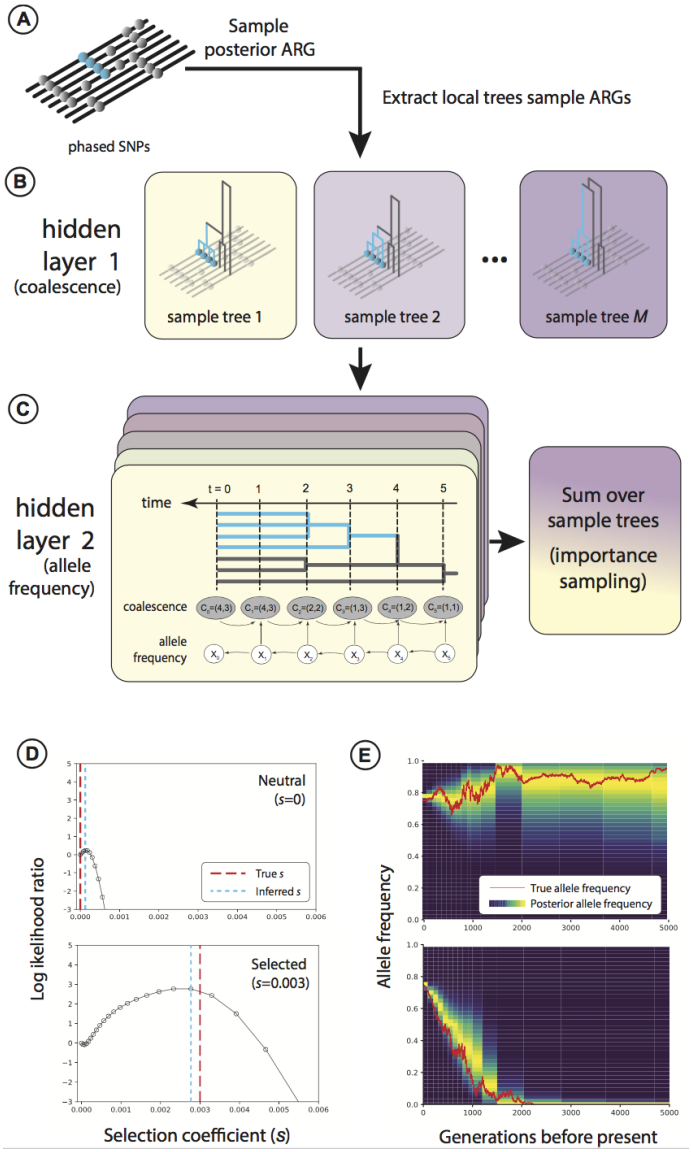


Figure 2.1: A: To apply our method for inferring selection, we begin by sampling the posterior ARG of a set of recombining chromosomes. B: For each sample ARG, we extract local trees at the site of interest (blue). C: For each sample local tree, we run an HMM to calculate the likelihood of selection, marginalizing out the hidden allele frequency trajectory based on coalescence in the sample tree. We later use the recursions performed in this step to calculate the posterior allele frequency trajectory. D: An example of the estimated likelihood function for an allele under neutrality (top) and selection (bottom). E: An example of the inferred allele frequency trajectory compared to the ground truth trajectory under neutrality (top) and selection (bottom). Both (D) and (E) are inferred from data simulated under a European demographic model with $n = 50$ haplotypes, conditioning on the derived allele segregating at 75% in the present day. with $s = 0$ and $s = 0.003$, respectively.

we also discretize X and N; we assume $X(t) = X_i$ for $t \in (t_i, t_{i+1}]$, and $N(t) = N_i$ for $t \in (t_i, t_{i+1}]$.

We use C to track the number of lineages remaining at these timepoints leading back into the past; as long as we keep track of the number of lineages belonging to each of the allelic classes, by exchangeability of fitness within an allelic class, we can model the likelihood function in the usual way, as independent of the topology given the waiting times. Hence, we define three simultaneous, related processes $C = (C^{\text{der}}, C^{\text{anc}}, C^{\text{mix}})$. The processes C^{der} and C^{anc} refer to coalescence within the derived and ancestral classes during the time going back from the time of sampling to the time of the mutation. The mixed process C^{mix} refers to coalescence going backwards from the time of the mutation. We call it the mixed process because it includes un-coalesced lineages from C^{anc}, as well as the lineage ancestral to all derived lineages. Assuming the infinite sites model, C^{mix} will have one additional lineage relative to C^{anc} at the time of the mutation, and will eventually reach $C^{\text{anc}} = C^{\text{mix}}$ once that lineage coalesces with one of the other lineages in the ancestral class. In Fig. 2.2 we illustrate the lines-of-descent process in the these three classes.

We model the probability of transitioning from $C_i \to C_{i+1}$ lineages during some time interval $[t_i, t_{i+1}]$ using a simple variation on Tavare's formula for the exact distribution of the number of lines of descent remaining after t generations [124] We use Tavare's formula in order to model the coalescent at discrete timepoints, allowing multiple coalescences at each epoch.

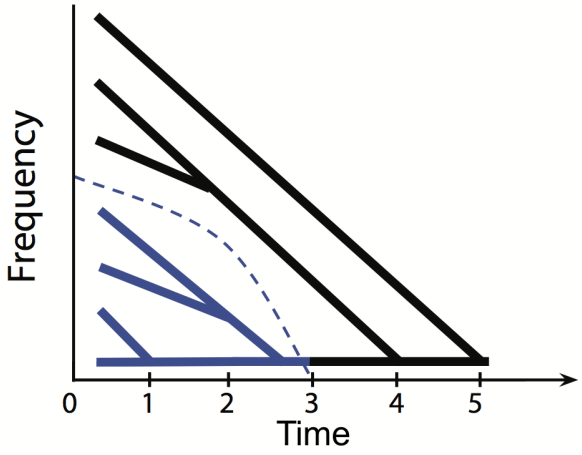
We write the likelihood of a trajectory X given C as

$$\mathbb{P}(C \mid X, N) = \prod_{i=0}^{K-1} \mathbb{P}(C_{i+1} \mid C_i, X_i, N_i) \tag{2.1}$$

More precisely, in terms of the derived, ancestral, and mixed processes,

$$\mathbb{P}(C \mid X, N) = \prod_{i=0}^{i^*-1} \mathbb{P}(C^{\text{der}}_{i+1} \mid C^{\text{der}}_i, X_i, N_i)\, \mathbb{P}(C^{\text{anc}}_{i+1} \mid C^{\text{anc}}_i, X_i, N_i) \times \prod_{i=i^*}^{K-1} \mathbb{P}(C^{\text{mix}}_{i+1} \mid C^{\text{mix}}_i, N_i) \tag{2.2}$$

where $i^* := \max\{i : X_i > 0\}$ denotes the index of the epoch during which the allele arose via mutation. Naturally, the mixed process—which we only keep track of while the derived allele is nonexistent—does not depend on X. We can write the transition probabilities using Tavare's formula [124]:



Time	0	1	2	3	4	5
C^{der}	4	3	2	1	1	1
C^{anc}	3	3	2	2	2	1
C^{mix}	4	4	3	3	2	1

Figure 2.2: Top: Coalescence conditioned on the allele frequency trajectory (dashed blue line). Blue lineages subtend the derived allele, whereas black lineages do not. Black lineages belong to the ancestral class while the derived allele has $X_t > 0$, and they belong to the mixed class while $X_t = 0$. Bottom: the numbers of derived, ancestral, and mixed lineages at each time point. Black numbers factor into the likelihood calculation, whereas gray numbers do not.

$$\mathbb{P}(C_{i+1}^{\text{class}} = b \mid C_i^{\text{class}} = a, Z_i = z_i^{\text{class}}) = \sum_{k=b}^{a} \left\{ \exp\left(\frac{-\binom{k}{2}(t_{i+1} - t_i)}{2z_i}\right) \times \frac{(2k-1)(-1)^{k-b}}{b!(k-b)!(k+b-1)} \prod_{l=0}^{k-1} \frac{(b+l)(a-l)}{(a+l)} \right\}$$

where

$$z_i^{\text{class}} = \begin{cases} N_i X_i & : \quad \text{class} = \text{der} \\ N_i(1 - X_i) & : \quad \text{class} = \text{anc} \\ N_i & : \quad \text{class} = \text{mix} \end{cases} \tag{2.3}$$

We note that this formula is known to be computationally unstable for large values of C, large values of N, and/or small values of $\Delta t_i = t_{i+1} - t_i$; under such conditions, the asymptotic distribution of $C_{i+1} \mid C_i = a$ (where a is, e.g., the number of derived lines of descent present at t_i) takes on a normal distribution [125]:

$$C_{i+1} \mid C_i = a \sim \mathcal{N}(\mu(\Delta t), \sigma^2(\Delta t)) \tag{2.4}$$

where

$$\mu(\Delta t) = \frac{2\eta}{\Delta t} \tag{2.5}$$

and

$$\sigma^2(\Delta t) = 2\eta/\Delta t(\eta+\beta)^2(1+\eta/(\eta+\beta)-\eta/\alpha-\eta/(\alpha+\beta)-2\eta)\beta^{-2} \tag{2.6}$$

where $\alpha = a\Delta t/2$, $\beta = -\Delta t/2$, and $\eta = \alpha\beta/[\alpha(e^{\beta}-1)+\beta e^{b}]$ [125]. In practice, for samples of $n = 50$ haplotypes under constant $N_e = 10^4$, we find this approximation is unnecessary; however, for the same sample size under a European demographic model, which exhibits very large recent N_e, we find it necessary to use this approximation during the roughly 10^3 generations preceding the present day, prior to which N_e, C, and our time discretization (and hence Δt) are sufficiently small that we change over to Tavare's exact formula [126].

Allele frequency transition probabilities

Our likelihood calculations require allele frequency transition distributions for different selection coefficients, population sizes, and spans of time. Rather than employ the more common approach of numerically calculating allele frequency transition distributions using the Wright-Fisher diffusion process with drift and selection (e.g., [7, 127]), we follow [128] and precompute allele frequency transition distributions on a grid of time spans (i.e., generations) and scaled selection coefficients (i.e., $\alpha = 2Ns$) using the Wright-Fisher model of reproduction in a finite population experiencing genetic drift and natural selection (see [7]). Specifically, for each value of α, we use simple matrix multiplication to produce allele frequency transition matrices for discrete frequencies in a haploid population of size $N = 2000$ at a number of generations spanning from $g = 1$ to $g = g_{\max}$ (corresponding to scaled drift times of $1/2000$ to $g' = g_{\max}/2000$) with some spacing chosen a priori; in practice, we use linear spacing for recent history and/or periods of population growth. We bin allele frequencies into d discrete frequency categories unevenly distributed between 0 and 1 such that extreme frequency bins outnumber intermediate frequency bins. To calculate allele frequency transition distributions for time spans and selection coefficients not contained in the grid of pre-computed values, we linearly interpolate between the nearest precomputed values. See [128] for details. Additionally, we condition the allele frequency process on the present-day frequency X_0 by using the following reweighting:

$$\mathbb{P}(X_i \mid X_{i+1}, X_0, s) = \frac{\mathbb{P}(X_i \mid X_{i+1}, s)\mathbb{P}(X_0 \mid X_i, s)}{\mathbb{P}(X_0 \mid X_{i+1}, s)}$$

where $\mathbb{P}(X_{i_1} \mid X_{i_2}, s)$ is the forward-time unconditional probability of transitioning from X_{i_2} to X_{i_1} (in coalescent time, $t_{i_2} > t_{i_1}$; in forward time, $t_{i_2} < t_{i_1}$).

Marginalizing the hidden allele frequency states

In the previous sections we showed how we obtain $\mathbb{P}(C \mid X)$ and $\mathbb{P}(X \mid s)$. The full likelihood of selection given the local tree G is thus

$$\mathrm{L}(s \mid G) \propto \mathbb{P}(C \mid s) = \sum_{x \in \mathcal{X}} \mathbb{P}(C \mid X = x)\mathbb{P}(X = x \mid s). \tag{2.7}$$

Naively, this involves a prohibitively large sum over d^{K-1} terms in $\mathcal{X}$, the space of possible trajectories. But due to the conditional independence of the likelihood, we can calculate the likelihood much faster using a recursion:

$$b_1(x_1) = \sum_{x_0} \mathbb{P}(C_1 \mid C_0, X_0 = x_0, N_0)\, \mathbb{P}(X_0 = x_0 \mid X_1 = x_1, N_0, s) \tag{2.8}$$

$$b_{i+1}(x_{i+1}) = \sum_{x_i} b_i(x_i)\, \mathbb{P}(C_{i+1} \mid C_i, X_i = x_i, N_i)\, \mathbb{P}(X_i = x_i \mid X_{i+1} = x_{i+1}, N_i, s) \tag{2.9}$$

and we can apply this recursion to calculate the likelihood function of s given G as

$$\mathrm{L}(s \mid G) \propto b_K(0). \tag{2.10}$$

The above is commonly known as the backward algorithm. In our model, the backward algorithm's recursion proceeds backwards through time. Alternatively, using the forward algorithm, with its recursion proceeding forwards in time:

$$f_{K-1}(x_{K-1}) = \mathbb{P}(X_{K-1} = x_{K-1} \mid X_K = 0, N_{K-1}, s) \tag{2.11}$$

$$f_{i-l}(x_{i-1}) = \mathbb{P}(X_{i-1} = x_{i-1} \mid X_i = x_i, N_{i-1}, s) \sum_{x_i} f_i(x_i)\mathbb{P}(C_{i+1} \mid C_i, X_i = x_i, N_i) \tag{2.12}$$

and we can apply this recursion to calculate the likelihood function of s given G as

$$\mathrm{L}(s \mid G) \propto \sum_{x_0} f_0(x_0) \tag{2.13}$$

To calculate the posterior probability of the allele frequency during the ith epoch X_i,

$$\mathbb{P}(X_i = x_i \mid C, s) = \frac{b_i(x_i) f_i(x_i)}{\sum_{x_i'} b_i(x_i') f_i(x_i')} \tag{2.14}$$

gives the posterior marginal of X_i using the familiar forward-backward algorithm.

Importance sampling to estimate the likelihood function

The above formulas pertain immediately only to the case in which the local tree is observed directly and without noise. In practical settings, the local tree is hidden to us and we must integrate over the space of possible local trees using sampling methods. Here we describe a novel importance sampling method to reweight posterior samples of the ARG to approximate the likelihood function of selection. Although we use s to express the argument of the likelihood function, we use this as shorthand for estimating the likelihood function of arbitrarily complex parameters; for example, one could estimate the selection coefficient s, as well as the time of selection's onset, t_s, before which the allele behaved neutrally.

We are given haplotype data D representing n haplotypes with l sites that are fixed for the derived allele. We wish to use D to infer the maximum-likelihood value of s for some locus $k \in \{1, 2, \ldots, l\}$ assuming that all other loci are selectively neutral (i.e. $s_j = 0 \ \forall j \in \{1, 2, \ldots, k-1, k+1, \ldots, l\}$). In other words, we restrict ourselves to testing simple hypotheses of the form "site k has selection coefficient s_k and all of its flanking sites are selectively neutral."

The likelihood of s under the data can be expressed as the expected value of the likelihood of the ARG $\mathcal{G}$ given the data D, with respect to the distribution of $\mathcal{G}$ given s:

$$\mathrm{L}(s) = \mathbb{E}_{\mathcal{G}|s}[\mathbb{P}(D \mid \mathcal{G}, s)] \tag{2.15}$$

At this stage, we introduce G, the discrete-time approximation of $\mathcal{G}$ (discussed in more detail by [115]), and we assume

$$\mathrm{L}(s) = \mathbb{E}_{G|s}[\mathbb{P}(D \mid G, s)] \tag{2.16}$$

By importance sampling, we are able to express the expectation over an alternative distribution $q(G)$, as long as $\mathbb{P}(G, D \mid s) > 0 \Rightarrow q(G) > 0$. Notice that this implies we can conduct sampling under $q(G)$ once, and reweight these samples for arbitrary values of s without having to conduct additional sampling. In other words, approximating $L(s)$ using importance sampling does not require sampling under each value of s at which you want to approximate $L(s)$.

In this paper we specifically consider the estimator given by $q(G) = \mathbb{P}(G \mid D, s = 0)$; i.e., the posterior ARG under selective neutrality. Later, we evaluate the performance of the estimator using the Markov chain Monte Carlo method ARGweaver, which samples from the posterior [115]. One can obtain the importance sampling estimate of the full likelihood $L(s)$ by expressing Eq. 2.16 as an expectation over a different distribution, i.e. the posterior distribution of the ARG (assuming selective neutrality):

$$L(s) = \mathbb{E}_{G|s}[\mathbb{P}(D \mid G, s)] = \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(D \mid G, s)\mathbb{P}(G \mid s)}{\mathbb{P}(G \mid D, s = 0)}\right] \tag{2.17}$$

We can express Eq. 2.17 using the Monte Carlo approximation

$$\widehat{L}(s) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{P}(D \mid G^{(m)}, s)\frac{\mathbb{P}(G^{(m)} \mid s)}{\mathbb{P}(G^{(m)} \mid D, s=0)} \to L(s) \tag{2.18}$$

where $G^{(m)} \sim \mathbb{P}(G \mid D, s=0),\ m = 1, 2, \ldots, M$, and "$\to$", here and in the following, means that the left-hand side converges almost surely to the right-hand side as M goes to infinity, assuming that a Law of Large Numbers for ergodic processes holds (the Birkhoff–Khinchin theorem).

Hence, if we sample genealogies from the posterior under selective neutrality, that is, $G^{(m)} \sim \mathbb{P}(G \mid D, s=0),\ m = 1, 2, \ldots, M$, then the right-hand side of Eq. 2.18 can be used as a Monte Carlo estimator of the likelihood function. However, in practice this estimator is highly unstable. However, a more stable estimator of the likelihood ratio $\frac{\mathrm{L}(s)}{\mathrm{L}(s=0)}$ can be derived. We can divide through Eq.2.17 by $L(s=0) = \mathbb{P}(D \mid s=0)$ to get

$$\frac{\mathrm{L}(s)}{\mathrm{L}(s=0)} = \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(D, G \mid s)}{\mathbb{P}(D, G \mid s=0)}\right] \tag{2.19}$$

Because we assume the data are conditionally independent of selection given the full ARG, we can simplify this as

$$\frac{\mathrm{L}(s)}{\mathrm{L}(s=0)} = \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(D \mid G)\mathbb{P}(G \mid s)}{\mathbb{P}(D \mid G)\mathbb{P}(G \mid s=0)}\right] = \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(G \mid s)}{\mathbb{P}(G \mid s=0)}\right] \tag{2.20}$$

A key development in our method is that although we sample the ARG of the entire sequence, we only calculate likelihoods using the marginal tree at the selected site, which we will call G_k. In doing so, we make a key approximation: for differing sweep parameters s and s', we assume that

$$\mathbb{P}(G_{\setminus k} \mid G_k, s) \approx \mathbb{P}(G_{\setminus k} \mid G_k, s') \tag{2.21}$$

That is, we assume that the rest of the ARG is approximately conditionally independent of s given the marginal tree at the selected site, G_k. Thus, we can reduce Eq. 2.20 to

$$\begin{aligned}
\frac{\mathrm{L}(s)}{\mathrm{L}(s=0)} &= \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(G \mid s)}{\mathbb{P}(G \mid s=0)}\right] \\
&= \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(G_{\setminus k} \mid G_k, s)}{\mathbb{P}(G_{\setminus k} \mid G_k, s=0)}\frac{\mathbb{P}(G_k \mid s)}{\mathbb{P}(G_k \mid s=0)}\right] \\
&\approx \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(G_k \mid s)}{\mathbb{P}(G_k \mid s=0)}\right]
\end{aligned}$$

which suggests the following importance sampling estimator using genealogies sampled from ARGweaver will converge almost surely to a close approximation to the likelihood ratio:

$$\widehat{\mathrm{LR}}(s) = \frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{P}(G_k^{(m)} \mid s)}{\mathbb{P}(G_k^{(m)} \mid s=0)} \to \mathbb{E}_{G|D,s=0}\left[\frac{\mathbb{P}(G_k \mid s)}{\mathbb{P}(G_k \mid s=0)}\right] \approx \frac{\mathrm{L}(s)}{\mathrm{L}(s=0)} \tag{2.22}$$

where $G^{(m)} \sim \mathbb{P}(G|D, s=0)$ for $m = 1, 2, \ldots, M$.

Finally, due to exchangeability of lineages within the derived and ancestral allelic classes, we can assume

$$\mathbb{P}(G_k \mid s) \propto \mathbb{P}(C_k \mid s) \Rightarrow \widehat{\mathrm{LR}}(s) = \frac{1}{M}\sum_{m=1}^{M}\Omega^{(m)}(s) \tag{2.23}$$

where

$$\Omega^{(m)}(s) := \frac{\mathbb{P}(C_k^{(m)} \mid s)}{\mathbb{P}(C_k^{(m)} \mid s=0)} \tag{2.24}$$

denotes the summand of the importance sampling estimator. That is, the topology within allelic classes is not important, and instead we need only the lines of descent process within each class.

We can maximize the likelihood ratio over different values of s to obtain the maximum-likelihood estimate of s

$$\hat{s} = \mathrm{argmax}_s \widehat{\mathrm{LR}}(s) \tag{2.25}$$

Finally, we can obtain an importance sampling estimate of $\pi(x_i \mid D, s)$, the posterior marginal of the allele frequency at timepoint i, X_i:

$$\begin{aligned}
\pi(X_i \mid D, s) &= \mathbb{E}_{G|D,s}[\mathbb{P}(X_i \mid G, D, s)] && (2.26)\\
&= \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid G, D, s)\frac{\mathbb{P}(G \mid D, s)}{\mathbb{P}(G \mid D, s=0)}\right] && (2.27)\\
&= \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid G, D, s)\frac{\mathbb{P}(G \mid s)}{\mathbb{P}(G \mid s=0)}\right] \times \frac{L(s)}{L(s=0)} && (2.28)\\
&\propto \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid G, D, s)\frac{\mathbb{P}(G \mid s)}{\mathbb{P}(G \mid s=0)}\right] && (2.29)\\
&\approx \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid G_k, G_{\setminus k}, D, s)\frac{\mathbb{P}(G_k \mid s)}{\mathbb{P}(G_k \mid s=0)}\right] && (2.30)\\
&\approx \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid G_k, s)\frac{\mathbb{P}(G_k \mid s)}{\mathbb{P}(G_k \mid s=0)}\right] && (2.31)\\
&= \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid C_k, s)\frac{\mathbb{P}(C_k \mid s)}{\mathbb{P}(C_k \mid s=0)}\right] && (2.32)
\end{aligned}$$

Hence,

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{P}(X_i \mid C_k^{(m)}, s)\Omega^{(m)}(s) \to \mathbb{E}_{G|D,s=0}\left[\mathbb{P}(X_i \mid C_k, s)\frac{\mathbb{P}(C_k \mid s)}{\mathbb{P}(C_k \mid s=0)}\right] \approx \kappa\pi(X_i \mid D, s) \tag{2.33}$$

where κ is the constant $[L(s)/L(s=0)]^{-1}$. Thus, our importance sampling estimate of the posterior marginal given s is

$$\hat{\pi}(x_i \mid D, s) := \frac{\sum_{m=1}^{M}\mathbb{P}(X_i \mid C_k^{(m)}, s)\Omega^{(m)}(s)}{\sum_{m=1}^{M}\Omega^{(m)}(s)} \tag{2.34}$$

where in the summand we use the posterior marginal established in Eq. 2.14. In practice, we fix $s = \hat{s}$. A concern is, therefore, that this estimator does not take uncertainty in the estimate of s into account. This problem can be addressed by using a Bayesian approach and allowing a prior distribution on s, $\pi(s)$, the posterior of the selection coefficient $\pi(s \mid D)$ follows

$$\pi(s \mid D) \propto \frac{L(s)}{L(s=0)}\pi(s) \approx \widehat{\mathrm{LR}}(s)\,\pi(s). \tag{2.35}$$

Then the estimate of the posterior marginal is given by

$$\hat{\pi}(x_i \mid D) = \int_{-\infty}^{\infty}\hat{\pi}(x_i \mid D, s)\pi(s|D)ds \tag{2.36}$$

which can be approximated by a sum over d discretized values of s, $\mathcal{S} = \{s_1, \ldots, s_d\}$ as

$$\hat{\pi}(x_i \mid D) := \sum_{s\in\mathcal{S}}\hat{\pi}(x_i \mid D, s)\tilde{\pi}(s|D) \tag{2.37}$$

where $\tilde{\pi}$ represents a probability mass function over s. In this this paper we assume positive directional selection with a dominance coefficient of $h = 1/2$, but our method can be extended easily to general values of h as well as negative selection.

The method is implemented in a computer package, `CLUES`, available for download at https://github.com/35ajstern/clues.

Simulations

To evaluate the power of CLUES to determine whether a site has been subject to selection, we simulated a dataset of $n = 25$ diploid individuals under two different demographic models; (1) a model of constant effective population size ($N = 10^4$), and (2) a model of European (CEU) demography [129]. We performed both sets of simulations using the

program `discoal` [130]. We set $\mu = 2r = 2.5 \times 10^{-8}$ mut/bp/gen, $L = 1 \times 10^5$ bp or 2×10^5 bp for the constant-size and CEU models, respectively, and simulated conditional on a variety of present-day frequencies and selection coefficients, the latter of which we ranged from weak to strong values. Under each condition, we simulated 100 independent iterations. We also sampled 1 ancient haplotype; because ARGweaver, which we used subsequently to sample the posterior ARG, does not incorporate any information about ancestral/derived states, it is best practice to add an ancient individual or outgroup to help polarize the the alleles. For the constant-size and CEU models, we used ancient sampling dates of 2×10^4 and 1.6×10^4 generations before present, respectively. Because `discoal` can only simulate piecewise-constant population sizes, we specified population sizes to take on the value of their harmonic mean over the epoch, calculated from the original CEU model.

Importantly, we conditioned simulations on the site of interest segregating at a particular frequency in the present day. Hence, when we considered the power to discriminate between neutral and selected alleles, we controlled the present-day frequency to be equal in both of these cases. Avoiding this step would otherwise upwardly bias estimates of the statistical power, due simply to the tendency for selected alleles to segregate at higher frequencies than neutral alleles [23]. (If the allele frequency in itself is also of interest, this part of the likelihood could trivially be added at a later stage, by simply using the stationary distribution of the allele frequency; see "Allele frequency transition probabilities" under Materials and Methods.) We then simulate the allele frequency backwards in time, from the present-day frequency, until the allele reaches a frequency of 0. Simulators such as `discoal` achieve this by using the conditional Wright-Fisher diffusion (see e.g. [131]). In the case where effective population size changes over time, running conditional simulations requires additional considerations because the probability of a mutation entering the population scales approximately linearly with population size. Naively sampling the trajectory backwards in time will therefore produce a bias, unless trajectories where the mutation occurs while N_e is low are somehow penalized. Thus, approaches such as reweighting sample trajectories using importance sampling have been used to correct this bias [132]. The program `discoal` implements a similar bias-correcting scheme using rejection sampling that rejects trajectories where the mutation occurs while N_e is low with higher probability than trajectories where the mutation occurs while N_e is high.

Next, we inferred the posterior ARG given the sequence data we simulated using ARGweaver [115]. This method works by proposing adjustments to an initial ARG, and randomly accepting or rejecting these proposals based on calculations of the prior probability of the proposed ARG, as well as its likelihood given the sequence data. Because the prior probability is based on the effective population size, we specified the same effective population size in the prior as we used to generate the sequence data. We found it important to adjust the proposal mechanism of ARGweaver; specifically, we adjusted resample window size and the number of resamples per window to achieve an acceptance rate of about 30-70%. In total, we sampled 3×10^3 ARGs for each simulation, discarding

the first 1×10^3 as a burn-in period, and subsequently thinning the remaining samples to reduce the computational burden of downstream analyses; we used a thinning rate of 100 samples, resulting in $M = 20$ approximately independent samples. Reducing the thinning rate would increase accuracy of the inference at the cost of additional computation to calculate the likelihood of each additional sample tree.

Using utilities in the ARGweaver package, we extracted local trees at the selected site (at the center of the locus) from these sample ARGs. We then analyze this final set of trees using CLUES . We also analyzed the same sequence data using nSL, H_{12}, and Tajima's D [79, 81, 62]. The nSL method is essentially equivalent to iHS [78], except nSL does not require specifying a genetic map; despite this, these methods have been shown to have very similar statistical power with a slight advantage of nSL under some conditions. H_{12} is a method to calculate haplotype homozygosity merging the two most common haplogroups; thus, it is a test for selection that is robust to the origin of a sweep, i.e. whether it is hard or soft. Tajima's D is a site frequency spectrum-based statistic which is sensitive to skews in the frequency distribution of linked alleles caused by hitchhiking on the partially swept selected allele. We used scripts provided by [106] to calculate D and H_{12}, using a window size of 100kb centered on the selected site. We compare testing for selection under these methods by comparing their power curves under both the constant N_e and CEU demography models (Figs. 2.3,2.4).

We also conducted a similar simulation study for detecting recent selection starting 100 generations ago. We simulated under the same CEU demographic model as previously described, but instead sample $n = 50$ diploids. We conducted ARG sampling and thinning as previously described, but in our analysis of the sample trees using CLUES , we calculated the likelihood for models of selection where $s = 0$ up until 100 generations ago, and $s \geq 0$ from that point until the present day. This sweep from standing variation (SSV) model differs from the hard sweep model we used previously, which assumes s is constant throughout history. Instead of optimizing the likelihood function just for s, we optimized jointly over two parameters, s and the onset of selection t_s, the latter of which represents the time of the onset of selection.

2.3 Results

Testing for selection

We found that across all scenarios, CLUES matches or exceeds the statistical power of the other methods evaluated (Figs. 2.3,2.4). As expected, all methods had highest power under large values of both the selection coefficient and the derived allele frequency (Fig 2.3I). Under these conditions, CLUES had 100% power at the 1% significance threshhold; the next most powerful method, nSL, had 68% power at the same significance level. CLUES also demonstrated improvement in power under weak selection; as the selection coefficient

was decreased, nSL retained about 20% power when $s = 0.003$ and <5% power when $s = 0.001$, and Tajima's D and H_{12} retained <5% power under both $s = 0.001, 0.003$ (Fig 2.3G,H). By contrast, CLUES retained approximately 45% and 90% power under $s = 0.001, 0.003$, respectively. We conclude that CLUES has high power across a wide regime of selection strengths, and has notably improved power over standard methods under weaker values of s.

We also considered the effect of present-day allele frequency on statistical power. Previous studies have shown a strong dependence of power on current allele frequency, with methods such as nSl and iHs having highest power at allele frequencies in the 70-90% range (see e.g. [78]). We tested for selection at alleles ranging in present day frequency from 25% to 75%, and while CLUES showed the expected pattern of increasing power with frequency, it also improved on the performance of other methods at lower frequencies. For example, under strong selection ($s = 0.01$), the power of CLUES changed from 100% to 90% to 85% as the frequency is decreased from 75% to 50% to 25% (Fig. 2.3C,F,I). By contrast, the power of the next most powerful method, H_{12}, dropped from approximately 65% to 45% to 15% (Fig. 2.3C,F,I). Under moderate selection ($s = 0.003$), these effects were even more drastic, with the power of CLUES and nSL (the next most powerful method in this regime) changing from 90% to 60% to 50% and 20% to 5% to <5%, respectively. We conclude that CLUES has high power compared to standard methods across a wide range of allele frequencies, with the most major improvements in performance occurring when the derived allele is at lower frequencies (<50%). We found that using the approximation due to Griffiths (Eq. 2.4, [125]) decreased power of CLUES by increasing variability of the null distribution of the likelihood ratios. Hence, for testing under nonequilibrium demography we used the exact lines-of-descent probabilities (Eq. 2.8). By contrast, as we will later show, we found the approximation given by Eq. 2.4 for $t \in [0, 1000]$ to improve estimation of allele frequency trajectories under this demographic model.

We also considered the same testing procedure under non-equilibrium demography, simulating under the previously described model of CEU demography (Fig. 2.32.4). We found in general reduced power to detect selection under this regime relative to the constant population size regime (Fig. 2.4I, cf. Fig. 2.3I), consistent with the well-known confounding of expanding population size with selection [41]. Nonetheless, CLUES demonstrated improved power relative to the competing methods across a wide range of selection coefficients (Fig. 2.4C,F,I), as well as across a wide range of derived allele frequencies (Fig. 2.4G,H,I).

Estimating selection coefficients

Using the simulations from the previous section to study statistical power in testing for selection, we used our estimate of the likelihood surface for s to estimate the value of the selection coefficient via maximum likelihood. We obtained selection coefficient estimates under importance sampling using ARGweaver (Fig. 2.5), as well as selection coefficient estimates based on the true local tree observed directly (Fig. A.1). Generally, the estimates

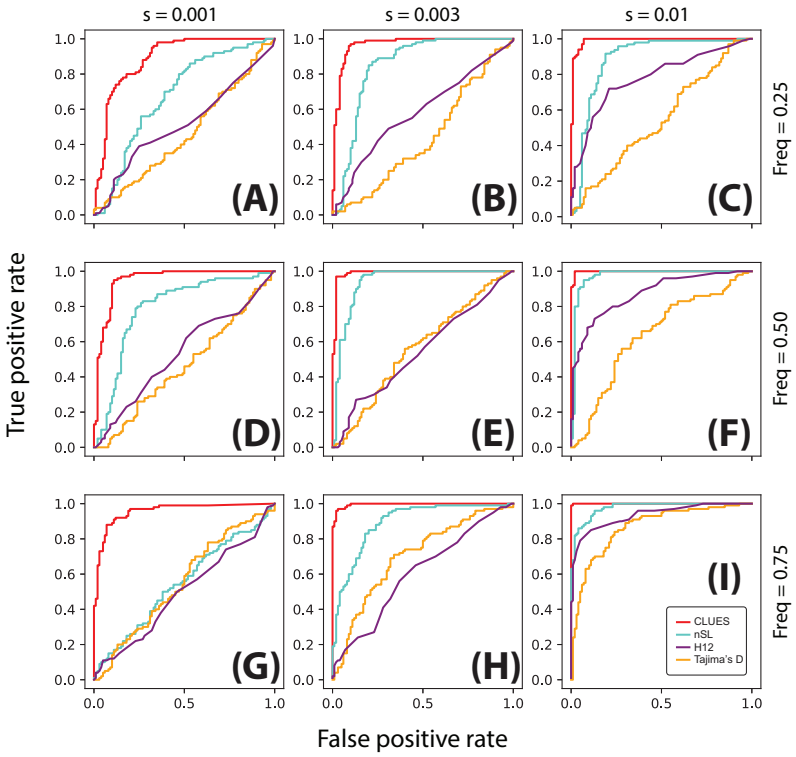


Figure 2.3: ROC curves illustrating performance of tests between selection and neutrality. Rows correspond to simulations conditioned on the same present-day allele frequency, and columns correspond to simulations with the same value of s. Simulations were performed under a model of constant effective population size ($N_e = 10^4$) using a locus of 100kb, $n = 25$ diploid individuals and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ recombinations/bp/gen.

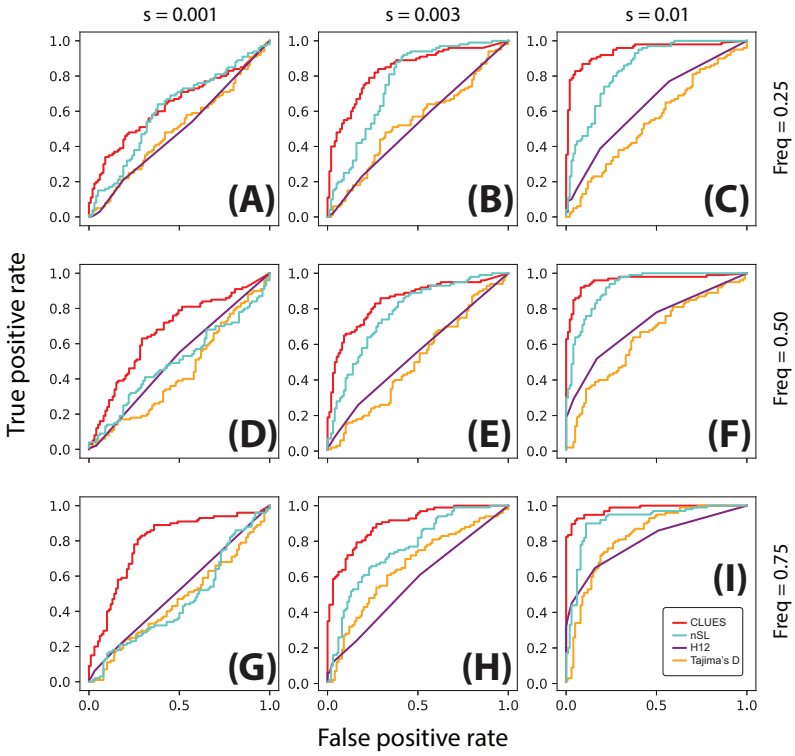


Figure 2.4: ROC curves illustrating performance of tests between selection and neutrality. Rows correspond to simulations conditioned on the same present-day allele frequency, and columns correspond to simulations with the same value of s. Simulations were performed under a model of European demography using a locus of 200kb, $n = 25$ diploid individuals and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ recombinations/bp/gen.

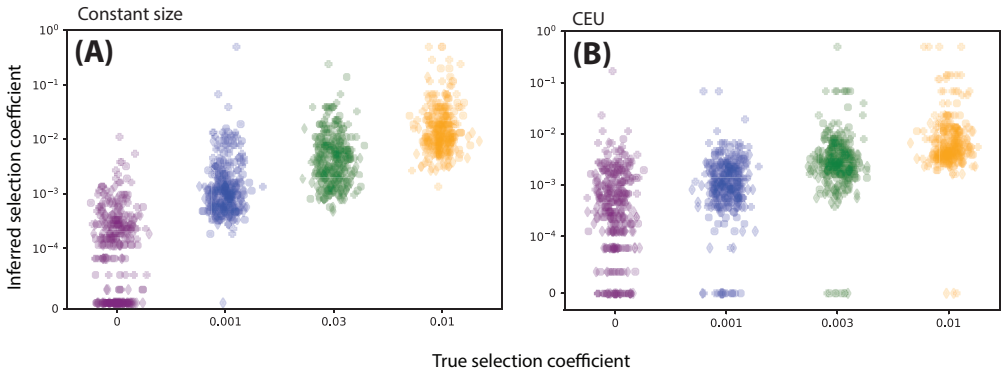


Figure 2.5: Inference of selection coefficients of varying strength using importance sampling method based on ARGweaver local trees. A: Constant population size. B: Tennessen CEU model. Marker shape denotes the present-day frequency conditioned upon in the simulation: +, 25%; ∘, 50%; ◇, 75%.

are approximately unbiased. For example, the mean estimates of $s = 0, 1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}$ were approximately $\overline{\hat{s}} = 1.9 \times 10^{-4}, 9.6 \times 10^{-4}, 3.2 \times 10^{-3}, 1.3 \times 10^{-2}$ when the present day frequency was fixed to 75% (Fig. 2.5A). Relative to inference when the true tree is observed, we found that the importance sampling estimates had increased variance, reflecting uncertainty in the tree. For example, we saw increased variability in the importance sampling vs. true tree estimates under constant population size (Fig. 2.5A vs. Fig A.1A), as well as under CEU demography (Fig. 2.5B vs. Fig. A.1B). This pattern is consistent with the additional uncertainty in s when the local tree is not observed directly. Notably, we found that importance sampling under a model of CEU demography yields estimates with a slight bias towards lower values of s, especially under strong selection (e.g. $s = 0.01$).

Inferring allele frequency trajectories

Using the same simulations and importance sampling estimates we obtained in the previous sections, we decoded the hidden Markov model (HMM) described in the section Materials & Methods. Specifically, we take $\hat{s}$, the maximum likelihood estimate of s, and plug it into the posterior marginal (Eq. 2.14) to obtain a probabilistic estimate of the allele frequency during a particular epoch; we do this independently for each epoch in our discrete-time model. To get a point estimate, we choose to use the posterior marginal mean; i.e., for each epoch, we choose the mean of the posterior marginal distribution. We illustrate the accuracy of these allele frequency trajectory estimates assuming the true local tree is observed and under importance sampling when the true tree is unknown in Fig. 2.6. We find that estimates of the allele frequency trajectory are generally unbiased for both true trees (Fig. 2.6 A,B) and importance sampling (Fig. 2.6 C,D), with increased

variance in the trajectory estimates in the importance sampling setting. We also illustrated variability in true vs. inferred trajectories controlling for s (Fig. A.6, here setting $s = 0$).

Whereas inference tended to be relatively accurate for high-frequency alleles (Fig. 2.6 B,D), when the derived allele was simulated conditioned on lower frequencies (e.g. 25%, Fig. 2.6 A,C), estimates tend to be downwardly biased. We tracked this bias to a lack of convergence in ARGweaver; specifically, we found that across different demographic scenarios and selection coefficients, ARGweaver can drastically overestimate the occurrence of very recent coalescences (in our case, in the last 100 generations; see Fig. A.5). Under constant population size, we see a nearly 7-fold excess in the number of recent coalescences inferred by ARGweaver. Naturally, this bias will affect estimates for low-frequency alleles more strongly, as fewer lineages subtend the derived allele, and thus a larger proportion of them are susceptible to this bias.

Because recombination rates vary substantially throughout the genomes of humans and other organisms, we also evaluated the accuracy of the estimates assuming $\mu = r$, larger than the $\mu = 2\rho$ setting we used in the other simulations, and estimation accuracy to be robust to this increase in recombination rate (Fig. A.2).

We also examined trajectory inference under non-equilibrium demography; i.e., the aforementioned model of CEU demography (Fig. A.3). Under the CEU model, we found trajectory estimates to have increased variance under importance sampling vs. true trees, but also a slight downward bias in estimating the selection coefficient under strong selection (i.e. $s = 0.01$; see Fig. 2.5B, Fig. A.3 D). As this bias does not occur under the true trees (Fig. A.1 B, Fig. A.3 B), we inspected the posterior trees sampled by ARGweaver for patterns consistent with this bias. We found that under this demographic model in particular, ARGweaver tends to under-sample trees with short times to most recent common ancestor (TMRCAs; see Fig. A.4). For reference, nearly 60% of runs under constant N_e contained even a single sample tree that had a TMRCA less than or equal to that of the true TMRCA (Fig. A.4 A). By comparison, under $s = 0.01$ and CEU demography, only 11% of ARGweaver runs met this criterion (Fig. A.4 B). Some bias is to be expected, as trees were sampled under a posterior distribution that assumes selective neutrality; however, these results suggest that, if ARGweaver is sampling from the true posterior assuming selective neutrality, then importance sampling estimates (of the selection coefficient, for example) will at least have much higher variance under the CEU model than under constant population size.

We further investigated whether uncertainty in s due to importance sampling variance drove the downward bias when estimating strong selection (Fig. 2.5B and Fig. A.3 D). First, we obtained importance sampling estimates of the trajectory fixing s to its true value (Fig. A.7 A). If uncertainty in s were the cause of the bias, then fixing the true value of s ought to correct for bias due to uncertainty. While we observe less bias in the estimates when fixing the true value of s, the bias is not totally eliminated. We observe a similar reduction in the bias of estimates under neutrality when we fix $s = 0$ (see Fig. A.6 B,E,H, vs. Fig. A.6 C,F,I). Thus, we conclude the bias is due to a lack of convergence in ARGweaver, which appears to be exacerbated in settings where strong selection is combined with

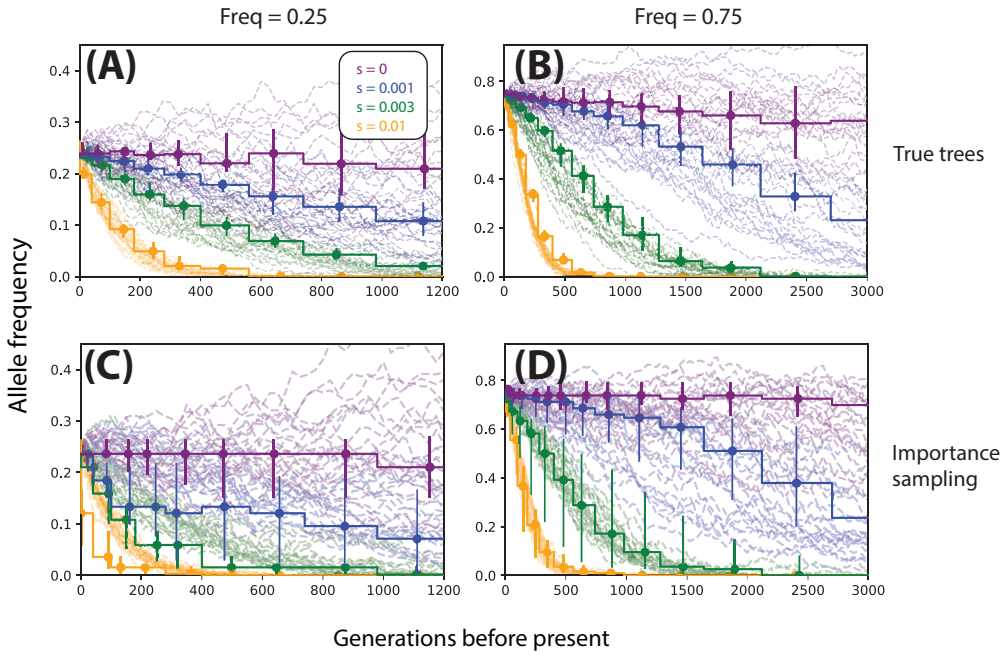


Figure 2.6: Allele frequency trajectories inferred from true trees (top row) and ARGweaver local trees (bottom row). Colored trajectories are inferred, black trajectories are the ground truth. Columns correspond to different initial allele frequencies (left: 25%, center: 50%, right: 75%), and rows/colors correspond to different selection coefficients. For each condition we show 25 randomly selected simulations and their corresponding inferences. All data are simulated under a model of constant effective population size ($N_e = 10^4$).

non-equilibrium demography.

We also investigated whether incorporating uncertainty in the estimate of s, rather than fixing $s = \hat{s}$, would improve the accuracy of trajectory inference. One strategy for modeling uncertainty in s is to apply a prior distribution to s. We found that marginalizing out s with respect to its posterior distribution (assuming a uniform prior on s) did not have a noticeable effect on inference for large values of s (Fig. A.7 B). This result is concordant with our observation that for large values of s, the likelihood surface peaks so strongly that the posterior remains tightly concentrated around the MLE $\hat{s}$. Hence, applying a prior distribution to s does not appear to be an adequate strategy to model uncertainty in s.

Inferring extremely recent selection

We applied our likelihood model of a sweep from a standing variant (SSV) to two types of datasets: selection from a standing variant starting 100 generations ago and selection with constant s (including $s = 0$), both described in 'Simulations' under Materials and Methods. We inferred trajectories under the best case scenario where the true trees are observed (Fig. 2.7A,B). We found that overall the method inferred the trajectory, as well as the strength and timing of selection, with highest accuracy when selection is strong (e.g. $s = 0.03$ in Fig. 2.7A,B). However, we found that as s took on smaller values ($s = 0.01$), many combinations of s and t_s had very similar likelihood (Fig. 2.7B), and thus estimates of s, t_s, and the allele frequency trajectory tended to be noisier than under very strong selection (Figs. 2.7A,B). Adding the extra parameter t_s did not cause overfitting when inferring the trajectories of hard sweeps (Fig. 2.7A). We also found good power to distinguish between hard vs. soft sweeps (i.e. sweeps from a standing variant), as apparent in the trajectories inferred in Fig. 2.7A. We calculated statistical power to test for a hard sweep using the statistic $\max_{s,t_s}\{L(s,t_s)\}/\max_{s}\{L(s,t_s=\infty)\}$; intuitively, this statistic is the ratio of the highest likelihood under any model with a SSV ($t_s \neq \infty$) to the highest likelihood of any hard sweep ($t_s = \infty$). At the 1% significance level we found 60% and 100% power to distinguish soft vs. hard sweeps with $s = 0.01, 0.03$, respectively.

We also performed importance sampling using ARGweaver and evaluated the power of the importance sampling estimates to detect recent selection vs. neutrality (Fig. 2.7C). Instead of comparing our method to nSL, which is not designed to detect signals of extremely recent selection, we compared to Singleton Density Score (SDS; [80]), as well as H_{12} and Tajima's D. We found that for lower values of s, all methods had generally low power. Although CLUES exhibited fairly high power (44%) to detect very strong recent selection ($s = 0.03$) —even outperforming SDS—we found that H_{12} has about the same power (45%) in this particular case. The lower power (<5%) of SDS is consistent with the fact that the method was explicitly designed to have high power for large datasets ($n > 1000$ for selection coefficients of this magnitude). Although we demonstrate that CLUES has substantial power to detect extremely recent selection, we found that importance sampling point estimates of s, t_s, and the trajectory were highly vulnerable to biases in the distribution sampled by ARGweaver (Fig. A.5). Specifically, we found that across various demographic and selection conditions, ARGweaver samples trees with substantially more recent coalescent events than in the true trees. Specifically, under the European demographic model with the settings used here to study recent selection, we find ARGweaver samples about a 4-fold excess of recent coalescent events (Fig. A.5 B). Clearly, this bias would produce a false signature of recent selection under neutral conditions. Thus, we did not further explore importance sampling estimates of s and the trajectory under the recent selection model. We conclude that potential ARG-sampling methods that avoid this bias will improve upon power to detect recent selection, as well as point estimates of the strength, timing of selection, and the allele frequency trajectory.

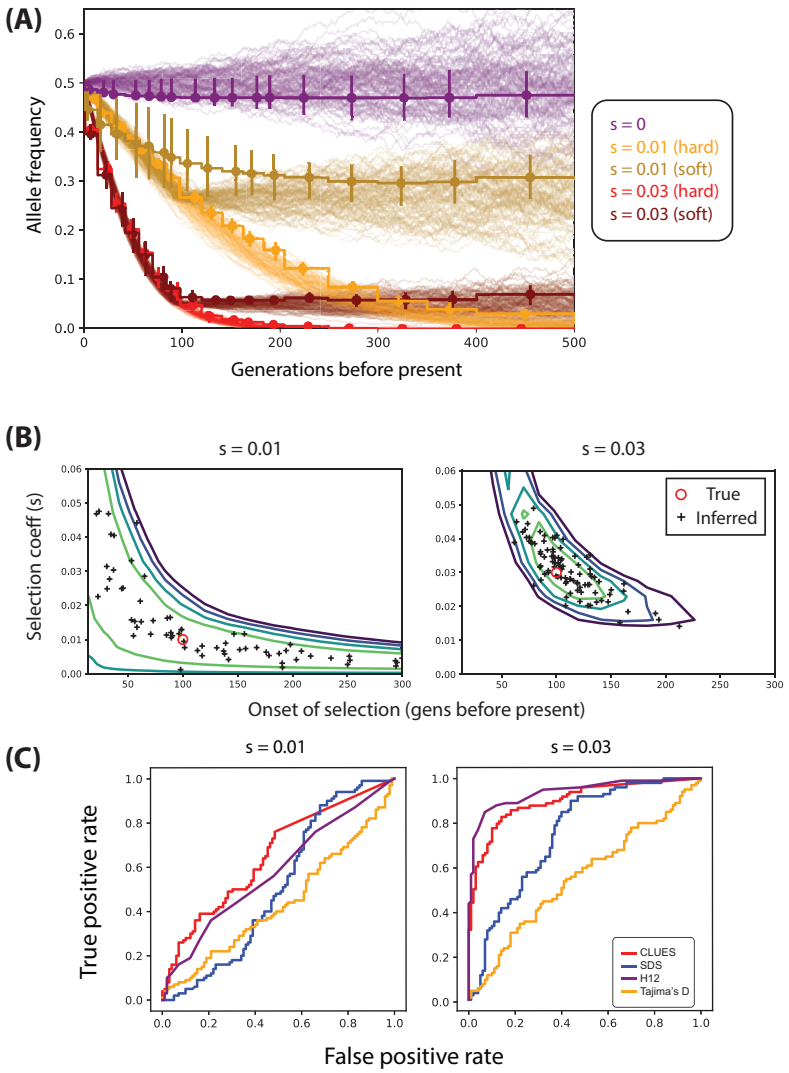


Figure 2.7: (A) Trajectories inferred from true trees under both hard sweeps and recent selection on a standing variant (i.e. soft sweeps) when both s and time of selection onset are unknown. (B) The log-likelihood surface for joint inference of s and onset of selection, averaged over 100 simulations, taking the true tree as observed. (C) ROC curves illustrating performance of tests between selection from a standing variant where onset of selection occurs 100 generations ago. We condition on a present day frequency of 50%.

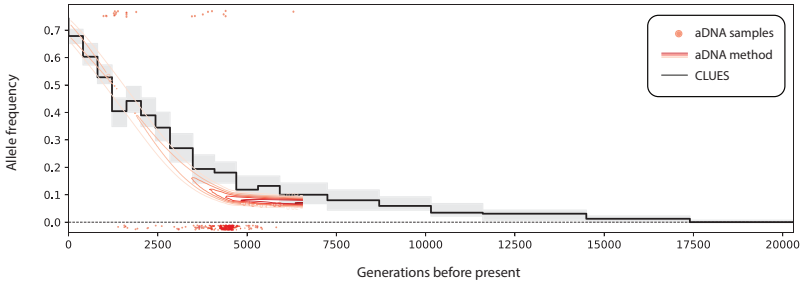


Figure 2.8: Comparison of inferred allele frequency trajectories for a sweep at rs4988235 (MCM6) in GBR under an ancient DNA (aDNA) based method vs. CLUES , which only uses contemporary modern data. Black curve is the posterior median allele frequency, whereas gray areas are a 95% posterior interval. The red surface is posterior of the frequency trajectory within Steppe ancestry conditioned on an ancient DNA time series, adapted from [136].

Analysis of a lactase persistence SNP

To assess performance of CLUES on empirical data, we applied our method to study selection acting on the SNP rs4988235 in the *MCM6 gene*, known to regulate the neighboring *LCT* gene and affect the lactase persistence trait. The derived allele (A) current segregates at approximately 72% in the 1000 Genomes Phase 3 reference panel (British in England and Scotland, henceforth GBR). We conducted sampling in ARGweaver assuming a model of European demography [129], using a 300kbp region centered around the focal SNP and polarizing alleles using the genomes of three ancient individuals (Altai Neandertal, Denisova, and Vindija Neandertal [133, 134, 135]). We sampled $M = 200$ ARGs, extracted local trees using tools in the ARGweaver package, and conducted importance sampling to estimate likelihood surfaces and trajectories using CLUES .

We found very strong evidence for selection on rs4988235 ($s = 0.0161$, $\log \mathrm{LR} = 131.82$). The trajectory as well as the value of the selection coefficient inferred by CLUES are consistent with previous estimates of the trajectory and $s = 0.018$ due to Mathieson and Mathieson (2018), illustrated in Fig. 2.8 [136]. Their method incorporates genomic times series spanning thousands of generations using an HMM-based approach, where hidden states are population-wide allele frequencies, observed states are genotypes of sampled ancient individuals, and transition probabilities are governed by the selection coefficient. Our approach, by contrast, does not utilize any ancient/timecourse data except for the 3 aforementioned ancient individuals, which we use to simply polarize the derived and ancestral states of each allele.

Analysis of pigmentation alleles

Using the same GBR panel from 1000 Genomes Phase 3, we analyzed a set of SNPs associated with pigmentation-related traits, some of which were previously identified as likely targets of recent selection [80]. We conducted sampling in ARGweaver assuming a model of European demography, using a 300kbp region centered around the focal SNP and sampling $M = 200$ approximately iid ARGs. We ran CLUES and estimated likelihood surfaces and allele frequency trajectories for these SNPs (Fig. 2.9). We found significant concordance between the SDS values and our likelihood ratio statistics paired for each SNP ($p = 1.7\times10^{-3}$, Spearman one-sided) [80]. We also illustrated the geographical distribution of these SNPs among diverse populations (Fig. A.8) using GGV [137].

We found several signals of very strong selection acting on rs619865 (*ASIP*, $s \approx 0.10$, Fig. 2.9I), rs12821256 (*KITLG*, $s \approx 0.016$, Fig. 2.9H), and rs1393350 (*TYR*, $s \approx 0.011$, Fig. 2.9J); these SNPs are significantly associated with freckling, blonde hair color, and freckling and blue/green eye color, respectively [138, 139, 140]. Interestingly, these SNPs all demonstrated a signal of selection mostly concentrated in the last ~5 kya. The geographical distribution of the frequency of these SNPs shows that the derived version of these variants are mostly concentrated in European populations, with minimal sharing with populations located in Africa and Asia (Fig. A.8 I,H,J). For example, *TYR* and *KITLG* segregate at a frequency ~20% in several European populations and have a frequency close to 0% in African and East Asian populations (Fig. A.8 J). These three SNPs are the only ones in this set of SNPs which have a frequency of nearly 0% across the African populations surveyed, with the exception of OCA2/HERC2 (Fig. A.8 A,H,I,J), consistent with our evidence for recent selection at these loci. The frequencies of these variants in GBR ranges from ~10-20%; by contrast, the only other variant in this set with comparable frequency in GBR (13%), rs35264875 (*TPCN2*), we find inconclusive evidence of selection (Fig. 2.9F), consistent with its comparably even geographical distribution relative to the aforementioned SNPs at *ASIP, KITLG*, and *TYR* (Fig. A.8 F).

At rs12896399 (*SLC24A4*, Fig. 2.9B), a SNP identified to be significantly associated with hair color [139], we found strong evidence for moderate selection ($s \approx 0.005$). This result is consistent with a previous analysis that suggested positive selection acted on this allele in Out-of-Africa (OoA) populations, based on its high allele differentiation relative to a YRI panel, and low haplotype diversity within CEU individuals [140]. Our results, paired with the apparent low levels of differentiation between European and Asian populations relative to differentiation between OoA populations and African populations at this locus (Fig. A.8 B) are consistent with our estimate that selection acted on *SLC24A4* as early as ~30 kya, during the OoA bottleneck as inferred by [141, 129].

Notably, we find moderate evidence for selection on rs12913832 (*OCA2/HERC2*, Fig. 2.9 A, Fig. A.9), a SNP previously shown to be causal for blue-brown eye color [142] and significantly associated with hair color [139]. This gene exhibits abberantly high differentiation across populations [143], consistent with a model of local adaptation of eye color. Compared to previous estimates based on ancient DNA samples [144], we estimate substantially

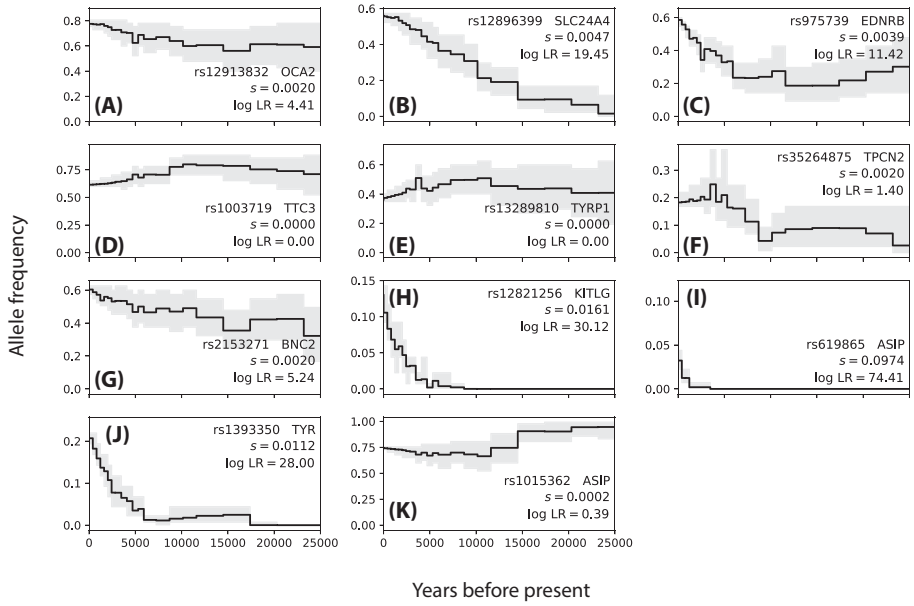


Figure 2.9: Allele frequencies trajectories inferred for 11 pigmentation-associated SNPs in GBR.

weaker selection acting on this gene ($s \approx 0.002$ vs. $s \approx 0.04$), and we find no evidence to support a recent increase in selection acting on this SNP (i.e., our method found a hard sweep to have higher likelihood than a SSV). Our estimate of moderate selection and lack of a recent change in the selection coefficient imply that selection on OCA2/HERC2 began at least ~50 kya, roughly the time of the start of the OoA bottleneck estimated by [141, 129]. Our analysis suggests that selection on *OCA2/HERC2* may have begun much earlier than previously suggested [144].

One surprising result is that we found no signal of selection acting at rs13289810 (*TYRP1*, $s \approx 0$, Fig. 2.9E). In Europeans, *TYRP1* is associated with hair and eye pigmentation [145, 146, 147, 148]. Some analyses of European populations have indicated evidence for positive selection on *TYRP1* [140, 146, 148]. Our results temper these claims, and appear consistent with the fairly even geographical distribution of rs13289810 frequency across European, African, and Middle Eastern populations (Fig. A.8 E).

2.4 Discussion

We have developed an approach to use modern population genomic data to approximate the full likelihood of selection acting on a locus. We use this approach to test for and

estimate the strength and timing of selection, as well as estimate the full allele frequency trajectory. The method is effective across a span of selection coefficients ($s = 0 - 0.01$), derived allele frequencies ($f = 25\% - 75\%$), and under multiple demographic models.

Our method draws on previously published methods to estimate the ancestral recombination graph (ARG). We chose to use ARGweaver because it is the only currently available method for sampling the posterior of the ARG; as shown in our derivation of the importance sampling estimates, we rely on sampling from the posterior in order to make rigorous guarantees regarding convergence and consistency of our estimators. Intuitively, it is important to model the uncertainty in the local tree in order to marginalize out this latent variable. We showed that estimates of the selection coefficient and the trajectory are generally accurate, barring scenarios where importance sampling is inefficient, or ARGweaver produces a bias in the inferred trees. In light of these biases, under certain conditions—primarily when the derived allele is at low frequencies ($\leq 25\%$)—importance sampling using ARGweaver trees has limited power to detect selection.

Another important limitation of ARGweaver is its computational cost; in order to study selection on short timescales, large sample sizes are necessary, often on the order of thousands of individuals [80]. The runtime of ARGweaver grows dramatically with increasing sample size; not only does the cost of the individual sampling steps increase with sample size, but also so does the size of the state space, necessitating more samples be taken in order to achieve convergence to the stationary distribution.

However, we see potential to make use of recent advances in inference of local trees in order to further advance approximate full-likelihood methods to infer selection (see e.g., [149, 150, 151, 152]; it is worth noting that some of these methods, such as [152], do not infer the ARG in a strict sense, but rather the sequence of local trees along a recombining locus). A major benefit of these methods is that they are far more scalable than ARGweaver, and hence offer more potential to study selection on short, punctuated timescales. However, they also possess several limitations: Firstly, several of these methods only infer topologies, rather than branch lengths [150, 151]. While it is possible to infer branch lengths condition on topology estimates, it is unclear how accurate these estimates would be. By contrast, methods that infer branch lengths along with topology entail a slight tradeoff in their scalability [149, 152]. Another limitation of these methods it that they only yield a point estimate of the local tree, rather than estimating uncertainty in the tree. Nonetheless, it may be feasible to quantify uncertainty in the local tree using a jackknife approach where the local tree is inferred over random subsets of the individuals.

It may also possible to make use of recent advances in inferring pairwise coalescence times (e.g., [153]) to build an approximation to the full likelihood. Recently, Albers & McVean proposed a composite likelihood method to estimate allele age by "sandwiching" the age using identity-by-descent tracts at the site of interest [154]. However, their method does not extend to inferring how the allele frequency changed over time, and does not explicitly model selection.

Currently our method assumes correct knowledge of the demographic history. The effects of latent or mis-specified population structure on inference of selection are well

known (e.g., [43]), but in future work one might try to determine the exact effects of mis-specification of effective population size on both inferring the local tree, and inferring selection conditional on the local tree. One approach to dealing with this is to extend the importance sampling approach we use to correct for selection to additionally correct for demography, when ARG sampling is performed under a mis-specified demographic model.

Furthermore, many aspects of our model of selection (e.g. coalescence, allele frequency transitions) assume a panmictic population. To extend our model to more complex demographic models would entail drastically increased computational cost (e.g., marginalizing allele frequencies corresponding to each population, rather than the allele frequency in a single population). Using a deterministic approximation of the allele frequency trajectory would circumvent this issue, but it would also raise new issues, such as how to model allele frequencies when $s = 0$.

Despite its limitations, the method presented here provides the first close approximation to a full likelihood function for the selection coefficient under simple models. As demonstrated by our simulations, full likelihood methods have the potential to greatly improve power to detect selection and estimate the strength of selection under a variety of conditions. It also provides a rigorous and accurate method for estimating allele frequency trajectories, and is the first to achieve so using modern data. As methods for inferring ARGs improve in the future, so too will the derived methods for detecting and quantifying selection and inferring allele frequency changes.

Chapter 3

Quantifying & disentangling selection on complex traits using whole-genome genealogies

This is work co-authored by Leo Speidel, Noah Zaitlen, and Rasmus Nielsen. It is in press at American Journal of Human Genetics, *and currently posted on* bioRxiv [155].

Abstract

We present a full-likelihood method to estimate and quantify polygenic adaptation from contemporary DNA sequence data. The method combines population genetic DNA sequence data and GWAS summary statistics from up to thousands of nucleotide sites in a joint likelihood function to estimate the strength of transient directional selection acting on a polygenic trait. Through population genetic simulations of polygenic trait architectures and GWAS, we show that the method substantially improves power over current methods. We examine the robustness of the method under uncorrected GWAS stratification, uncertainty and ascertainment bias in the GWAS estimates of SNP effects, uncertainty in the identification of causal SNPs, allelic heterogeneity, negative selection, and low GWAS sample size. The method can quantify selection acting on correlated traits, fully controlling for pleiotropy even among traits with strong genetic correlation ($|r_g|$ = 80%; c.f. schizophrenia and bipolar disorder) while retaining high power to attribute selection to the causal trait. We apply the method to study 56 human polygenic traits for signs of recent adaptation. We find signals of directional selection on pigmentation (tanning, sunburn, hair, P=5.5e-15, 1.1e-11, 2.2e-6, respectively), life history traits (age at first birth, EduYears, P=2.5e-4, 2.6e-4, respectively), glycated hemoglobin (HbA1c, P=1.2e-3), bone mineral density (P=1.1e-3), and neuroticism (P=5.5e-3). We also conduct joint testing of 137 pairs of genetically correlated traits. We find evidence of widespread correlated response acting on these traits (2.6-fold enrichment over the null expectation, P=1.5e-7). We find that for

several traits previously reported as adaptive, such as educational attainment and hair color, a significant proportion of the signal of selection on these traits can be attributed to correlated response, vs direct selection (P=2.9e-6, 1.7e-4, respectively). Lastly, our joint test uncovers antagonistic selection that has acted to increase type 2 diabetes (T2D) risk and decrease HbA1c (P=1.5e-5).

3.1 Introduction

Genome-wide association studies (GWAS) have shown that many human complex traits, spanning anthropometric, behavioral, metabolic, and many other domains, are highly polygenic [156, 157, 158]. GWAS findings have strongly indicated the action of purifying and/or stabilizing selection acting pervasively on complex traits [159, 160, 161]. Some work has also utilized biobank data to measure the fitness effects of phenotypes using direct measurements of reproductive success [162]. However, there are few, if any, validated genomic signals of directional polygenic adaptation in humans.

Several factors have contributed to this uncertainty. Chief among them, polygenicity can allow adaptation to occur rapidly with extremely subtle changes in allele frequencies [163]. Classic population genetics-based methods to detect adaptation using nucleotide data have historically been designed to detect selective sweeps with strong selection coefficients, unlikely to occur under polygenic architecture [1]. Polygenic adaptation, after a shift in the fitness optimum, can occur rapidly while causal variants only undergo subtle changes in allele frequency [40]. After a transient period during which the mean of the trait changes directionally, a new optimum is reached and the effect of selection will then largely be to reduce the variance around the mean [164]. However, identifying the genomic footprints of the transient period of directional selection is of substantial interest because it provides evidence of adaptation.

To this end, the advent of GWAS has ushered in a series of methods which take advantage of the availability of allele effect estimates by aggregating subtle signals of selection across association-tested loci. For example, some methods (e.g., the Q_X test) compare differences in population-specific polygenic scores – an aggregate of allele frequencies and allele effect estimates – across populations, and tests whether they deviate from a null model of genetic drift [165]. Other methods have generalized this test, e.g. to identify and map polygenic adaptations to branches of an admixture graph [166]. Whereas the aforementioned methods exploit between-population differentiation to detect polygenic adaptation, another class of methods is based on within-population variation. For example, selection scans based on singleton density score (SDS) have demonstrated utility in detecting polygenic adaptation via the correlation of SNPs' effect estimates and their SDSs [167]. Another test looks for dependence of derived allele frequencies (DAF) on SNP effect estimates [168].

Several powerful tests for selection were developed to take advantage of recent advances in ancestral recombination graph (ARG [169]) and whole-genome genealogy in-

ference. Such methods enjoy better power in detecting selection as the ARG, if observed directly, fully summarizes the effects of selection on linked nucleotide data. We note that several methods, notably ARGweaver [170] infer the strictly-defined ARG; by contrast, methods such as Relate [171] infer a series of trees summarizing ancestral histories spanning chunks of the genome. For example, the T_X test estimates changes in the population mean polygenic score over time by using the coalescent tree at a polymorphic site as a proxy for its allele frequency trajectory; the sum of these trajectories weighted by corresponding allelic effect estimates forms an estimate of the polygenic score's trajectory [172]. Speidel, et al. (2019) also designed non-parametric test for selection based on coalescence rates of derived- and ancestral-allele-carrying lineages calculated empirically from the coalescent tree inferred by Relate.19 However, these methods ultimately treat the coalescent tree as a fixed, observed variable, where it is actually hidden and highly uncertain. Furthermore, most methods infer the tree under a neutral model, and thus provide biased estimates of the genealogy under selection.

To address these issues, we recently developed a full-likelihood method, CLUES, to test for selection and estimate allele frequency trajectories [116]. The method works by stochastically integrating over both the latent ARG using Markov Chain Monte Carlo, and the latent allele frequency trajectory using a dynamic programming algorithm, and then using importance sampling to estimate the likelihood function of a focal SNP's selection coefficient, correcting for biases in the ARG due to sampling under a neutral model.

Beyond the issue of statistical power, tests for polygenic adaptation can in general be biased when they rely on GWAS containing uncorrected stratification. For example, several groups found strong signals of height adaptation in Europe [165, 166, 80, 173, 174, 175]; later, it was shown that summary statistics from the underlying meta-analysis (GIANT, a.k.a. Genetic Investigation of ANthropometric Traits) were systematically biased due to uncorrected stratification, and subsequent tests for selection on height failed to be reproduced using properly corrected summary statistics 20,25,26. However, beyond this case, the extent to which other signals of polygenic selection may be inflated by uncorrected stratification is an open question. Here, we investigate the robustness of the new likelihood method to uncorrected stratification and compare it to another state-of-the-art method (tSDS), showing that our likelihood method is less prone to bias but has substantially improved power.

Another issue faced by current methods to detect polygenic adaptation is confounding due to pleiotropy. For example, direct selection on one trait may cause a false signal of selection on another, genetically-correlated trait. While a variant of the Q_X test has been proposed for the purpose of controls for pleiotropy, this method relies of signals of between-population differentiation to test for selection, and is not directly applicable to test multiple traits jointly [175].

Here, we present a full-likelihood method (Polygenic Adaptation Likelihood Method, PALM) that uses population DNA sequence data and GWAS summary statistics to estimate direct selection acting on a polygenic trait. We demonstrate robustness by exploring potential sources of bias, including uncorrected GWAS stratification. We also introduce

a variant on our method which controls for pleiotropy by testing ≥2 traits for selection jointly. We show our method not only fully controls for this bias, but retains high power to distinguish direct selection from correlated response even in traits with strong genetic correlation (up to 80%), and has unique power to detect cases of antagonistic selection on genetically correlated traits. We explore the behavior of the test when traits with causal fitness effects are excluded to illustrate limitations and proper interpretation of these selection and correlated response estimates.

3.2 Model

Linking SNP effects to selection coefficients

Let β be the effect of a SNP on a trait. We model the selection coefficient acting on this SNP using the Lande approximation $s \approx \beta\omega$, where ω is the selection gradient, the derivative of fitness with respect to trait value. If β is measured in phenotypic standard deviations, then ω is the so-called selection intensity. Chevin and Hospital (2008) showed that for a neutral 'tag' SNP with frequency $u = 1 - v$ and genotypic correlation r to a SNP with selection coefficient s, and allele frequencies p and $q = 1 - p$, to a first approximation the linked neutral SNP effectively undergoes selection with $s_{tag} \approx rs\sqrt{pq/uv}$ [176]. Applying this principle to the multivariate Lande approximation, we find that $s_{tag} \approx \beta_{tag}\omega$, where $\beta_{tag} = r\sqrt{pq/uv}$ is the marginal effect of the tag SNP, assuming no linkage disequilibrium between the tag SNP and any other causal SNP other.

Inferring the seleciton gradient using a full-likelihood model

Our likelihood model builds heavily on Stern, et al. (2019), which developed importance sampling approaches for estimating the likelihood function of the selection coefficient acting on a SNP, $L^{SNP}(s)$ [116]. Let $\beta_{(i)}$ be the effect of SNP i on the trait. Based on these SNP-level selection likelihoods, we model the likelihood function for the selection differential acting on a trait as the product of the SNP likelihoods evaluated at selection coefficients under the Lande approximation:

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}(\beta_{(i)}\omega) \tag{3.1}$$

We provide details for calculating this likelihood function using an importance sampling approach in Appendix B. Given this likelihood function, we estimate ω using its maximum-likelihood estimate. This model is used by our so-called marginal test PALM.

Fitness effects of multiple traits

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

$$s \approx \beta^T \omega \tag{3.2}$$

where ω now is a vector of selection gradients for each of the d traits. The results of Chevin and Hospital (2008) apply directly given this approximation for the selection coefficient, and we now express the likelihood of the selection gradient as

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}(\beta_{(i)}^T \omega) \tag{3.3}$$

We can solve for the maximum-likelihood estimate of ω jointly using standard optimization. This model is used by our joint test J-PALM.

Simulations

Pleiotropic polygenic trait architecture

We sample effect sizes jointly for $d = 23$ polygenic traits with previously estimated SNP heritability and genetic correlations [177, 178]. We consider different values of polygenicity (M, the number of causal SNPs) and degrees of pleiotropy (ϱ, the probability that a causal SNP is pleiotropic). Let G be the additive genetic covariance matrix (diagonal entries are the SNP heritabilities $g_{ii} = h_i^2$ for each trait i). Then the genetic correlation of traits i, j is $r_{g,ij} = g_{ij}/\sqrt{h_i^2 h_j^2}$. Under our simulation model, we assume that if a SNP is pleiotropic, then $\beta \sim MVN\left(0, G^*/\left(M\nu\right)\right)$, where ${g^*}_{ii} = g_{ii} \cdot (1 - (1-\varrho)/d)/\varrho$,${g^*}_{i\neq j} = g_{i\neq j}/\varrho$. If a SNP is non-pleiotropic and is causal for trait j, then $\beta_j \sim N\left(0, h_j^2/\left(M\nu\right)\right)$ where $h_j^2 = g_{jj}$, and $\beta_{\neq j} = 0$. We assume that if a SNP is non-pleiotropic, it is causal for a particular trait j with uniform probability $1/d$. Under this model, we can see that averaging over pleiotropic and non-pleiotropic loci, we recover the overall genetic covariance G:

$$\sigma_{\beta_j}^2 = (1-\varrho)/d \cdot h_j^2 + \varrho \cdot (1 - (1-\varrho)/d)/\varrho \cdot h_j^2 = h_j^2 = g_{jj}, \tag{3.4}$$

$$\sigma_{\beta_i,\beta_j} = 0 + \varrho \cdot 1/\varrho \cdot g_{i\neq j} = g_{i\neq j} \tag{3.5}$$

Note that here β is standardized by the phenotypic variance, but not the genotypic variance. Thus we normalize the variance by a factor of $\nu = 2 \cdot E\left[pq\right]$, assuming some

stationary distribution for the allele frequency $p = 1 - q$. Assuming the neutral stationary distribution $f(p) \propto 1/p$, which yields $\nu = 4\log N_e$, where N_e is the diploid effective population size. This choice of ν ensures $E\left[\sum_{k=1}^{M} 2\beta_k^2 p_k q_k\right] = h^2$ under the nominal allele frequency spectrum. The equation holds because we assume independence of effects and allele frequencies; we also performed simulations where β and p are allowed to depend strongly on each other due to purifying selection.

Simulation of confounding due to population structure and uncorrected GWAS stratification

Previous estimates of selection to increase height in Europe have been biased by a combination of uncorrected stratification and GWAS and systematic differences in the coalescent rate at SNPs that depended on their allele frequency difference in 1000 Genomes (1KG) British (GBR) vs. Southern Italy (TSI) populations [179, 180]. We developed a simulation model based on empirical data from the 1KG data in order to assess the robustness of our method compared to tSDS-based tests for polygenic selection [167]. We model uncorrected stratification in summary statistics for a simulated polygenic trait architecture by drawing random SNP effects

$$\beta \sim N\left(0, h^2/(M\nu) \cdot I\right) \tag{3.6}$$

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

$$\phi = X\beta + S + \epsilon \tag{3.7}$$

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

$$E\left[\hat{\beta} \mid X\right] = \beta + X^T S/(2Npq) \tag{3.8}$$

$$= \beta + 2\sigma_s\left(\sqrt{N_1 N_2} f_1 - \sqrt{N_1 N_2} \cdot (N/N_2 \cdot p - N_1/N_2 \cdot f_1)\right)/(2Npq) \tag{3.9}$$

$$= \beta + \sqrt{N_1/N_2}\sigma_s (f_1 - p)/(pq) \tag{3.10}$$

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

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

$$\beta \sim MVN(0, h^2/(Mv) \cdot I) \tag{3.11}$$

$$\hat{\beta} \mid \beta \sim N(\beta + \sqrt{N_1/N_2}\sigma_s (f_1 - p)/(pq), \sigma_e^2/N \cdot I) \tag{3.12}$$

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

Population genetic model of selection and ascertainment bias via GWAS

Given β, we simulate selection following the multivariate Lande approximation (see Model). Because we simulate polygenic architectures of $M \geq 100$ without assuming no linkage betweened causal loci, outr assumption of infinitesimal genetic architecture is appropriate. (We also explore the performance of our model when we allow LD between causal SNPs; see Fig. B.4). We then simulate the trajectory of the allele forward in time using a normal approximation to the Wright-Fisher model with selection, i.e.

$$p_{t+1} \sim N\left(p_t + s p_t (1 - p_t), p_t (1 - p_t)/4N_e\right), \tag{3.13}$$

where s is calculated using the multivariate Lande approximation. For most of our simulations, we simulate forward for 50 generations (i.e., we assume selection began 50 generations before the present), unless otherwise stated. Let p be the present-day allele frequency. We simulate the ascertainment of this SNP in a GWAS by simulating the SNP Z-scores $Z \sim MVN\left(\sqrt{2Npq}\beta, E\right)$, where $E_{ii} = 1, E_{i \neq j} = \rho_e$, where ρ_e is a term that allows for cross-trait correlations in environmental noise. (Note that here Z is the usual Z-score of $\hat{\beta}$, not to be confused with the selection Z-score we introduce later.) Unless stated otherwise, we set $N = 10^5, \rho_e = 0.1$ in all simulations. We use a p-value threshold of 5×10^{-8} to ascertain a SNP; this must be surpassed by at least one trait. If a SNP is ascertained, we simulate its trajectory backwards in time using the normal approximation to the neutral Wright-Fisher diffusion conditional on loss, $p_{t-1} \sim N\left(p_t(1 - 1/4N_e), p_t (1 - p_t)/4N_e\right)$. We use the coalescent simulator `mssel` to simulate a sample of haplotypes conditional on this allele frequency trajectory [172]. We use $n = 400$ haplotypes and $\mu = r = 10^{-8}$/bp/gen and simulate regions of 1Mbp, centered on the causal SNP at the position 5×10^5.

To simulate ascertainment of non-causal SNPs in a GWAS, we take the trait with the top Z-score at the causal SNP and jointly simulate Z-scores for that trait for all linked SNPs within a 200kbp window centered on the causal SNP and surpassing a MAF threshold (MAF ≥ 0.01). We ascertain the SNP with the top Z-score (sometimes the causal SNP),

and then simulate the Z-scores for all traits, conditioned on the Z-score for the one aforementioned trait. We simulate this way rather than jointly simulating Z-scores for all traits at all SNPs because for two reasons; the top SNP will typically have the same top trait association as the causal, and jointly simulating all trait-by-SNP Z-scores increases computational time by >400 for the parameters we used.

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

Inference of local genealogies

Given a set of simulated haplotypes, we use the software package Relate19 to infer local genealogies along the sequence. Using positions of the SNPs ascertained through GWAS, we use the add-on module SampleBranchLengths to draw $m = 5,000$ MCMC samples of the branch lengths of the local tree at the ascertained sites. We then extract coalescence times from these MCMC samples (thinned down to m=500 approximately independent samples), and partition the coalescence times for each sample tree based on whether they occur between lineages subtending the derived/ancestral alleles. We note that Relate, unlike ARGweaver, does not sample over different ARG or tree topologies, and it samples branch lengths for two distinct local trees independently, conditional on the observed data.

Comparisons to tSDS in simulations

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

3.3 Results

Simulations

Overview of simulations

We conducted evolutionary simulations of polygenic adaptation acting on a wide range of multi-trait polygenic architectures. Our simulated traits are based on SNP heritability and genetic correlation estimates for 23 real human traits [177, 178]; unless otherwise stated, we simulate positive selection on/test for selection on a trait modeled after the heritability of schizophrenia (h^2=0.45), and in most of our pleiotropy analyses we used parameters based on schizophrenia and its genetic correlation with 3 other traits: bipolar disorder, major depression, and anorexia. In most of our analysis we refer to these traits as Trait I/II/III/IV (corresponding to models of schizophrenia/bipolar/depression/anorexia, respectively). As our method is based on aggregating population genetic signals of selection with GWAS summary statistics, we also simulated GWAS in samples of modern-day individuals ($N = 10^5$). Our simulated summary statistics incorporate all of the following sources of bias found in GWAS, unless stated otherwise: random noise in the effect estimates; Winner's Curse bias in the effect estimates (unless stated otherwise, we ascertain SNPs with associations $P < 5 \times 10^{-8}$ for at least 1 trait analyzed); uncertainty in the location of the causal SNP (we ascertain the top GWAS hit throughout the linked region); and environmentally correlated noise across traits (only relevant to simulations of pleiotropic architectures). Furthermore, we also simulate a number of scenarios which violate our model assumptions, to assess our method's robustness: these include uncorrected GWAS stratification; purifying/stabilizing selection; underpowered/uneven GWAS sample sizes; and allelic heterogeneity (i.e., multiple linked causal SNPs).

For each causal locus, we simulate haplotype data for a sample of $n = 400$ 1Mbp-long chromosomes (mutation and recombination rates $\mu = r = 10^{-8}$ and effective population size $N_e = 10^4$ unless stated otherwise), on which we applied Relate, a state-of-the-art method for tree inference, to infer the coalescent tree at SNPs ascertained in this GWAS [171]. However, we point out that our approach can be applied to any pre-existing method for estimating/sampling these trees (e.g. ARGweaver [170]). We then conduct importance sampling to estimate the likelihood function of the selection gradient – i.e., the effect of a unit increase in phenotypic values on fitness – for individual traits (i.e., estimated marginally), as well as sets of genetically correlated traits (i.e., estimated jointly). Our method, Polygenic Adaptation Likelihood Method (PALM), can be used to estimate ω for polygenic traits.

Improved power to detect selection and estimates of the selection gradient

We ran PALM to test for selection on our simulations of polygenic trait architectures, described above (and in more detail in Appendix B). We estimate the selection gradient

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

Table 3.1: Summary statistics for the accuracy and calibration of selection gradient estimates. Mean s.e. is the mean nominal standard error. Simulations are the same as used in Fig 3.1.

and standardize this quantity by its standard error, estimated through block-bootstrap, to conduct a Wald test on whether the selection gradient is non-zero.

First, we conducted simulations at different values of the selection gradient, ranging from neutral ($\omega = 0$) to strong($\omega = 0.1$, average change of mean phenotype of ~2 standard deviations), and compared the statistical power of PALM to that of tSDS (Fig 3.1). We simulate 5Mb haplotypes for a trait with polygenicity (i.e., number of causal SNPs) $M = 1,000$; we sample $n = 178$ haplotypes for PALM and $n = 6,390$ for tSDS, corresponding to the sample sizes we used in our application to 1000 Genomes British (GBR) individuals vs the sample used by Field, et al. (2016) from the UK10K. Here we ascertain only causal SNPs, but SNP effects are still estimated through an association test (unless otherwise stated, all other simulations incorporate uncertainty in the causal SNP). Both methods are well calibrated under the null ($\omega = 0$, Fig 3.1A). But we find that despite having a much smaller sample size, PALM has substantially improved power to detect selection at all levels (Fig 3.1A), especially at weaker values of the selection gradient, where tSDS has essentially no power ($\omega \leq 0.05$). PALM is also capable of estimating the selection gradient (Fig 3.1A, Tab 3.1). These estimates are well-calibrated, with empirical standard errors closely matching estimated standard errors, except when the selection gradient is exceptionally strong ($\omega = 0.1$) (Tab 3.1).

We also examined the calibration and power of the marginal test in simulations of a polygenic trait with varying polygenicity (Fig. 1D). Across a wide range of polygenicities, PALM is well-powered to detect selection (> 90% for $100 \leq M \leq 1000$), with slightly lower power for extremely polygenic architectures (~ 65% for $M = 10^4$) and the false positive rate (FPR) was well-calibrated in all circumstances (Fig. 3.1D). In comparisons to tSDS, we found substantially improved statistical power across this range of polygenicity values (Fig.3.1D). We also conducted similar tests for a short pulse of selection ($\omega = 0.05$ for 35 generations, or ~1000 years assuming 29 years/generation) under a model of British demography [152]; we found that overall power was comparable to that of constant population size simulations with $\omega = 0.025$, consistent with previous work showing that the product of selection strength and timespan largely determines statistical power

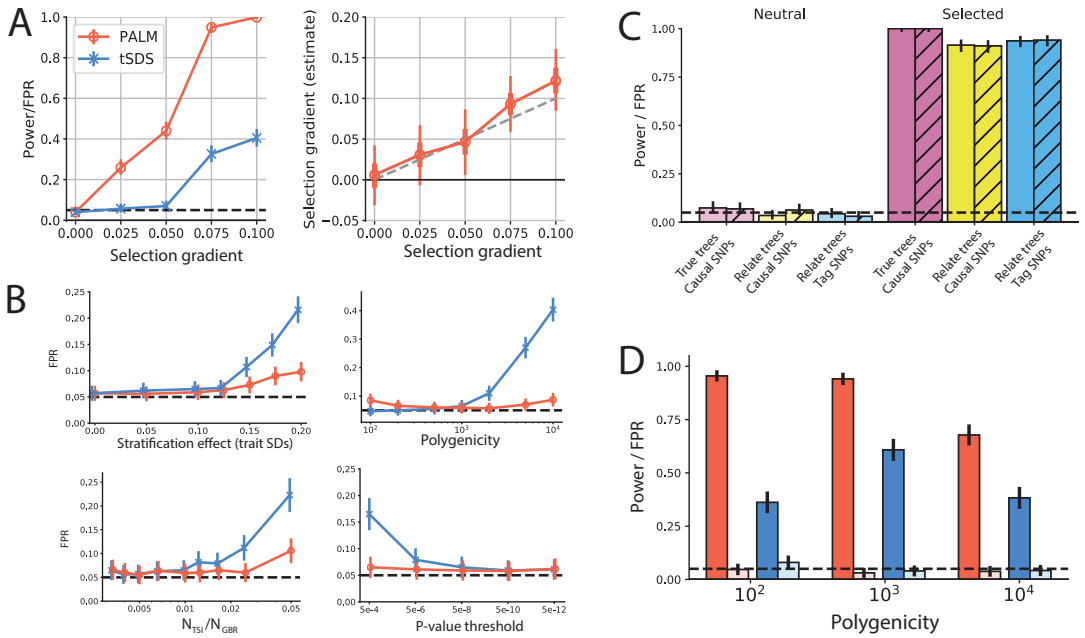


Figure 3.1: (A) Left: Power/false positive rate (FPR) of PALM and tSDS. Right: PALM selection gradient estimates. Error bars denote 25-75th percentiles (thick) and 5-95th percentiles (thin) of estimates; see Tab 3.1 for more details of moments and error. Markers and colors in (A) also apply to (B,D). (B) False positive rate of PALM and tSDS applied to neutral simulations with uncorrected population stratification, simulated using 1000 Genomes data. We used baseline values of $\sigma_S = 0.1, N_{TSI}/N_{GBR} = 1\%, M = 10^3, h^2 = 50\%$,using SNPs ascertained at $P < 5 \times 10^{-8}$. (C) Comparison of PALM using true vs Relate-inferred trees; causal vs GWAS-ascertained tag SNPs; and true marginal SNP effects (solid) vs GWAS-estimated SNP effects (hatched). (D) Varying polygenicity (M) of the polygenic trait. Baseline parameters for all simulations except (C) were our constant-size model with $M = 10^3$, with Scz under positive selection and testing Scz for selection. In (A,B) we use Relate-inferred trees and estimated SNP effects at the causal SNPs; in D we use Relate-inferred trees and estimated effects at tag SNPs. In all panels, we use a 5% nominal FPR (dashed horizontal line) and simulated 10^3 replicates. Error bars denote 95% Bonferroni-corrected confidence intervals.

(Fig. B.2).

Robustness to uncorrected GWAS stratification

We compared the power curve to the false positive rate (FPR) of both methods under a model of uncorrected GWAS stratification (Fig 1B). We simulated polygenic trait architectures and GWAS such that estimated SNP effects ($\hat{\beta}$) were both systematically biased and correlated with differences in the coalescence rate, stratified by DAF (e.g., SDS), matching the findings of [179, 180] that allele frequency differentiation between British (GBR) and Toscani in Italia (TSI) individuals was positively correlated with both $\hat{\beta}$ and SDS (Fig B.3).

To model this scenario, we ascertained a set of 40,320 SNPs with MAF $>$ 0.5% in the UKBB and SDS calculated by Field et al. (2016) using the UK10K cohort [167]. We then sampled coalescence times at these SNPs in 1KG Phase 3 British (GBR) individuals using Relate. For each SNP, we simulated GWAS summary statistics by assuming that the GWAS cohort is comprised of some ratio, N_{TSI}/N_{GBR}, of TSI to GBR individuals, where population identity determines an individual's stratified effect. This induces a correlation between SNP effects and the difference in allele frequency between TSI and GBR. Baseline parameter values were $\sigma_S = 0.1$, $N_{TSI}/N_{GBR} = 1\%$, $M = 1,000$, and $P = 5 \times 10^{-8}$. We varied the strength of the stratified effect (σ_S, in phenotypic standard deviations) and found that both methods are well-calibrated when σ_S is sufficiently small, but as σ_S grows past 0.1 the FPR of tSDS was inflated over 100% more than that of PALM (Fig 3.1B).

We stress that this disparity is most likely not caused by higher sensitivity of tSDS, as we simulated polygenic adaptation under similar parameters and found PALM was better-powered to detect selection, with up to 8x improvement in power for smaller values of the selection gradient (Fig. 1A). We also found that for highly polygenic traits (e.g. $M = 2 \times 10^3$), the tSDS test is overconfident ($>$ 10% at 5% nominal), while PALM remains well-calibrated (Fig. 3.1B). We observe the same pattern as we increase the size of the cohort subgroup receiving the stratified effect (N_{TSI}/N_{GBR}); at $N_{TSI}/N_{GBR} = 2.5\%$ the tSDS test is overconfident ($>$ 10% at 5% nominal), while PALM remains well-calibrated (Fig. 3.1B).

Lastly, we tested the sensitivity of these methods to the stringency of the P-value threshold used, and found that the tSDS test was increasingly overconfident as the threshold was relaxed, whereas, PALM was well-calibrated regardless of P-value threshold (Fig. 3.1B).These results suggest that PALM is more robust to uncorrected stratification than the tSDS test, while obtaining superior statistical power even at lower sample sizes. However, we emphasize that PALM, like any other available test, is not fully robust to the effects of uncontrolled population stratification. Sufficiently strong uncorrected population stratification can lead to false inferences of polygenic selection when there is none.

Robustness to ascertainment bias and uncertainty in GWAS estimates

Next, we considered the effects of different levels of uncertainty and ascertainment on performance of PALM (Fig. 3.1C). We considered the effects of conditioning on the true local tree vs using Relate-inferred trees combined with importance sampling, conditioning on the true marginal SNP effect vs estimating this effect with noise in a GWAS; and conditioning on the causal SNP vs taking the top tag SNP in a local GWAS on linked SNPs. PALM was well-calibrated both using true trees and importance sampling, with highest statistical power (100%) using true trees and a slight drop in power under importance sampling (90-92%) (Fig. 3.1C). Our test was well-calibrated despite bias (from Winner's Curse) and noise in the estimated SNP effects, with no discernible difference from using the true SNP effects (Fig 3.1C); however, for smaller sample sizes ($N << 10^5$) this may not be the case. Lastly, using the causal SNPs vs GWAS-ascertained tag SNPs did not diminish test power, and FPR remained well-calibrated (Fig 3.1C). We also explored the effects of GWAS sample size, which will affect the ascertainment process, and hence the degree of bias and uncertainty in ascertained SNP effect estimates (Tab. B2). We considered two different GWAS sizes; $N = 10^4$ and 10^5. We found that under lower sample size, the test was slightly inflated (e.g. empirical FPR of 3.1% (±1.4%) and 7.0% (±1.6%) at $N = 10^5, 10^4$ for Trait II respectively, where parentheses denote 95% CIs; Tab. B2). In terms of power, the test is still well-powered at lower sample sizes, but there is a noticeable drop (94.1% (±1.4%) and 69.0% (±3.0%) at at $N = 10^5, 10^4$ respectively; Tab. B2).

Robustness to model violations

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

As our model and baseline simulations assume a single causal SNP per linked locus, we conducted simulations of allelic heterogeneity (Supp. Fig 4) using forward simulations in SLiM 31. We simulated a trait architecture with $h^2 = 50\%$ and a mutational target of 100×1 Mbp linked loci, considering two cases: (1) 5% of incoming mutations are causal, and (2) 50% of incoming mutations are causal. In each of these scenarios we conducted simulations with neutral evolution and adaptation. We found that in each case, the test is well-calibrated under the null, and well-powered to detect selection (Fig. B.4).

Lastly, we explored the time specificity of PALM's test for selection. Testing under a nominal model of selection in the last 50 generations, we consider the rate at which

PALM's estimate of selection timing can be biased by older selection (Fig. B.5). We found that as selection recedes into the past, the FPR decays towards the nominal rate, with limited confounding when the pulse of selection occurred 200-250 generations ago.

Pleiotropy can cause bias in tests for polygenic adaptation

Traits with no fitness effect can undergo correlated response due to direct selection on pleiotropically related traits. Without accounting for pleiotropy, standard tests for polygenic adaptation cannot be interpreted as statements regarding direct selection. To illustrate how pleiotropy can affect tests for polygenic adaptation, we simulated pleiotropic trait architectures for 23 traits based on estimates of SNP heritability and genetic correlation for real human traits [177]. This builds largely off our aforementioned simulation approach, with the introduction of a parameter ϱ, the degree of pleiotropy, i.e. the probability that a causal SNP is pleiotropic. As a brief illustration of how pleiotropy causes bias in polygenic selection estimates, we used our pleiotropic traits simulations to estimate maximum-likelihood selection coefficients for SNPs ascertained for associations to two genetically correlated traits, Trait I and II, modeled after schizophrenia and bipolar disorder ($r_g \approx 80\%$; Fig. B.6). We simulate a pulse of selection to increase Trait I ($\omega = 0.05$, approximately +1 SD change in population mean over 50 generations, Tab. B1); Trait 2 has no causal effect on fitness. Selection coefficients were estimated by taking the maximum-likelihood estimate of s for each SNP independently, where the likelihood is estimated using our importance sampling approach. Here we show results for polygenicity $M = 1000$ and degree of pleiotropy $\varrho = 60\%$ (Fig. B.6).

Under the Lande approximation $s \approx \beta^T \omega$, we expect a non-constant linear relationship between $\hat{\beta}$ and $\hat{s}$ for traits under selection. But due to the strong correlation between these two traits, it is difficult to disentangle which of the traits has a causal effect on fitness (Fig. B.6A). We performed an ad-hoc test for a systematic relationship between $\hat{\beta}$ and $\hat{s}$ (Spearman test) to detect polygenic adaptation; while this test is well-powered to detect selection on Trait I, it is prone to spurious hits for selection on Trait II, which has no effect on fitness (Fig. B.6B). Thus, marginal tests for selection on traits can be significantly biased due to pleiotropy (in this case, genetic correlation).

Joint test for polygenic adaptation controls for pleiotropy

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

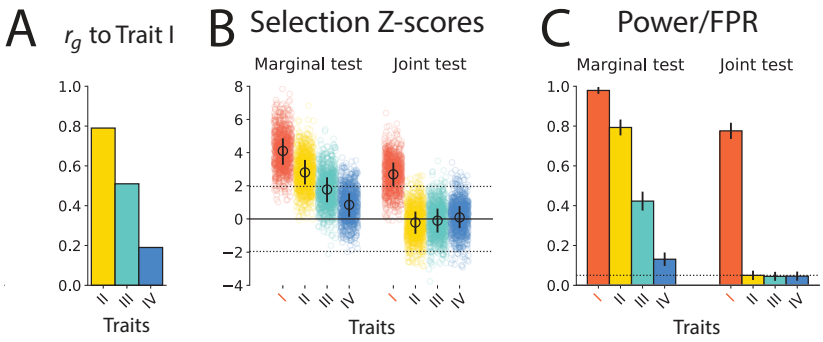


Figure 3.2: **Joint testing for polygenic adaptation controls for pleiotropy** We simulated a cluster of four traits (I-IV) modeled after (A) real human heritability and genetic correlation estimates for schizophrenia (I), bipolar disorder (II), major depression (III), and anorexia (IV), with selection to increase Trait I in the last 50 generations. (B,C) We ran marginal and joint tests for selection on these four traits. While marginal selection tests were well-powered, they were strongly biased by even fairly low genetic correlations. (B,C) Conducting a joint test fully controls for genetic correlations while retaining high power to detect and isolate selection on Trait I. Simulations (1,000 replicates) were done under our constant effective population size model with $\varrho = 60\%$, $M = 1,000$, with Trait I under positive selection.

We applied both our marginal test PALM and our joint test J-PALM to our cluster of four simulated traits, Traits I-IV, modeled after SNP heritabilities and genetic correlations for four psychiatric traits: schizophrenia, bipolar disorder, major depression and anorexia (Fig 3.2A). All traits have significantly positive genetic correlation to one another; here we highlight their genetic correlations to the selected trait, Trait I (Fig 3.2A; genetic correlations and SNP heritabilities directly from [177]). We assume a pulse of recent selection for increased Trait I prevalence, with all other traits selectively neutral. We tested traits marginally and jointly (i.e., all four simultaneously) for selection (Fig 3.2B,C). We found that marginal estimates are biased and cause inflation of the false positive rate (FPR) when testing for selection (Fig 3.2B,C). This bias largely follows the genetic correlation of the estimand trait to the selected trait (Fig 3.2A,B). Here we show results for polygenicity M=1000 and degree of pleiotropy $\varrho = 100\%$ (Fig 3.2), but the results are similar for differing degrees of pleiotropy (holding r_g constant), such as $\varrho = 60\%$ (Fig B.7). This highlights that genetic correlation, regardless of the degree of pleiotropy, is the main cause of bias in marginal estimates of the selection gradient.

Furthermore, our results show that if any trait in a genetically correlated cluster is under selection, marginal estimates of the selection gradient for the other traits is typically highly inflated. For example, a genetic correlation as low as $r_g = 19\%$ is sufficient to inflate the

FPR for a neutral trait by nearly 150% (Fig 3.2A,C). Most traits studied in GWAS have large genetic correlations; Watanabe, et al. (2019) found an average $|r_g|$ = 16% across 155,403 human trait pairs, with 15.5% of trait pairs significant (average $|r_g|$ = 38%) [181]. The extent of strong genetic correlation suggests that if any single heritable trait has evolved under selection, it is likely to cause substantial ripple effects in terms of bias of selection estimates on other heritable traits. By contrast, estimates of selection obtained via our joint test, fully correct for these biases, if the relevant selected trait is included in the analysis (Fig 2B,C). We applied the joint test to the same set of simulations and find it can reliably detect and attribute selection to Trait I (Fig 3.2B,C). The joint test preserved ~80% power even with the leading genetic correlate, Trait II, having r_g=79.4% to Trait I, and produces well-calibrated FPR regardless of r_g (Fig 3.2C).

We explored performance of J-PALM under a wide array of simulation scenarios of different polygenic architectures and types of selection (Fig. 3.3), varying the degree of pleiotropy ϱ (Fig 3.2A), r_g to the selected trait (Fig 3.2B), polygenicity M (Fig 3.2C), and antagonistic selection (Fig 3.2D). Baseline values of parameters used were positive selection on Scz with other traits neutral, jointly testing Trait I and Trait III (r_g = 51%), ϱ = 60%, and $M = 1,000$. All of our pleiotropic simulations include an environmental noise correlation across traits of ρ_e = 10%. Across this range of pleiotropic and polygenic architectures, we established that the joint test is well calibrated when no traits are under selection (Fig. B.8). Across different degrees of pleiotropy ($40\% \leq \varrho \leq 100\%$), we found J-PALM was well-calibrated and had good power to detect and attribute selection to Trait I (Fig 3.3A).

Across a range of levels of polygenicity ($100 \leq M \leq 10,000$), PALM was well calibrated and had good power to detect and attribute selection to Trait I (> 75% for $M \leq 3,000$), although the power is somewhat attenuated for extremely polygenic architectures (~ 40% for $M = 10,000$) (Fig 3.3B). This pattern is also found in the marginal tests on the same data, and there is only a modest reduction in power when switching to the joint test (Fig 3.1C, Fig 3.3B). We note that the reduction in power is sensitive to the strength of genetic correlation; joint test of Trait I vs Trait II (r_g = 79%) had greater reduction in power from the marginal test than that of Trait I vs Trait III (Fig 3.1C, Fig 3.3B,C, Fig B.9). Our method fully corrects the biases suffered by marginal tests for polygenic adaptation, while retaining good power to detect adaptation even when genetic correlation is strong.

We also examined what happened when selection acted on different traits in the cluster, jointly testing each selected trait with Trait II (Fig 3C). The test is well-calibrated for all traits, but has less power to attribute selection to traits with a high genetic correlation to Trait II (e.g. Trait I, h^2 = 45%, r_g = 7%), or low heritability (e.g. Trait III, h^2 = 17%, r_g = 4%) (Fig 3.1E, Fig 3.3C). By contrast, traits with high heritability and/or low genetic correlation to Trait II (e.g. Trait IV, h^2 = 49%, r_g = 11%) have little loss in power in the joint test (Fig 3.1E, Fig 3.3C).

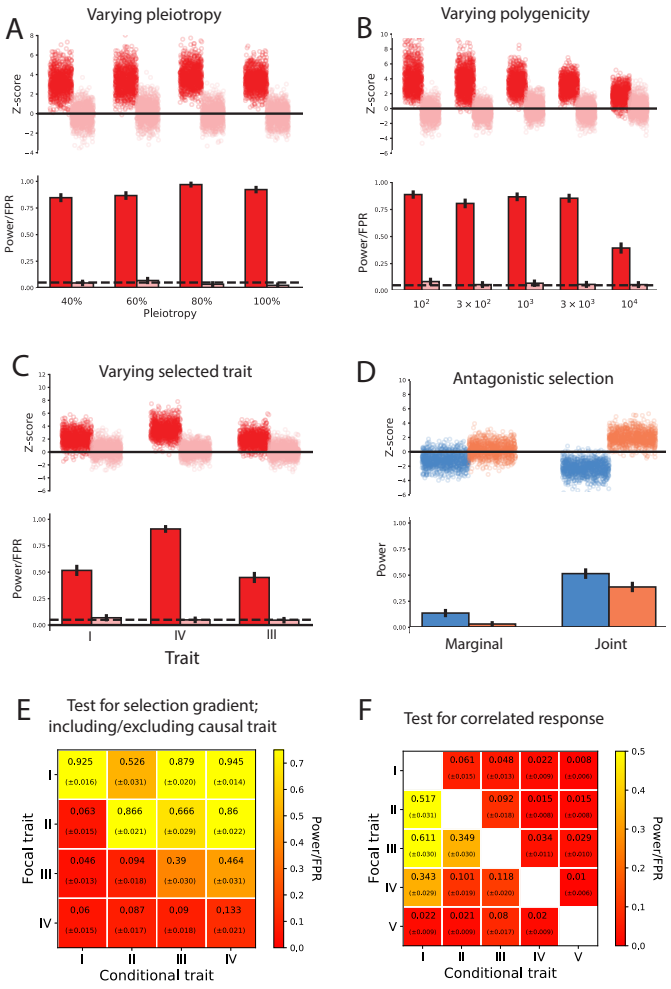


Figure 3.3: **Simulations of joint testing power and calibration** (A) Differing the degree of pleiotropy ϱ, (B) the trait truly under selection, (C) the polygenicity M of the traits, (D) antagonistic selection on two traits with positive genetic correlation, (E) pairwise tests for selection (Trait I under selection), (F) pairwise tests for correlated response (Trait I under selection). (A-D) Red/pink/blue bars indicate estimates of selection for traits under positive selection/neutrality/negative selection, (E-F) Heatmap is colored by positive rate (also text in boxes; standard errors in parentheses). Dashed horizontal lines indicate 5% nominal significance level and black lines indicate 95% Bonferroni-corrected confidence intervals. Baseline parameters for all simulations (1,000 replicates under each scenario) were our constant-size model with $\varrho = 60\%$, $M = 1,000$, with Trait I under positive selection. In panels (A,B) and (D) joint tests are performed on Trait I/Trait III and Trait I/Trait II, respectively. (E) Diagonal elements correspond to marginal test for selection.

Detecting antagonistic selection

We also considered the possibility of antagonistic selection (i.e., selection to both increase Trait I and decrease Trait II, Fig. 3.3D). We hypothesized that marginal tests would be underpowered to detect this mode of selection acting on traits with strong genetic correlation, and that joint testing might uncover this signal. Indeed, power to detect selection in this regime is quite low using marginal testing, with 3-13% power at a 5% threshold (Fig 3.3D). However, the joint testing boosts power significantly, with 40-51% power at a 5% threshold (Fig
reffig:Ch3Fig3D). We also tested the opposite scenario, where Trait I and Trait II are both under positive (complementary) selection; we found the joint test is well-powered to detect that multiple genetically correlated traits are under selection (Fig. B.10). Thus, J-PALM provides several gains in power over the marginal test, such as uncovering antagonistic selection that is 'cancelled out' by genetic correlation, or confirming multiple traits are under selection.

Interpretation and limitations of the joint test

We also considered how our joint test performs when the causal trait (i.e., a trait with a causal effect on fitness) is excluded from the model. We conducted pairwise joint tests on each pair of Traits I-IV in simulations with Trait I under selection and all other traits neutral (Fig. 3.3E). Rows correspond to the trait for which the selection test is performed (the focal trait), and columns correspond to the other trait included in the joint model (the conditional trait). We also considered other scenarios, such as all traits neutral, complementary selection, and antagonistic selection (Fig. B.11).

As we demonstrated previously, when the causal trait (Trait I) is included, the selection test is well-calibrated for neutral traits (Fig. 3.3E). However, we find that when Trait I is excluded, the selection test has high positive rates for traits that have no causal fitness effect, but are strongly genetically correlated with the causal trait (e.g. Trait II). In general, our results demonstrate that selection tends to be attributed to the trait with the strongest genetic correlation to the causal trait (e.g., Trait II); other traits with genetic correlation to the causal trait (e.g. Trait III) have some minor inflation of the positive rate, but selection is predominantly attributed to the closest proxy for the causal trait. These results highlight an important limitation of our model: Namely,the selection gradient estimates are not to be interpreted as causal fitness effects. As our simulated results show, this proposition is generally false when a trait with causal fitness effect and nonzero genetic correlation is excluded.

Testing for correlated response

Our method can also test for correlated response to selection, i.e., whether a trait has evolved (at least in part) due to selection on some other genetically correlated trait. We

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

We conducted tests of correlated response on each pair of traits I-V (we introduce Trait V, which has $r_g = 0\%$ to Trait I) (Fig. 3.3F). We found that the test for correlated response of Trait I is null, concordant with all other traits in the simulation being neutral (Fig. 3.3F). We also saw that Trait V, which has no genetic correlation to the directly selected trait, the test is null, concordant with the necessity of genetic correlation to drive correlated response (Fig. 3.3F). We saw that tests for correlated response generally grew in their power as r_g to Trait I increased. However, power is slightly lower for $r_g = 80\%$ than $r_g = 50\%$ (i.e., testing Trait II vs. Trait III for correlated response to Trait I) (Fig. 3.3F). This may indicate that for strongly genetically correlated traits, it is often ambiguous which one of the traits is evolving in correlated response. The test is also well-calibrated under neutral simulations (Fig. B.12A), and well-powered to detect more complex forms of correlated response such as antagonistic and complementary selection (Fig. B.12B,C). We also explored the performance of the correlated response test, along with the joint test for selection, in a K-way model with Traits I-IV tested jointly (Fig. B.13). Our results indicate that our test statistic R can be used to detect whether a trait has been under correlated response; however, it is incorrect to make strongly causal interpretations of the test (e.g., "Trait III is under correlated response to Trait II").

Effect of small of uneven GWAS sample size

We tested the effect of GWAS sample size on the joint test, considering not only lower sample size, but also uneven sample sizes (Tab. B2). Similar to the effect of lower sample size on the marginal test, we found that lower sample size for both traits reduced power and slightly inflated the FPR; e.g., testing for selection jointly on Trait I vs Trait II (simulating selection to increase Trait I), we found that at $N = 10^4$ for Trait I and Trait II, the FPR for Trait II reached 8.0% (±1.8%) (Tab. B2). However, this was not always the case; e.g., for $N_I = 10^5$, $N_{II} = 10^4$, the FPR for Trait II was calibrated properly (4.6% ±1.4%) (Tab. B2).

Power to assign selection to the causal trait was reduced when that trait's GWAS was

underpowered; e.g., 51.6% (±1.6%) to 45.7% (±1.6%) when N_I was dropped from 10^5 to 10^4 ($N_{II} = 10^5$) (Tab. B2). Interestingly, we found an even bigger drop in power associated with reduced sample size for the correlated trait (Trait II); when N_{II} was reduced from 10^5 to 10^4 ($N_I = 10^4$), power to detect selection on Trait I dropped from 45.7% (±1.6%) to 27.7% (±1.4%) (Tab. B2). These results indicate that as long as sample size is reasonably large, estimates are well-calibrated; furthermore, by increasing sample size of GWAS for one trait, the joint test is able to leverage that towards improving power to detect selection on other traits that have overlapping genetic architecture.

Empirical analysis of trait evolution in individuals of British ancestry

We analyzed 56 GWASs of metabolic, anthropometric, life history, behavioral, pigmentation- and immune response-related traits in humans (54 from the UKBB; see Tab. B3 for details) for signs of polygenic adaptation. We used GWAS summary statistics that were nominally corrected for population structure using either a linear mixed model [182] or fixed PCs (K=20 PCs) [183], and in some cases a family history-based approach [184] to boost power for under-powered UKBB traits, such as type 2 diabetes. All traits used had at least 25 genome-wide significant (GWS) loci($P < 5 \times 10^{-8}$) in independent LD blocks [185]. For all of our empirical analyses, we used coalescent trees sampled using Relate for a sample of British ancestry (GBR, $n = 89$) from the 1000 Genomes Project, assuming pre-established estimates of GBR demographic history [152, 186]. We specifically tested for selection in the last 2000 years (i.e., 68.95 generations, assuming a generation time of 29 years). The selection gradient (ω) was estimated using maximum-likelihood, with standard errors estimated by block-bootstrapping. We first tested traits marginally for polygenic adaptation (Fig. 3.4). We include SNPs by pruning for LD using independent LD blocks, choosing the SNP with the lowest p-value in each independent block, and excluding blocks that do not have a SNP exceeding this threshold [185].

Marginal tests for selection

We report our estimates of the selection gradient (Fig. 3.4) normalized by their standard errors, highlighting significant traits (FDR = 0.05) and other traits of interest, with results also presented in Tab. B4. In the marginal tests with PALM, we found strong signals of selection acting to decrease pigmentation (Fig. 3.4, Tab. B4). We reported traits with selection gradient p-value exceeding a multiple testing-corrected threshold (FDR = 0.05, Benjamini-Hochberg). Tanning showed the strongest signal of directional (in this case, negative) selection among all tested traits ($\omega = -0.357\,(\pm 0.046)\,, P = 5.5 \times 10^{-15}$; standard errors in parentheses). Sunburn $\left(\omega = +0.356\ (\pm 0.052)\,, P = 1.1 \times 10^{-11}\right)$ and hair color ($\omega = +0.128\,(\pm 0.027)\,, P = 2.2 \times 10^{-6}$) also showed significant positive selection. Several life history traits also showed significant selection; e.g. age at first birth ($\omega = +0.0546\,(\pm 0.0149)\,, P = 2.5 \times 10^{-4}$) and EduYears ($\omega = +0.389\,(\pm 0.0107)\,, P = 2.6 \times 10^{-4}$).

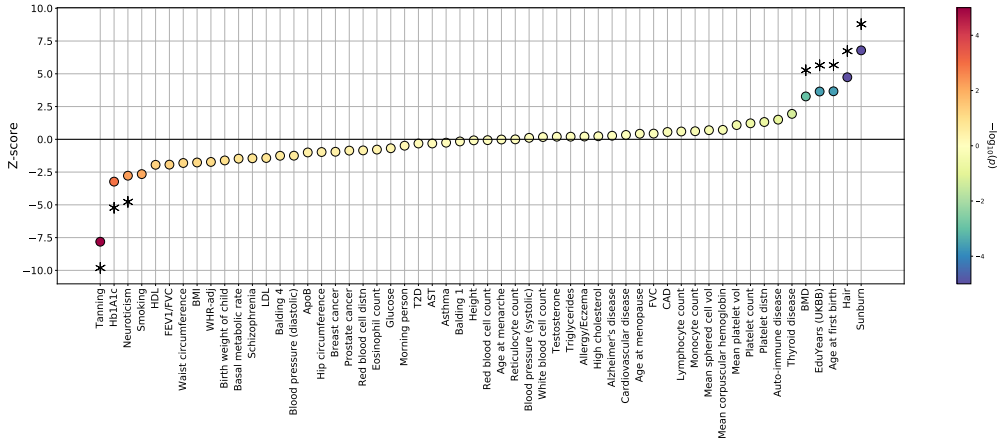


Figure 3.4: **Estimates of the selection gradient on 56 human traits** The selection gradient ($\hat{\omega}$) was estimated using 1000 Genomes Great British (GBR) individuals and summary statistics from various GWASs, with standard errors ($\widehat{se}_{\omega}$) estimated via block-bootstrap ($Z = \hat{\omega}/\widehat{se}_{\omega}$). Starred traits indicate significance at FDR = 0.05.

We also found significant selection acting on an anthropometric trait, bone mineral density heel-T Z-score (BMD, $\omega = +0.0728\,(\pm 0.0222)$,$P = 1.1 \times 10^{-3}$), and negative selection acting on glycated hemoglobin levels (HbA1c, $\omega = -0.0167\,(\pm 0.00518)$, $P = 1.2 \times 10^{-3}$) as well as neuroticism ($\omega = -0.0706\,(\pm 0.0254)$, $P = 5.5 \times 10^{-3}$).

Several traits of interest to have no or inconclusive evidence of directional selection. We found no evidence for any recent directional selection on height ($\omega = -0.00148 \times 10^{-3}$ (± 0.0190), $P = 0.938$). We also find inconclusive evidence for selection on body mass index (BMI, ($\omega = -0.0585\ (\pm 0.0331)$, $P = 0.077$), in contrast to previous findings that BMI has been under significant selection to decrease [168].

Joint tests for selection

We analyzed 137 trait pairs (Bonferroni $P_{r_g} < 0.005$ and $|r_g| > 0.2$)[181] using J-PALM to examine if marginal signals of selection were due to a correlated response to selection on another trait (Table 3.2, Tab. B5). To aid clarity, we introduce the notion of focal vs conditional traits in a joint test. For example, if we estimate the selection gradient of Trait 1 and Trait 2, (ω_1, ω_2), then ω_1 is the estimate for Trait 1 (the focal trait), jointly tested estimated with Trait 2 (the conditional trait); similarly ω_2 is the estimate for Trait 2 (the focal trait), jointly tested estimated with Trait 1 (the conditional trait). We establish significance of correlated response using a Wald test on the statistic R, the difference in the joint and marginal selection estimates for a focal trait, where the joint analysis is performed with some other conditional trait (see "Testing for correlated response" and

Traits		Marginal test		Joint test			
Focal	Conditional	Z	P_Z	Z	P_Z	R	P_R
Hair	Tanning	4.74	2.2e-06	1.91	0.056	-3.77	1.7e-04*
EduYears	Sunburn	3.65	2.7e-04	2.33	0.020	-4.68	2.9e-06*
HbA1c	T2D	-3.23	1.2e-03	-4.41	1.0e-05	-3.17	1.6e-03*
HbA1c	BP (Diastolic)	-3.23	1.2e-03	-1.95	0.051	2.36	0.019
T2D	HbA1c	-0.32	0.75	2.75	6.0e-03	4.34	1.5e-05*
T2D	BP (Diastolic)	-0.32	0.75	0.28	0.78	2.10	0.036

Table 3.2: **Selected trait pairs under correlated response in Great British ancestry** Selection on the focal trait is estimated jointly with the conditional trait. We report the Z-scores under both the marginal and joint tests, as well as the R statistic of the difference in joint vs marginal selection gradient estimates, and their P-values. T2D = Type 2 diabetes, HbA1c = glycated hemoglobin, BP = blood pressure. Asterisk (*) denotes significance at FDR = 0.05 ($n = 2{\times}137 = 274$ tests on 137 trait pairs with Bonferroni-significant $P_{r_g} < 0.005/\frac{56{\cdot}55}{2} and |r_g > 0.20|$).

Appendix B for more details). Selected results are presented in Table 3.2, and results for the full analysis of all 137 trait pairs are available in Tab. B5.

We found several significant signals (FDR = 0.05) of correlated response (Table 3.2, full results in Tab. B5). For example, although EduYears had strong evidence for selection in the marginal test ($P_{marginal} = 2.6 \times 10^{-4}$), we found after conditioning on sunburn ability ($r_g = 0.24, P = 2.3 \times 10^{-4}$)[181] a significant attenuation of this estimate ($P_{joint} = 0.020, P_R = 2.6 \times 10^{-6}$). These results suggest that a large part of the signal of selection on EduYears is likely due to indirect selection via correlated response, vs direct selection. However, we stress that these results do not provide evidence of direct selection on the conditional trait, here e.g. childhood sunburn occasions (sunburn) (see e.g. Fig. 3.3E).

We also find significant attenuation of selection signals for pigmentation traits in our joint analyses (Table 3.2). In our joint analysis of hair color and tanning ($r_g = -0.17, P = 3.6{\times}10^{-3}$) [181], we found that after conditioning on tanning, there is no residual evidence for direct selection on hair color ($P_{marginal} = 2.2 \times 10^{-6}$; $P_{joint} = 0.056$; $P_R = 1.7 \times 10^{-4}$). (The same caveat above regarding the interpretation of correlated response applies here to tanning ability).

We identified one case in which the joint analysis uncovers selection acting on a trait that did not show significant selection marginally; we found that type 2 diabetes (T2D), conditioning on HbA1c ($r_g = 0.69$)[187], shows significant selection to increase in prevalence ($P_{marginal} = 0.75$; $P_{joint} = 0.0060$; $P_R = 1.5 \times 10^{-5}$; see Table 3.2). Estimates of negative selection on HbA1c are also enhanced after accounting for T2D ($P_{marginal} = 1.2{\times}10^{-3}$; $P_{joint} = 1.0{\times}10^{-5}$; $P_R = 0.0016$; see Table 3.2). This 'cancelling-out' effect of opposing

selection on T2D and HbA1c, two traits with strong (but not perfect) positive genetic correlation, is the second-strongest signal of correlated response in our joint analyses. We confirmed that the separability of these two phenotypes is not due to phenotype mis-specification; T2D status was confirmed by doctor's diagnosis strictly after 30 years of age, in order to avoid the possibility that T1D individuals mistakenly self-report as T2D-diagnosed [184].

We also illustrate our estimates of selection coefficients for ascertained T2D/HbA1c SNPs, found independently of one another, and their fit to our inferred model of antagonistic selection on T2D/HbA1c (Fig. 3.5A). In general, T2D-increasing and/or HbA1c-decreasing SNPs are under positive selection, and vise versa; however, a subset of HbA1c-increasing SNPs show extremely strong signs of positive selection ($s > 0.03$); these SNPs tend to have visibly higher positive effects on T2D than other SNPs with comparable HbA1c effect. In a joint analysis of HbA1c and diastolic blood pressure (as a proxy for hypertension), our estimate of direct selection on HbA1c was significantly attenuated at a nominal level($P = 0.019$, Table 3.2), although it did not meet our FDR cutoff. We also did a joint analysis of T2D and diastolic blood pressure, finding a significant boost in the estimate of direct selection on T2D ($P = 0.036$, Table 3.2), although it did not meet our FDR cutoff.

Lastly, we tested our set of R statistics among the pairs of genetically correlated traits for enrichment in the tail over the null (Fig. 3.5B). At the nominal 5% FPR level, we found significant (2.6-fold) enrichment for correlated response acting on these traits ($P = 1.5 \times 10^{-7}$, one-sided binomial test), suggesting that many additional traits in this analysis have evolved under indirect selection due to correlated response.

3.4 Discussion

We have presented a method, PALM, for estimating the directional selection gradient acting on a polygenic trait. Our method works by estimating likelihood functions for the selection coefficients of a set of GWAS SNPs, and then aggregating these functions along with GWAS-estimated SNP effects to find the likelihood of the selection gradient. Through simulations, we showed that PALM offers improved power over current methods across a range of selection gradients($\omega = 0.025 - 0.10$) and polygenicities ($M = 10^2 - 10^4$), and is the first method to our knowledge that can estimate ω from nucleotide data. We conducted robustness checks and showed that PALM is robust to typical sources of uncertainty and bias in GWAS summary statistics (e.g. sampling variation, ascertainment bias/Winner's Curse) allelic heterogeneity, purifying selection, and underpowered GWAS.

We also introduced a method, J-PALM, to jointly estimate the selection gradient on multiple traits in order to control for pleiotropy. We showed that, across a wide range of polygenic architectures ($M = 10^2 - 10^4$, $\varrho = 40\% - 100\%$), J-PALM can reliably detect and assign selection to the causal trait when it is considered in the analysis, and can be used to uncover genetically correlated traits under antagonistic selection where the marginal

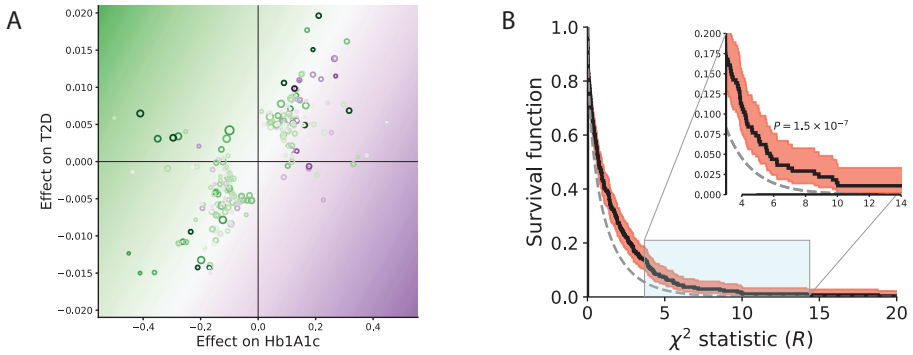


Figure 3.5: **Correlated response in real traits** (A) Expanded view of antagonistic selection on glycated hemoglobin (HbA1c) vs type 2 diabetes (T2D). We estimate individual SNP selection coefficients by taking the maximum-likelihood estimate $\hat{s}$ for each SNP. We plot this value against the joint SNP effect estimates for HbA1c and T2D. Colored lines represent isocontours of $s(\beta) = \beta_{HbA1c}\hat{\omega}_{HbA1c} + \beta_{T2D}\hat{\omega}_{T2D}$, the estimate of the Lande transformation from SNP effects to selection coefficients, where $\hat{\omega}$ is inferred jointly for the two traits (Tab. 3.2). Purple-green color gradient illustrates expected selection coefficients under $\hat{\omega}$ (background) vs. individual SNP selection coefficient estimates (rings). Ring diameter is proportional to SNP selection log-likelihood ratio. (B) Enrichment of correlated response in analysis of genetically-correlated traits. Enrichment in the tails of the distribution of our test statistic for correlated response R ($P = 1.5 \times 10^{-7}$, binomial test) which had 2.6-fold enrichment at the nominal 5% level. We assessed $n = 2 \times 137 = 274$ estimates of correlated response on 137 trait pairs with Bonferroni-significant $P_{r_g} < 0.005/\frac{56}{2}$ and $|r_g > 0.20|$. Red area indicates pointwise 95% CI of the survival curve.

approach (e.g. PALM) is underpowered. We considered several additional sources of bias unique to multi-trait analyses (i.e. uneven GWAS sample sizes, correlation in trait environmental noise) and found J-PALM robust to these as well.

We note several areas in which the study of polygenic adaptation can be advanced. Our operative model of polygenic adaptation is based on the Lande approximation, which over long time-courses will overestimate the efficiency of adaptation under stabilizing selection with a shift in the optimum [164, 188]. A model that incorporates these dynamics will potentially be better suited to detecting polygenic adaptation over longer time-courses, such as analyses of ancient DNA samples. Furthermore, under stabilizing selection more SNP heritability is expected to be sequestered to low-frequency alleles, and so common SNPs are expected to change less under adaptation than in our simulation model [159,

164].

Advances might also be made through more nuanced models that make fuller use of GWAS summary statistics and LD among GWAS marker. We showed our thresholding and pruning scheme for selecting sites did not substantially decrease our method's power. Pre-existing methods for fine-mapping or ascertaining pleiotropic loci might increase power even further [189]. It is also possible that for traits with extremely high polygenicity and/or low heritability, it will be necessary to utilize summary statistics that are sub-significant, and account for uncertainty in the location of the causal site. While in this paper we explored a thresholding and pruning scheme, which previous work and our own simulations show to be robust for stringent thresholding [179, 180], we have not established how results would differ for an LD clumping approach, or how misspecification of the LD reference panel (vs. the GWAS and/or population genetic cohort) affects our results.

We showed that PALM is substantially less prone to bias due to uncorrected GWAS stratification than comparable methods such as tSDS. However, we stress that PALM can nonetheless be biased under sufficiently strong uncorrected stratification. Other forms of stratification that we did not explore, such as gene-by-environment ($G \times E$) interactions, may be more difficult to account for via standard kinship-based approaches; however, new methods have recently arisen to this particular end [190].

Another limitation of our model is the interpretation of the estimates of the selection gradient and correlated response. We showed through simulations that when a genetically correlated trait with causal fitness effect is excluded from the analysis, estimates of direct selection have no causal interpretation. To address this, we introduced the notion of an effective selection gradient, which depends on which traits are modeled together. Estimates of the effective selection gradient allow us to determine whether a focal trait has evolved under correlated response another trait; however, this does not have the causal interpretation that the focal trait is under correlation response to a particular conditional trait.

Applying PALM to study evolution of 56 human traits in British ancestry, we found 8 traits under significant directional selection, recovering several previously-reported targets, such as pigmentation traits, educational attainment, and glycated hemoglobin (HbA1c), in agreement with previous findings of selection on these traits in Europe [167, 168, 144]. We also report several novel targets of directional selection, such as increased bone mineral density and decreased neuroticism. Despite historical claims of selection to increase height in Europe [173], we found no evidence for selection to increase height, consistent with recent analyses which showed that signals of directional selection on height have been drastically inflated by uncorrected population structure in GWAS summary statistics [179, 180].

We applied our joint test J-PALM to study 137 pairs of genetically correlated traits for signatures of correlated response. We found a highly significant enrichment of correlated response acting on these traits. Particularly, we found significant correlated response acting on pigmentation and life history traits (hair color, educational attainment). We showed that signal of selection on traits such as hair color and educational attainment,

which have been widely reported to date [167, 168, 144, 191], are due in significant part to correlated response to selection on other traits, vs direct selection acting on these traits.

One proposed theory for the diversification and increase of blonde hair color in Europe is sexual selection [192, 193]. However, our results do not support this, as we show that evidence for selection on hair color is attributable mostly to correlated response, beyond which there is little evidence for direct selection on this trait. This echos previous analysis showing selection at individual hair color loci may be indirect, via their pleiotropic effects (e.g. blonde hair gene *KITLG* responding to selection for tolerance to climate and UV radiation [194]), and conflict with arguments that hair color has been under direct sexual selection.

In our marginal test for selection, we detected significant selection to increase educational attainment, consistent with some previous work [168]. However, in a joint test with sunburn (i.e., "childhood sunburn occasions," the number of times the individual was sunburned as a child), strong signals of selection to increase educational attainment were significantly obviated. We conclude that signals of selection on educational attainment are driven significantly by correlated response. We caution that "childhood sunburn occasions" is a survey question, and is likely affected by many exogeneous factors beyond skin pigmentation (e.g., opportunity to visit the beach or use sunscreen). We propose that gene-by-environment ($G \times E$) interactions may be driving these signals of correlated response. Lewontin (1970), responding to Jensen (1969), pointed out that then-current estimates of intelligence quotient (IQ) heritability were inflated by $G \times E$ [195, 196]. Indeed, in modern-day GWAS, we see that educational attainment polygenic scores in the UKBB are only 50% as predictive in adoptees as in non-adoptees, indicating a significant role of $G \times E$ in the expression of educational attainment, as well as estimates of its heritability and genetic correlations [197]. The role of $G \times E$ or indirect genetic effects has been further illustrated by the discrepancy of sibling-based vs standard GWAS estimates of SNP effects on educational attainment50. Hence, genetic correlation of sunburn and educational attainment may be overestimated (e.g., $\widehat{r_g} = 0.24$ using UKBB GWAS [181]). We do not have data to elucidate the mechanism of this proposed $G \times E$ interaction, but hypothesize that educational opportunities and other environmental influences could be affected by skin pigmentation. Even in the absence of $G \times E$, we stress that our results are not interpretable as evidence of direct selection on "childhood sunburn occasions"–let alone skin pigmentation–following from our simulation study. Lastly, the inferred correlation between the traits and/or the signals of selection could be affected by uncorrected GWAS stratification [179, 180].

We found one case of significant antagonistic selection: T2D shows significant selection to increase, but this signal was initially occluded by the positive genetic correlation of T2D with negatively-selected glycated hemoglobin (HbA1c). Our joint analysis with J-PALM disentangles this antagonism between T2D and HbA1c, revealing latent adaptation of T2D. T2D is a complex disease with a complex etiology, involving obesity and various metabolic risk factors. Selection may have favored some of these factors under previous environmental conditions where both obesity and diets rich in simple sugars

were uncommon (also known as the thrifty gene hypothesis) [198]. HbA1c is a biomarker commonly used to not only diagnose pre-diabetes/diabetes, but also to monitor chronic hyperglycemia as a risk factor for vascular damage [199]. T2D and HbA1c are strongly, although imperfectly genetically correlated (r_g = 69%). Although this may seem peculiar as HbA1c is a diagnostic criterion for T2D, we speculate the distinction between these phenotypes could be driven by variation in HbA1c above and/or below diagnostic thresholds, or variation of other molecular traits (e.g. fasting glucose) that are also used as diagnostic criteria. , and HbA1c is also associated with hypertension and other cardiovascular disease independently of T2D incidence [187]. It is therefore possible that selection might have favored some of the traits underlying increased T2D risk, but acted against some of the more specific negative effects of T2D which now are measured by HbA1c [187, 199, 200]. These results provide evidence in support of the thrifty gene hypothesis [198].

Chapter 4

Finding polygenic gradients through ancestry disequilibrium regression

This work was co-advised by Noah Zaitlen and Rasmus Nielsen. It is unpublished.

Abstract

Differences in phenotypes between populations/ancestry groups can be caused by either environmental effects or genetic 'gradients'. Here, we present a method which detects genetic gradients (and distinguishes them from environmental effects) in a GWAS cohort. The method works by regressing uncorrected GWAS summary statistics on ancestry disequilibrium scores, a variation on the familiar LD score which we introduce here. Through simulations of complex trait architectures, we show our method is well-calibrated and well-powered to detect genetic gradients along major principal components. We also apply our method to 46 GWAS of human traits, finding evidence for polygenic adaptation along major genetic gradients in Europe, including behavior (e.g. anorexia), metabolic phenotypes (HbA1c, LDL cholesterol), and autoimmune diseases (celiac, lupus, Crohn's disease); our results that geographical variation of these traits in Europe is significantly driven by causal genetic factors.

4.1 Introduction

Stratification in association studies can be caused by dependence between ancestry and phenotype. However, the mechanism of this dependence can vary meaningfully. One such mechanism is environmental stratification (henceforth ES): when ancestry is correlated with exposure to an environmental variable that affects the phenotype, variants that are ancestry-specific will have inflated associations [201]. In light of this, many methods have been successfully deployed to control for stratification in genome-wide association study (GWAS) summary statistics (e.g., [202, 203, 204].

Another potential cause of stratification is correlation of causal genetic factors with ancestry (genetic stratification, henceforth GS). Techniques such as admixture mapping have established that certain complex traits are caused by ancestry-specific genetic variation [205, 206, 207]. However, admixture mapping detects locus-specific associations; many traits may have complex architectures with ancestry-phenotype associations that are too subtle to detect marginally. Furthermore, under both ES and GS, we expect similar inflation of associations for all ancestry-specific SNPs, whether or not they are causal, and thus it has been unclear how to distinguish these sources of stratification from GWAS summary statistics [177].

To this end, we present a method that uses GWAS summary statistics to detect polygenic gradients with respect to principal components, as a proxy for ancestry. Our method regresses summary statistics on a variation on the familiar LD score, which we term ancestry disequilibrium scores (AD scores) [177]. Our AD scores measure the cumulative ancestry tagged by a SNP. Conditional on the SNP's frequency, AD scores are dependent on GS, but not ES, allowing us to distinguish the two scenarios. Similar approaches defining a so-called signed LD profile have been used to estimate directional effects of functional annotations on polygenic risk (e.g. SLDP [208]). In fact, our proposed method reduces to a straightforward application of SLDP. Liu et al. 2018 proposed another method to estimate polygenic gradients, but their method is based on joint (vs marginal) GWAS estimates obtained via the inverse LD matrix, and assumes that population stratification is corrected for in the GWAS [209].

4.2 Model

Under our model, phenotypic variance can be partitioned into 4 sources:

1. Genetic variance (isotropic, i.e., irrespective of ancestry)
2. Genetic variance (due to ancestry)
3. Random noise (isotropic, i.e., irrespective of ancestry)
4. Environmental stratification (due to ancestry)

As a note: our model is very closely related to that described by Reshef, *et al.* (2018). We refer the reader to the Supplemental Texts of Reshef, *et al.* (2018) as well as Bulik-Sullivan *et al.* (2015) to round out the understanding of the random-effects models, and how their parameters are identifiable via the relationship between summary statistics (here, Z scores; in LDSC, χ^2 statistics) and LD patterns (here, linear combinations of signed LD R; in LDSC, linear combinations of R^2).

Our model assumes an additive phenotype with SNP heritability h^2 and a specified number of causal SNPs M (heretofore 'polygenicity'). We model phenotypes as a linear

function of mean-centered/variance-standardized genotypes $X \in \mathbb{R}^{N \times M}$ and covariates $U^{(K)} \in \mathbb{R}^{N \times K}$,where $K < \min(M, N)$:

$$\phi = X\beta + U^{(K)}\eta + \epsilon \tag{4.1}$$

Parameters that control the phenotype conditional on X and $U^{(K)}$ are β and η, which are random causal genetic effects and environmental (covariate) effects, respectively.

SNP effects model

Under our model, we assume that genetic variance is decomposed into two contributions: (1) Variance in mean effect, which depends on SNP loadings; and (2) isotropic variance, i.e. residual variance in effects, irrespective of SNP loadings.

To model this, we introduce a parameter γ, a length-K vector that specifies the degree to which causal genetic variation loads onto each principal component; and a statistic W, the so-called SNP 'loadings' onto covariates 1 through K. (See the subsequent section 'SNP loadings' for how these are calculated, e.g. when using principal component loadings.) We assume

$$E[\beta \mid W] = W\gamma/M, \quad \text{Cov}(\beta \mid W) = \nu/M \cdot I \tag{4.2}$$

where $\nu = h^2 - {h^2}_\gamma$.

Proposition: The total heritability (in the merger of all the ancestry groups) is h^2 if we set ${h^2}_\gamma = ||\Lambda\gamma||^2$.
Proof: (Provided $\text{Var}(\phi) = 1$,) The heritability h^2 under additivity follows

$$h^2 = \text{Var}\ (X\beta) \tag{4.3}$$

Since $E[X] = 0$,

$$= E[\beta^T X^T X\beta] \tag{4.4}$$
$$= E[E[\beta^T X^T X\beta \mid X]] \tag{4.5}$$

Notice that $\beta \mid X$ has the same distribution as $\widetilde{\beta} + X^T U\gamma$, where $\widetilde{\beta} \sim MVN(0, \nu/M \cdot I)$. Thus

$${h^2}_g = E[\gamma^T U^T XX^T XX^T U\gamma + E[\widetilde{\beta}^T X^T X\widetilde{\beta} \mid X]] \tag{4.6}$$

Let $R = X^T X$, the LD matrix. Then

$$E[\widetilde{\beta}^T X^T X \widetilde{\beta} \mid X]] = \text{trace}(E[\beta\beta^T]E[X^T X]) \tag{4.7}$$
$$= (\nu/M)\text{trace}(IR) = \nu \tag{4.8}$$
$$h^2{}_g = E[\gamma^T U^T X X^T X X^T U \gamma] + \nu \tag{4.9}$$
$$= \sum_{p=1}^{K} E[\lambda^2{}_p]\gamma_p^2 + \nu \tag{4.10}$$

And $\sum_{p=1}^{K} E[\lambda^2{}_p]\gamma_p^2 \to ||\Lambda\gamma||^2$ as $n \to \infty$, where Λ is the realized diagonal matrix of eigenvalues of XX^T. Then choosing $h^2{}_\gamma = ||\Lambda\gamma||^2$ we recover the desired total heritability h^2.

SNP loadings

Let $X = U\Sigma V^T$ be the singular value decomposition (SVD) of X, where U, Σ, V^T have dimensions $N \times N$, $N \times M$, and $M \times M$, respectively. The matrices U and V are orthonormal and Σ is diagonal matrix; note that since $M > N$ (usually the number of genotyped SNPs is much greater than the sample size), the trailing $M - N$ diagonal elements of Σ are zeros.

The so-called SNP loadings of are given by W, a $M \times K$ matrix of SNP loadings on PCs 1 through K:

$$W := X^T U = V\Sigma^T \Sigma U^T U = V\Sigma^T \tag{4.11}$$

Accounting for practical settings of PCA

In practice, PCA covariates are calculated using approximately independent SNPs, pruned for LD and filtered by MAF. Hence, here we reserve X to refer to the full genotype matrix, and here $Z \in \mathbb{R}^{N \times M'}$, where $M' \ll M$) are genotypes filtered and used to calculate covariates. Now we redefine U, Σ, V^T accordingly: $Z = U\Sigma V^T$. Let $W = X^T U$ be the loadings of the full genotypes onto covariates U.

Proposition 2: For general SNP loadings W (fixed/non-random), the total heritability is h^2 if we choose $h^2{}_\gamma = \gamma W^T R W \gamma$.

Proof: The heritability h^2 under additivity follows

$$h^2{}_g = E[\gamma^T W^T X^T X W \gamma] + \nu \tag{4.12}$$

Since we assume W is fixed,

$$h^2{}_g = \gamma^T W^T \, E[X^T X] W \gamma + \nu \tag{4.13}$$
$$= \gamma^T W^T R W \gamma + \nu. \tag{4.14}$$

Loading genetic correlations (random loadings, no filtering)

First we consider the case where $X = U\Sigma V^T$ and thus loadings W are random. We proved the amount of phenotypic variance due to genetic stratification is $||\Lambda\gamma||^2$. Further, the genetic correlation of phenotypes ϕ and SNP loading scores onto PC p, XX^TU_p (the latter is the contribution of genetic stratification to the polygenic score) is

$$r_p = corr(X\beta, XX^TU_p) \tag{4.15}$$
$$\text{Cov}(X\beta, XX^TU_p) = E[\beta]^TX^TXX^TUp \tag{4.16}$$
$$= \gamma^T\Lambda^2U_p \tag{4.17}$$
$$\text{Var}(X\beta) = h^2,\ \text{Var}((XX^TU_p) = {\lambda_p}^2 \tag{4.18}$$

Thus,

$$r_p = \frac{\gamma^T\Lambda^2U_p}{\lambda_p\sqrt{h^2}} \tag{4.19}$$

And the proportion of the genetic variance explained by PC p is

$$\frac{{h^2}_{\gamma,p}}{{h^2}_g} = \frac{(\lambda\gamma)^2}{h^2} \tag{4.20}$$
$$\tag{4.21}$$

Loading genetic correlations (fixed loadings, with filtering)

Next, we consider the case where $Z = U\Sigma V^T$ and Z is a filtered genotype matrix (say, obtained by filtering X by MAF and pruning for LD). We assume annotations (loadings W) are fixed/nonrandom, consistent with the approach of LD score approaches (e.g., Reshef 2018). Let $r_p = corr(X\beta, XW_p)$. Since $E[X] = 0$,

$$\text{Cov}(X\beta, XW_p) = E[\beta]^TE[X^TX]Wp \tag{4.22}$$
$$= (W\gamma)^TRW_p \tag{4.23}$$
$$\text{Var}(X\beta) = h^2,\ \text{Var}(XW_p) = {W_p}^TRW_p \tag{4.24}$$

Thus,

$$r_p = \sum_{k=1}^{K}\gamma_k \cdot \frac{{W_k}^TRW_p}{\sqrt{h^2 \cdot {W_p}^TRW_p}} \tag{4.25}$$

And the proportion of heritability explained by PC p is

$$\frac{{h^2}_{\gamma,p}}{h^2} = \gamma^2 \cdot \frac{{W_p}^T R W_p}{h^2} \tag{4.26}$$

Identifiability of ES vs. GS

Note that under a model where $\beta \sim MVN(\mu, \nu)$, then this reduces to a model where η, γ are unidentifiable (hat tip to Andy Dahl and Po-Ru Loh for pointing out this issue); $\beta = \gamma + \mu$, where $\gamma \sim MVN(0, \nu)$:

$$\begin{aligned}
\phi &= X(\mu + \gamma) + U\eta \ + \epsilon && (4.27)\\
&= U\Lambda\gamma + X\gamma + U\eta + \epsilon && (4.28)\\
&= X\gamma + U(\Lambda\gamma \ + \ \eta) \ + \epsilon && (4.29)\\
&= X\gamma + U\eta\prime \ + \epsilon && (4.30)
\end{aligned}$$

Hence a model with (γ, η) is unidentifiable from one with $(0, \Lambda\gamma + \eta)$. However, if we allow for sparsity,

$$\beta \sim MVN(\frac{1}{\sqrt{Mq}} W\gamma, \frac{1}{Mq}\nu \cdot I) \text{ w.p. } q, \quad \beta = 0 \text{ otherwise} \tag{4.31}$$

then

$$\phi = \ X(\frac{1}{\sqrt{Mq}} W\gamma + \delta) + U\eta + \epsilon \tag{4.32}$$

where $\delta \sim MVN(0, \frac{1}{Mq}\nu \cdot I)$, and

$$\phi = X(\delta \cdot Y) + U(\frac{1}{\sqrt{Mq}}\Lambda\gamma \cdot Y + \eta) \ + \epsilon \tag{4.33}$$

Where $Y_i \sim \text{Bern}(Mq)$ for $i = 1, 2, \ldots, M$. Thus, if we assume sparsity of β, then γ, η are identifiable.

4.3 Estimating genetic gradients

Now, we describe the behavior of the SNP effect estimates in a GWAS. Note that we do not assume the inclusion of any covariates in the GWAS. The Z-score for the SNP is $Z_{j=}{X_j}^\top \phi/\sqrt{N}$. This has expectation

$$E[\ Z_j \mid X] = E[X_j^T(X\beta + U\eta + \epsilon)/N|X] \tag{4.34}$$

$$= E[\sum_{k=1}^{M} N r_{jk}\beta_k + W_j\eta)/\sqrt{N}|\ X] \tag{4.35}$$

Because $E[\beta_k \mid X_k] = Wk\gamma/Mq$,

$$= \frac{1}{\sqrt{N}}\,(\frac{N}{\sqrt{Mq}}\sum_{k=1}^{M} r_{jk}W_k\gamma + W_j\eta) \tag{4.36}$$

Where $r_{jk} = \frac{1}{N}X_j^T X_k$ is the genotypic correlation coefficient for SNPs j, k. Finally, we get

$$= \sqrt{\frac{N}{Mq}}\sum_{p=1}^{K}\gamma_p s_{jp} + \frac{1}{\sqrt{N}}\sum_{p=1}^{K}\eta_p w_{jp} \tag{4.37}$$

Where

$$s_{jp} := \sum_{k=1}^{M} w_{kp} r_{jk} \tag{4.38}$$

is similar to an annotation-weighted LD score (see Gazal, et al. (2017)), where instead of squared LD ${r_{jk}}^2$, we sum over signed LD r_{jk}, and the annotations are SNP loadings onto PCs 1 through K. We call s_{jp} the ancestry disequilibrium (AD) score of SNP j on principal component p.

The equation above shows that bias in the Z-scores can be driven by an LD-dependent term that captures genetic stratification, and a term independent of LD that captures environmental stratification. We can estimate γ and η jointly in a standard multiple linear regression, and estimate standard errors through block-jackknife.

From Reshef, et al. (2018) we know that

$$\mathrm{Cov}(Z \mid \gamma)\ \approx {\sigma_e}^2 R^2 + R/N \tag{4.39}$$

where R is the LD matrix, ${\sigma_e}^2$ is the proportion of phenotypic variance due to random noise, and N is the sample size. This covariance follows from a model with random genotypes and causal genetic effects. Similar to Reshef, et al. (2018), we use the covariance matrix to account for dependence and heteroscedasticity in the residuals of Z regressed onto AD scores.

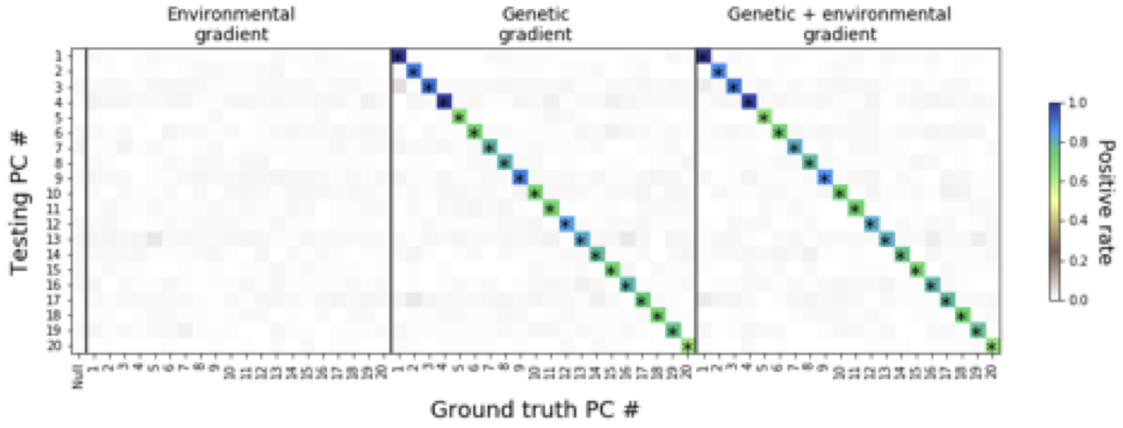


Figure 4.1: **Performance of ADR in simulations** Testing for genetic stratification on simulations under the null (no stratification; leftmost column), purely environmental stratification (leftmost box), purely genetic stratification (middle box), and genetic + environmental stratification (rightmost box) simulations. Stars indicate positive rate significantly >0.005. For each simulation scenario (row), all PCs are tested jointly for a genetic stratification effect. Marginal estimates used in simulation are estimated using K=40 top PCs as covariates. Simulations were performed using 1000 Genomes European individuals (resampled up to cohort size $N = 300,000$), simulating trait architectures on chromosome 3 and PC covariates estimated LD-pruned SNPs genomewide. We assume 10,000 causal SNPs.

4.4 Results

Simulations

To validate the ADR method, we conducted simulations of a heritable phenotype with $h^2 = 50\%$ under 4 conditions: null simulations, environmental stratification (ES), genetic stratification (GS), and a mix of the two sources of variation (Fig. 4.1, Fig. 4.2). When included in the simulations, we assume that 10% of complex trait variation is driven by each form of stratification. Each simulation imposes ES and/or GS on a single PC at a time (e.g., PC1), but tests all PCs 1-20 jointly (columns of Fig. 4.1). Across all PCs, ADR has > 50% power (nominal $\alpha = 5\%$) to detect genetic differentiation in GS simulations, with or without ES imposed on top of that (Fig. 4.1) while being well-calibrated under the null (Fig. 4.1, Fig. 4.2).

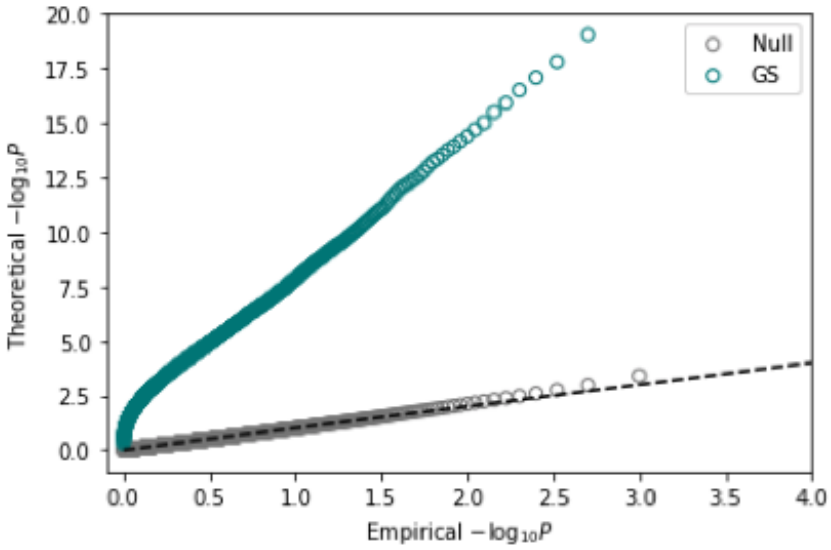


Figure 4.2: **QQ plot from simulations** Null simulations are aggregate χ^2 statistics from all null tests (i.e. any test where GS is not occurring on the tested PC). GS simulations are aggregate χ^2 statistics from all where GS is occurring on the tested PC.

Analysis of human GWAS

We applied our method using the same AD scores derived for our simulation experiments. We used SLDP [208] to estimate directional effects of ancestry on 46 complex traits in humans. We computed ancestry disequilibrium scores using SLDP, with annotations corresponding to SNP loadings on each PC independently. Along PCs 1-4 (illustrated in Fig. 4.3), we tested for genetic gradients by estimating these directional effects and computing two-sided permutation-based P-values using SLDP (Fig. 4.4, Fig. 4.5, Table 4.1). In total, we find 8 traits that are significantly genetically differentiated with $P < 0.05$ (Bonferroni) (Table 4.1).

Stratified analysis using functional annotations

One way to validate the significant genetic gradients identified in the previous section is to demonstrate that they are enriched in regions of the genomic with penetrant biological effects. To investigate this, I re-estimated genetic gradients, but this time I stratified the estimates by using binary functional annotation (e.g. 'repressed', 'non-synonymous') (Fig. 4.6). To calculate the AD scores for covariate p, stratified on annotation a, I simply take

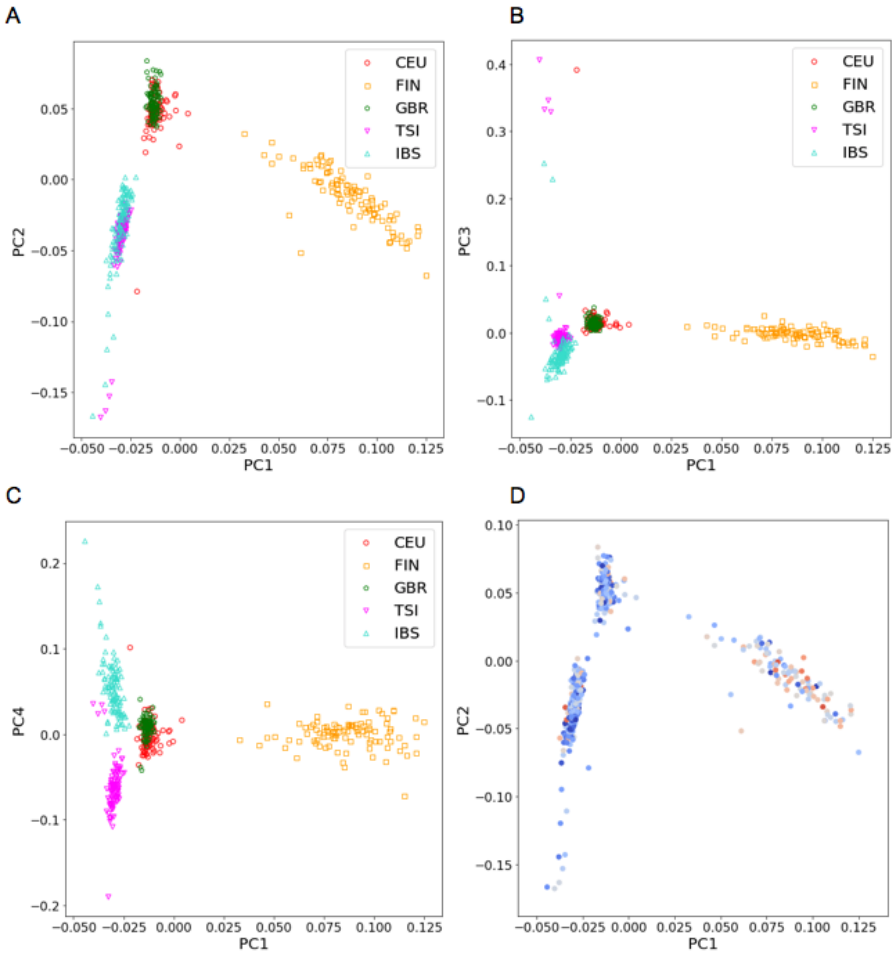


Figure 4.3: (A-C) PCA of European 1000 Genomes individuals used in simulations & AD score calculations. (D) Simulated phenotypes (represented by red-blue colors) vs. PC1 & 2; here $\eta = 0.2$ (i.e., there is an environmental effect). Linear regression of the phenotypes on PC1 is highly significant ($P = 10^{-17}$).

Trait	PC1	PC2	PC3	PC4
Anorexia	9.17*	16.37*	6.26*	7.15*
ENIGMA2_MeanPutamen	8.63*	5.45*	4.43*	2.45
Lupus	6.95*	3.67*	2.51	1.92
LDL	3.09	6.71*	3.20	4.74*
HbA1C	2.81	5.30*	1.84	1.96
Celiac	1.08	3.01	4.77*	-4.04*
cov_EDU_YEARS	-5.90*	4.25*	4.47*	-8.17*
Primary_biliary_cirrhosis	0.34	1.17	1.21	2.30
CD	0.07	1.71	2.08	-0.32
Rheumatoid_Arthritis	0.37	0.50	1.89	0.66
pigment_HAIR	-0.21	1.85	-0.51	0.82
blood_EOSINOPHIL_COUNT	1.11	1.24	1.82	1.17
body_HEIGHTz	0.01	1.79	-0.97	0.17
disease_RESPIRATORY_ENT	1.57	1.09	0.18	0.02
disease_DERMATOLOGY	-2.03	1.54	-0.13	-1.02
disease_ALLERGY_ECZEMA_DIAGNOSED	-0.63	-0.74	-1.44	1.53
blood_HIGH_LIGHT_SCATTER_RETICULOCYTE_COUNT	-0.57	1.47	0.76	-0.79
disease_HI_CHOL_SELF_REP	1.07	0.48	1.39	1.46
cov_SMOKING_STATUS	1.29	-0.19	0.20	1.42
Autism	-0.15	-0.53	-0.56	1.05
ALZ	1.02	0.79	0.01	-0.34
disease_T2D	1.02	0.63	0.47	-0.90
body_BALDING1	0.96	-0.16	-0.05	-1.53
disease_AID_SURE	-1.45	0.81	0.93	-1.75
UC	-1.09	0.46	0.93	-0.22
bp_SYSTOLICadjMEDz	-1.34	-0.32	0.90	-0.06
mental_NEUROTICISM	-5.19*	-1.43	0.87	-0.53
body_WHRadjBMIz	0.78	-1.04	-2.50	-0.85
repro_MENOPAUSE_AGE	-1.23	0.74	-0.28	-1.29
blood_WHITE_COUNT	-2.26	-2.62	-1.25	0.71
pigment_SUNBURN	0.60	0.42	-0.52	0.71
lung_FEV1FVCzSMOKE	0.54	-0.32	-1.09	0.23
blood_RBC_DISTRIB_WIDTH	-1.54	0.49	0.29	-1.55
blood_RED_COUNT	-0.64	-1.56	0.43	0.19
blood_PLATELET_COUNT	-0.14	-1.37	0.22	-0.22
lung_FVCzSMOKE	-1.60	0.17	0.12	-1.10
bmd_HEEL_TSCOREz	-2.67	-3.58	-2.29	0.15
Coronary_Artery_Disease	-1.86	-1.50	0.08	-1.73
Schizophrenia	-0.37	-0.76	-0.42	0.04
disease_HYPOTHYROIDISM_SELF_REP	-3.09	-2.10	-1.73	-0.41
body_BMIz	-1.03	-2.47	-0.55	-0.73
blood_LYMPHOCYTE_COUNT	-0.64	-3.15	-1.31	-2.67
repro_MENARCHE_AGE	-2.63	-2.47	-1.08	-0.81
blood_MONOCYTE_COUNT	-2.38	-1.70	-1.55	-0.94

Table 4.1: **Testing for adaptation in genetic stratification of 46 human traits** Z-scores of trait genetic stratification estimates along PCs 1-4. Stars indicate tests with Bonferroni $P < 0.05$.

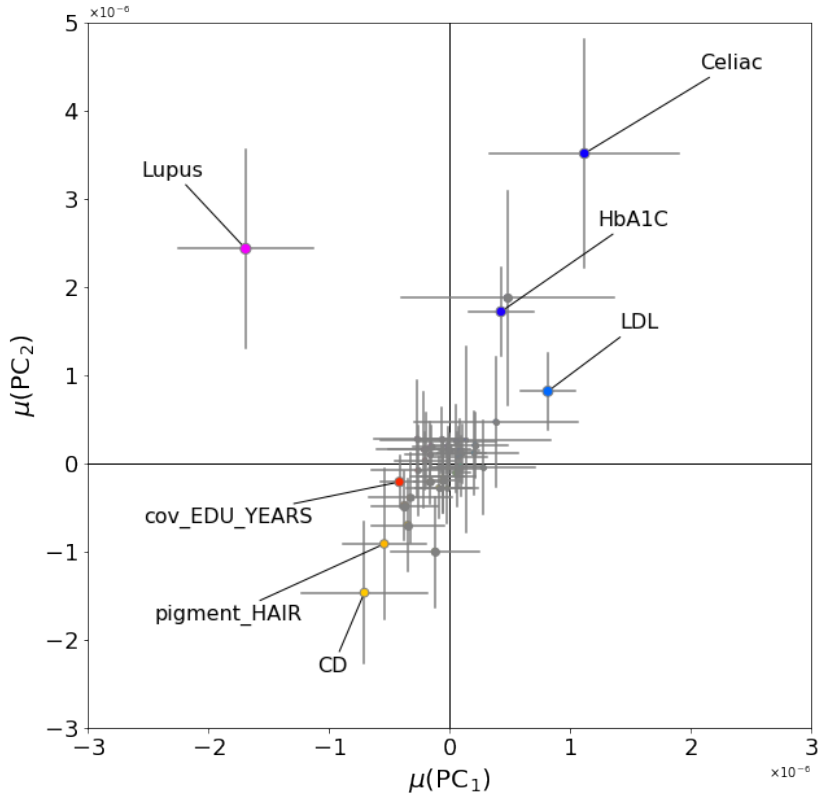


Figure 4.4: **Genetic stratification effect estimates for 46 human traits (PC1 vs. PC2).** Estimates for PC1 vs. PC2. Colored dots are significant at $P < 0.05$, Bonferroni-corrected. Gray lines are 95% CIs. Colors represent the angle of (μ_x, μ_y). P-values computed used permutation test (see e.g., Reshef 2018)

$$s_{jp,a} = \sum_{k=1}^{M} w_{jp} r_{jk} \cdot I(A_k = a), \tag{4.40}$$

where $I(A_k = a)$ is simply an indicator that SNP k has the annotation a. Using SLDP, I computed r_p (see Reshef et al 2018 for detailed explanation of r_f, their analogous quantity); r_p is essentially a measure of how correlated effect sizes are with SNP PC loadings. Let $r_p(a)$ be the functional correlation for PC p, stratified on annotation a. In this work, I used annotations used by Weissbrod, *et al.* (2020).

I then calculated a statistic D, which is basically just a comparison between r_p for a particular annotation a and a baseline annotation (I use 'Repressed'):

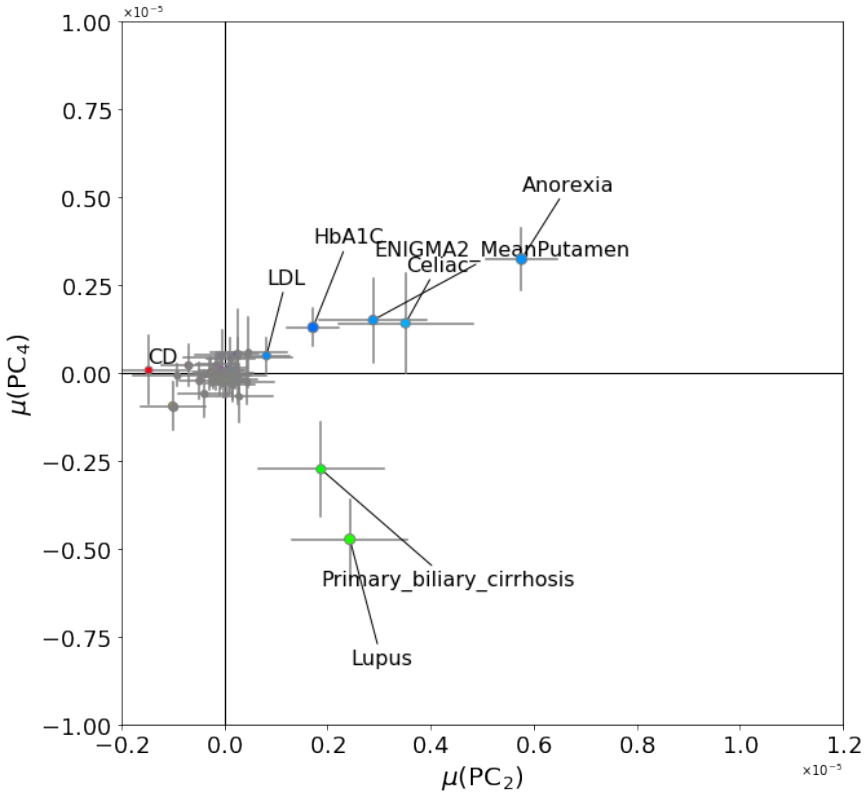


Figure 4.5: **Genetic stratification effect estimates for 46 human traits (PC2 vs. PC4).** Estimates for PC2 vs. PC4. Colored dots are significant at $P < 0.05$, Bonferroni-corrected. Gray lines are 95% CIs. Colors represent the angle of (μ_x, μ_y). P-values computed used permutation test (see e.g., Reshef 2018)

$$D := \text{sign}\Big(r_p(\text{Repressed})(r_p(a) - r_p(\text{Repressed})\Big) \tag{4.41}$$

Under the null that there is no enrichment of genetic gradients in particular functional annotations, we should expect $E[D] = 0$. By contrast, we see that for many axes of population variation (e.g., PC1 & 2), there are significant enrichments for genetic gradients in non-synonymous, synonymous, and other functional annotations (e.g. methylation sites) (Fig. 4.6). For other regions of less penetrant effect (e.g. intron), we see smaller/insignificant enrichment signal (Fig. 4.6).

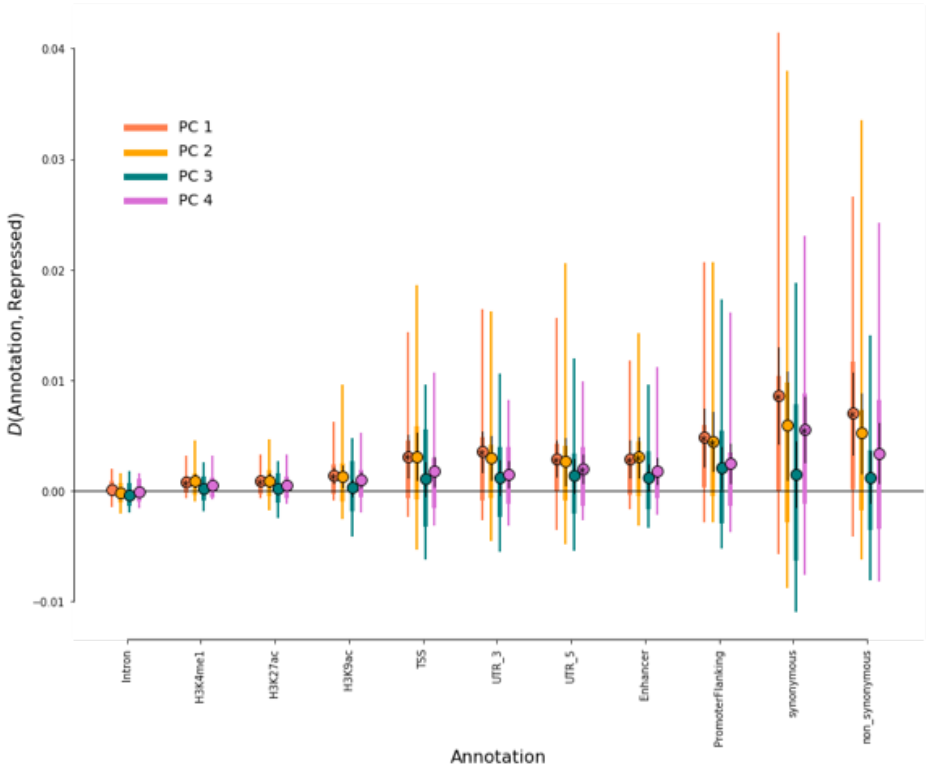


Figure 4.6: **Meta-analysis of functional annotation-stratified analyses for 46 human traits.** SLDP (Reshef, *et al* 2018) results across 46 complex traits, estimating directional effects of ancestry (PCs 1-4 in Europe), stratified by binary functional annotations. Dots indicate sample mean D across traits, whereas black lines show a 95% CI on mean D, and thick/thin colored lines indicate IQR/5-95 percentile on D. Starred dots indicate $P < 0.05$ after Bonferroni correction with $N = 4 \times 46 \times 11 = 2024$.

4.5 Conclusion & future work

Here we have developed a method to infer genetic gradients (and distinguish them from environmental effects) causing complex trait differences amongst ancestry groups. We validate this via bespoke simulations of complex trait variation based on empirical genetic data from Europeans in the 1000 Genomes Project. We then applied our approach to an analysis of GWAS of 46 human traits (summary statistics from Reshef *et al*, 2018) and find a number of traits whose genetic basis significantly differentiated along axes of population structure (i.e. major PCs) within Europe. We validate our findings through an analysis stratified on functional annotations, which show enrichment of this genetic differentiation particularly in functional regions (e.g. protein-coding, promoter-flanking).

Looking forward, many analyses should be considered, including:

- Simulations
  - Redo simulations using SLDP (Reshef *et al* 2018) rather than bespoke method, for consistency with empirical analysis.
  - MAF and LD-dependent architectures with different levels of sparsity (SLDP does control for MAF, but not levels of LD *a la* Gazal, et al. 2017)
  - Environmental *variance* effects
  - Genetic correlations and genetic causality (see below, special case of $G \times E$)
  - $G \times E$ simulations of the form $G \rightarrow \phi_1$, $\phi_1 \times E \rightarrow \phi_2$, and ϕ_1 has genetic differences between populations.
  - Binary annotation effects under simulation with known enrichments.
  - Compare to Xuanyao Liu's method (PopDiff)
  - What happens if you don't use PCs in your gwas, or if the PCs are chosen badly? In general what if pop structure is not 100% corrected for?
  - Choice of population structure correction (e.g.LMMs, fixed PCs) (we only looked at the latter so far)
- Additional analyses
  - Indirect effects: Re-analyse traits (especially EA!) using sib-based or other GWAS to control for indirect effects
  - Extend the number of phenotypes and populations examined
  - Extend ancestry factors considered (e.g. use measures of local ancestry based on ancient genomes, e.g. genetic effect of Steppe ancestry)
  - Use these to consider timing/age of selection?
  - Measure PC loading (and F_{st}) differences under annotations
  - Once we have a set of phenotypes we believe have changed due to genetic differences, is there another test we can perform to validate without using the effect size estimates from the GWAS? E.g., is F_{st} at non-synonymous variants that are robustly GWAS significant (and fine-mapped?) different than F_{st} at general non-synonymous variants?

4.6 Methods

Principal components analysis (PCA)

We use genotype calls for European 1000 Genomes Phase 3 individuals. We prune LD in PLINK with `-indep-pairwise 100 5 0.15`. We also exclude regions of high LD (`https:`

//genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD)) and the MHC region (https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37). We run PCA on these data using PLINK:

```
./plink2 --bed 1000G_merged \
    --maf 0.05 \
    --hwe 1e-5 \
    --pca 40 \
    --out 1000G_merged
```

Then, we run the following to obtain SNP loadings on PCs 1-40:

```
./plink2 --bed 1000G_merged \
    --variant-score 1000G_merged.covar \
    --out loadings
```

Ancestry disequilibrium scores

We calculate ancestry disequilibrium (AD) scores using SLDP [208] with loadings from the previous step applied as the annotations of interest. We use 1000 Genomes Phase 3 European indivudals (https://data.broadinstitute.org/alkesgroup/LDSCORE/1kg_eur.tar.bz2) and a window size of 1cM around the focal SNP. Analysis is restricted to $\sim 10^7$ HapMap3 SNPs [177].

LD matrix calculation and approximation

To improve efficiency of the regression of our AD score regression, we account for heteroscedasticity and off-diagonal covariances by performing a generalize least squares (GLS) regression. Following Reshef, et al. (2018), the covariance matrix follows

$$\text{Cov}(Z \mid \gamma) \approx \sigma_e^2 R^2 + R/N \tag{4.42}$$

where ${\sigma_e}^2 = 1-h^2$ and R is the signed LD matrix. We make a block-diagonal approximation to R following Berisa & Pickrell (2016), assuming that SNPs between LD blocks are in linkage equilibrium. Additionally, we make a low-rank approximation to the blockwise LD matrices, storing just enough of the top principal components such that the >95% of the variance is accounted for. This captures the vast majority of LD patterns in each block, and significantly improves computational efficiency, since by this criterion the requisite number of PCs is usually ~2 orders of magnitude lower than the number of SNPs in the block.

Simulation of genetic architecture and GWAS

We use genotypes X and PCs U obtained from 1000 Genomes (1KG) Phase 3 Europeans and simulate polygenic trait architectures atop this data. (See Methods: PCA). Assume L total polymorphisms in this data. We assume a polygenic trait with $M = Y_1 + Y_2 + ... + Y_L$ sparsely causal sites with normally-distributed effects, where $Y_i \sim \text{Bern}(q)$ i.i.d. for $i = 1, 2, .., L$:

$$\beta_i \sim N\Big(\frac{W_i g}{\sqrt{Lq}}, \frac{h^2 - ||g||^2}{Lq}\Big) \text{ if } Y_i = 1, \text{ otherwise } \beta_i = 0 \tag{4.43}$$

where g is a vector of K parameters specifying genetic gradients along the top K principal components, such that $0 \leq ||\Lambda g||^2 \leq h^2 \leq 1 - ||\eta||^2$, where η is a parameter specifying ancestry-dependent environmental effects. The vector W_i is the SNP loading of i onto PCs 1 through K.

Since it is not feasible to simulate genetic architectures in extremely large cohorts, we simulate GWAS in this smaller cohort ($N_{1KG} = 489$) but modify the sampling noise to yield summary statistics with the desired noise level (e.g., $N = 10^5$). Since the standard error of scales with $1/\sqrt{N}$, we reduce the standard deviation of isotropic random noise ϵ by a factor of $c = \sqrt{N_{1KG}/N}$.

We simulate phenotypes as

$$\phi \sim MVN(X\beta + U\eta,\ cI) \tag{4.44}$$

where $X\beta$ is a vector of polygenic scores and the scalar c is chosen to standardize $\text{Var}(\phi) = 1$. We compute polygenic scores using PLINK 2:

```
./plink2 --bed 1000G_merged \
    --score betas.txt \
    --out scores
```

We then run association testing in PLINK 2:

```
./plink2 --bed 1000G_merged \
    --glm hide-covar \
    --covar 1000G_merged.eigenvec \
    --pheno phenos.txt \
    --out sumstats \
```

Note that we include covariates when we run association testing! This is because in preliminary work, we attempted to estimate not only γ (the genetic gradient), but also η (the environmental gradient), and when the totally uncorrected GWAS is used, this resulted in very overdispersed estimates of η.

Calculating AD scores

Using the SLDP package (https://github.com/yakirr/sldp), we performed `preprocessannot` to produce AD scores for all of our traits along major PCs, whose SNP loadings we calculated in the aforementioned section:

```
preprocessannot --sannot-chr pc<p>/ \
--bfile-chr refpanel/plink_files/1000G.EUR.QC. \
--print-snps refpanel/1000G_hm3_noMHC.rsid \
--ld-blocks refpanel/pickrell_ldblocks.hg19.eur.bed \
--chroms <n>
```

for $n = 1, 2, \ldots, 22$ and $p = 1, 2, 3, 4$. The folder pc<p>/ must contain `<n>.sannot.gz`, a file with the SNP loading of each SNP on chromosome n.

We did this again for the stratified analysis, setting the weighting to be 0 at SNPs that do not have the specified functional annotation.

See SLDP wiki and data page (https://data.broadinstitute.org/alkesgroup/SLDP/) for more details and refpanel downloads.

Estimating genetic gradients

Again using the SLDP package, we performed `sldp`:

```
sldp --pss-chr sumstats/complex/ALZ.KG3.95/ \
--sannot-chr ../ADR/annot/pc<p>/ \
--background-sannot-chr background/maf5/ \
--outfile-stem sumstats/complex/ALZ.KG3.95.pc<p> \
--ld-blocks refpanel/pickrell_ldblocks.hg19.eur.bed \
--svd-stem svd/svds_95percent \
--bfile-reg-chr refpanel/plink_files/1000G.EUR.QC.hm3_noMHC. \
--seed 0
```

where `ALZ.KG3.95/` is an example folder containing pre-processed summary statistics (for Alzheimer's disease) for SLDP, available on the SLDP data page. Here p again is the desired PC.

Note: I forget what happened, but there was an issue with the downloadable SVDs (of the LD matrix R) from the SLDP data page. I think I had to recompute these from scratch. Send me an email if you need them.

Bibliography

[1] Aaron J Stern and Rasmus Nielsen. "Detecting Natural Selection". In: *Handbook of Statistical Genomics: Two Volume Set* (2019), pp. 397–40.

[2] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[3] Sewall Wright. "Evolution in Mendelian populations". In: *Genetics* 16.2 (1931), pp. 97–159.

[4] Motoo Kimura. "Solution of a process of random genetic drift with a continuous model". In: *Proceedings of the National Academy of Sciences* 41.3 (1955), pp. 144–150.

[5] Motoo Kimura. "STOCHASTIC PROCESSES AND DISTRIBUTION OF GENE FREQUENCIES UNDER NATURAL SELECTIONi". In: *Population Genetics: The Nature and Causes of Genetic Variability in Populations* 20 (1955), p. 33.

[6] Motoo Kimura. "On the probability of fixation of mutant genes in a population". In: *Genetics* 47.6 (1962), pp. 713–719.

[7] Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*. Vol. 27. Springer Science & Business Media, 2012.

[8] Deborah Charlesworth. "Balancing selection and its effects on sequences in nearby genome regions". In: *PLoS genetics* 2.4 (2006), e64.

[9] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. "Contrasting the genetic architecture of 30 complex traits from summary association data". In: *The American Journal of Human Genetics* 99.1 (2016), pp. 139–153.

[10] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. "An expanded view of complex traits: from polygenic to omnigenic". In: *Cell* 169.7 (2017), pp. 1177–1186.

[11] Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. "Estimation of 2Nes from temporal allele frequency data." In: *Genetics* 179.1 (May 2008), pp. 497–502. ISSN: 0016-6731. DOI: 10.1534/genetics.107.085019. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2390626%5C&tool=pmcentrez%5C&rendertype=abstract.

[12] Gregory I Lang et al. "Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations". In: *Nature* 500.7464 (2013), p. 571.

[13] Benjamin H Good et al. "The dynamics of molecular evolution over 60,000 generations". In: *Nature* 551.7678 (2017), p. 45.

[14] Iosif Lazaridis, Nick Patterson, and *et al Mittnik Alissa*. "Ancient human genomes suggest three ancestral populations for present-day Europeans." In: *Nature* 513.7518 (Sept. 2014), pp. 409–13. ISSN: 1476-4687. DOI: 10.1038/nature13673. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4170574%5C&tool=pmcentrez%5C&rendertype=abstract.

[15] Iain Mathieson et al. "Genome-wide patterns of selection in". In: *Nature* 528.7583 (2015), pp. 499–503. ISSN: 0028-0836. DOI: 10.1038/nature16152. URL: http://dx.doi.org/10.1038/nature16152.

[16] Anna-Sapfo Malaspinas et al. "Estimating allele age and selection coefficient from time-serial data." In: *Genetics* 192.2 (Oct. 2012), pp. 599–607. ISSN: 1943-2631. DOI: 10.1534/genetics.112.140939. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3454883%5C&tool=pmcentrez%5C&rendertype=abstract.

[17] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. "Identifying signatures of selection in genetic time series." In: *Genetics* 196.2 (Feb. 2014), pp. 509–22. ISSN: 1943-2631. DOI: 10.1534/genetics.113.158220. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3914623%5C&tool=pmcentrez%5C&rendertype=abstract.

[18] Joshua G Schraiber, Steven N Evans, and Montgomery Slatkin. "Bayesian Inference of Natural Selection from Allele Frequency Time Series." In: *Genetics* 203.1 (May 2016), pp. 493–511. ISSN: 1943-2631. DOI: 10.1534/genetics.116.187278. URL: http://www.ncbi.nlm.nih.gov/pubmed/27010022.

[19] Austin L Hughes and Masatoshi Nei. "Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection". In: *Nature* 335.6186 (1988), p. 167.

[20] Ruth Hershberg and Dmitri A Petrov. "Selection on codon bias". In: *Annual review of genetics* 42 (2008), pp. 287–299.

[21] Rotem Sorek and Gil Ast. "Intronic sequences flanking alternatively spliced exons are conserved between human and mouse". In: *Genome Research* 13.7 (2003), pp. 1631–1637.

[22] Adam J Hockenberry et al. "Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands". In: *Molecular biology and evolution* (2017), msx310.

[23] Sewall Wright. "The distribution of gene frequencies under irreversible mutation". In: *Proceedings of the National Academy of Sciences* 24.7 (1938), pp. 253–259.

[24] Adam Eyre-Walker and Peter D Keightley. "The distribution of fitness effects of new mutations". In: *Nature Reviews Genetics* 8.8 (2007), p. 610.

[25] Stanley A Sawyer and Daniel L Hartl. "Population genetics of polymorphism and divergence." In: *Genetics* 132.4 (1992), pp. 1161–1176.

[26] Adam R Boyko et al. "Assessing the evolutionary impact of amino acid mutations in the human genome". In: *PLoS genetics* 4.5 (2008), e1000083.

[27] Yun-Xin Fu. "Statistical properties of segregating sites". In: *Theoretical population biology* 48.2 (1995), pp. 172–197.

[28] RC Lewontin and Jesse Krakauer. "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms". In: *Genetics* 74.1 (1973), pp. 175–195.

[29] Enrique Santiago and Armando Caballero. "Variation after a selective sweep in a subdivided population". In: *Genetics* 169.1 (2005), pp. 475–483.

[30] John Maynard Smith and John Haigh. "The hitch-hiking effect of a favourable gene". In: *Genetics Research* 23.1 (1974), pp. 23–35.

[31] Norman L Kaplan, RR Hudson, and CH Langley. "The" hitchhiking effect" revisited." In: *Genetics* 123.4 (1989), pp. 887–899.

[32] John M Braverman et al. "The hitchhiking effect on the site frequency spectrum of DNA polymorphisms." In: *Genetics* 140.2 (1995), pp. 783–796.

[33] Pardis C Sabeti et al. "Detecting recent positive selection in the human genome from haplotype structure". In: 419.October (2002). DOI: 10.1038/nature01027.1..

[34] Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. "Natural selection and the distribution of identity-by-descent in the human genome". In: *Genetics* 186.1 (2010), pp. 295–308.

[35] Yuseob Kim and Rasmus Nielsen. "Linkage disequilibrium as a signature of selective sweeps". In: *Genetics* 167.3 (2004), pp. 1513–1524.

[36] Joachim Hermisson and Pleuni S Pennings. "Soft sweeps: molecular population genetics of adaptation from standing genetic variation". In: *Genetics* 169.4 (2005), pp. 2335–2352.

[37] Molly Przeworski, Graham Coop, and Jeffrey D Wall. "The signature of positive selection on standing genetic variation". In: *Evolution* 59.11 (2005), pp. 2312–2323.

[38] Nandita R Garud et al. "Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps." In: *PLoS genetics* 11.2 (Mar. 2015), e1005004. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005004. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4338236%5C&tool=pmcentrez%5C&rendertype=abstract.

[39] Richard R Hudson and Norman L Kaplan. "The Coalescent Process in Models With Selection and Recombination". In: 840 (1988), pp. 831–840.

[40] Jonathan K Pritchard, Joseph K Pickrell, and Graham Coop. "The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation". In: *Current biology* 20.4 (2010), R208–R215.

[41] Rasmus Nielsen et al. "Genomic scans for selective sweeps using SNP data". In: (2005), pp. 1566–1575. DOI: 10.1101/gr.4252305..

[42] Kosuke M Teshima, Graham Coop, and Molly Przeworski. "How reliable are empirical genomic scans for selective sweeps ?" In: 773 (2006), pp. 702–712. DOI: 10.1101/gr.5105206.702.

[43] Nicolas Galtier, Frantz Depaulis, and Nicholas H Barton. "Detecting bottlenecks and selective sweeps from DNA sequence polymorphism". In: *Genetics* 155.2 (2000), pp. 981–987.

[44] Thomas Städler et al. "The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations". In: *Genetics* 182.1 (2009), pp. 205–216.

[45] Daniel R Schrider et al. "Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps". In: *Genetics* 200.1 (2015), pp. 267–284.

[46] Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo. "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes." In: *Molecular biology and evolution* 2.2 (1985), pp. 150–174.

[47] Masatoshi Nei and Takashi Gojobori. "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." In: *Molecular biology and evolution* 3.5 (1986), pp. 418–426.

[48] Spencer V Muse and Brandon S Gaut. "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome." In: *Molecular biology and evolution* 11.5 (1994), pp. 715–724.

[49] Nick Goldman and Ziheng Yang. "A codon-based model of nucleotide substitution for protein-coding DNA sequences." In: *Molecular biology and evolution* 11.5 (1994), pp. 725–736.

[50] Ziheng Yang. "PAML 4: phylogenetic analysis by maximum likelihood". In: *Molecular biology and evolution* 24.8 (2007), pp. 1586–1591.

[51] Sergei L Kosakovsky Pond and Spencer V Muse. "HyPhy: hypothesis testing using phylogenies". In: *Statistical methods in molecular evolution*. Springer, 2005, pp. 125–181.

[52] Ziheng Yang. "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution." In: *Molecular biology and evolution* 15.5 (1998), pp. 568–573.

[53] Rasmus Nielsen and Ziheng Yang. "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene". In: *Genetics* 148.3 (1998), pp. 929–936.

[54] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level". In: *Molecular biology and evolution* 22.12 (2005), pp. 2472–2479.

[55] Wendy SW Wong et al. "Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites". In: *Genetics* 168.2 (2004), pp. 1041–1051.

[56] John H McDonald and Martin Kreitman. "Adaptive protein evolution at the Adh locus in Drosophila". In: *Nature* 351.6328 (1991), p. 652.

[57] Colin D Meiklejohn, Kristi L Montooth, and David M Rand. "Positive and negative selection on the mitochondrial genome". In: *Trends in Genetics* 23.6 (2007), pp. 259–263.

[58] John H. McDonald and Martin Kreitman. "Adaptive protein evolution at the Adh locus in Drosphila". In: *Nature* 351 (1991), pp. 652–654.

[59] Adam Eyre-Walker. "Changing effective population size and the McDonald-Kreitman test". In: *Genetics* 162.4 (2002), pp. 2017–2024.

[60] Richard R Hudson, Martin Kreitman, and Montserrat Aguadé. "A test of neutral molecular evolution based on nucleotide data". In: *Genetics* 116.1 (1987), pp. 153–159.

[61] Scott H Williamson et al. "Simultaneous inference of selection and population growth from patterns of variation in the human genome". In: *Proceedings of the National Academy of Sciences* 102.22 (2005), pp. 7882–7887.

[62] Fumio Tajima. "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism." In: *Genetics* 123.3 (1989), pp. 585–595.

[63] Yun-Xin Fu and Wen-Hsiung Li. "Statistical tests of neutrality of mutations." In: *Genetics* 133.3 (1993), pp. 693–709.

[64] Justin C Fay and Chung-i Wu. "Hitchhiking Under Positive Darwinian Selection". In: (2000).

[65] Katherine M Siewert and Benjamin Franklin Voight. "Detecting Long-term Balancing Selection using Allele Frequency Correlation". In: *bioRxiv* (2017), p. 112870.

[66] Guillaume Achaz. "Frequency spectrum neutrality tests: one for all and all for one". In: *Genetics* 183.1 (2009), pp. 249–258.

[67] R. C. Lewontin and Jesse Krakauer. "DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY O F THE SELECTIVE NEUTRALITY OF and LEWONTIN". In: *Genetics* (1973), pp. 175–195.

[68] Kent E Holsinger and Bruce S Weir. "Genetics in geographically structured populations: defining, estimating and interpreting F ST". In: *Nature Reviews Genetics* 10.9 (2009), p. 639.

[69] Sewall Wright. "The genetical structure of populations". In: *Annals of Human Genetics* 15.1 (1949), pp. 323–354.

[70] Mark A Beaumont and David J Balding. "Identifying adaptive genetic divergence among populations from genome scans". In: *Molecular ecology* 13.4 (2004), pp. 969–980.

[71] Xin Yi et al. "Sequencing of 50 human exomes reveals adaptation to high altitude". In: *Science* 329.5987 (2010), pp. 75–78.

[72] Graham Coop et al. "Using environmental correlations to identify loci underlying local adaptation". In: *Genetics* 185.4 (2010), pp. 1411–1423.

[73] Russell Lande. "Neutral theory of quantitative genetic variance in an island model with local extinction and colonization". In: *Evolution* 46.2 (1992), pp. 381–389.

[74] Ken Spitze. "Population structure in Daphnia obtusa: quantitative genetic and allozymic variation." In: *Genetics* 135.2 (1993), pp. 367–374.

[75] Jeremy J Berg and Graham Coop. "A population genetic signal of polygenic adaptation". In: *PLoS genetics* 10.8 (2014), e1004412.

[76] Fernando Racimo, Jeremy J Berg, and Joseph K Pickrell. "Detecting polygenic adaptation in admixture graphs". In: *Genetics* (2018), genetics–300489.

[77] Pardis C Sabeti et al. "The case for selection at CCR5-Δ32". In: *PLoS biology* 3.11 (2005), e378.

[78] Benjamin F Voight et al. "A map of recent positive selection in the human genome." In: *PLoS biology* 4.3 (Mar. 2006), e72. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0040072. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1382018%5C&tool=pmcentrez%5C&rendertype=abstract.

[79] Anna Ferrer-Admetlla et al. "On detecting incomplete soft or hard selective sweeps using haplotype structure". In: *Molecular biology and evolution* 31.5 (2014), pp. 1275–1291.

[80] Yair Field et al. "Detection of human adaptation during the past 2 , 000 years". In: (2016), pp. 1–18.

[81] Nandita R Garud et al. "Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps". In: *PLoS Genet* 11.2 (2015), e1005004.

[82] Stephen M Krone and Claudia Neuhauser. "Ancestral processes with selection". In: *Theoretical population biology* 51.3 (1997), pp. 210–237.

[83] Montgomery Slatkin. "Simulating genealogies of selected alleles in a population of variable size". In: *Genetical Research* 78 (2000), pp. 49–57.

[84] Graham Coop and Robert C Griffiths. "Ancestral inference on gene trees under selection." In: *Theoretical population biology* 66.3 (Nov. 2004), pp. 219–32. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2004.06.006. URL: http://www.ncbi.nlm.nih.gov/pubmed/15465123.

[85] Robert C Griffiths and Paul Marjoram. "Ancestral inference from samples of DNA sequences with recombination". In: *Journal of Computational Biology* 3.4 (1996), pp. 479–502.

[86] Wolfgang Stephan, Thomas HE Wiehe, and Marcus W Lenz. "The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory". In: *Theoretical Population Biology* 41.2 (1992), pp. 237–254.

[87] Yuseob Kim and Wolfgang Stephan. "Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome". In: 777.February (2002), pp. 765–777.

[88] Rick Durrett and Jason Schweinsberg. "A coalescent model for the effect of advantageous mutations on the genealogy of a population". In: 115 (2005), pp. 1628–1657. DOI: 10.1016/j.spa.2005.04.009.

[89] Ha My T Vy and Yuseob Kim. "A composite-likelihood method for detecting incomplete selective sweep from population genomic data". In: *Genetics* 200.2 (2015), pp. 633–649.

[90] Hua Chen, Nick Patterson, and David Reich. "Population differentiation as a test for selective sweeps". In: *Genome research* 20.3 (2010), pp. 393–402.

[91] Lan Zhu and Carlos D Bustamante. "A composite-likelihood approach for detecting directional selection from DNA sequence data". In: *Genetics* 170.3 (2005), pp. 1411–1421.

[92] Mark A Beaumont, Wenyang Zhang, and David J Balding. "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4 (2002), pp. 2025–2035.

[93] Benjamin M Peter, Emilia Huerta-Sanchez, and Rasmus Nielsen. "Distinguishing between selective sweeps from standing variation and from a de novo mutation". In: *PLoS genetics* 8.10 (2012), e1003011.

[94] Louise Ormond et al. "Inferring the age of a fixed beneficial allele". In: *Molecular ecology* 25.1 (2016), pp. 157–169.

[95] Kimberly F McManus et al. "Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans". In: *PLoS genetics* 13.3 (2017), e1006560.

[96] Sara Sheehan and Yun S Song. "Deep learning for population genetic inference". In: *PLoS computational biology* 12.3 (2016), e1004845.

[97] Lauren Alpert Sugden et al. "Localization of adaptive variants in human genomes using averaged one-dependence estimation". In: *Nature communications* 9.1 (2018), p. 703.

[98] John Novembre et al. "Genes mirror geography within Europe". In: *Nature* 456.7218 (2008), p. 98.

[99] Alkes L Price et al. "New approaches to population stratification in genome-wide association studies". In: *Nature Reviews Genetics* 11.7 (2010), p. 459.

[100] Daniel R Schrider and Andrew D Kern. "Supervised Machine Learning for Population Genetics: A New Paradigm". In: *Trends in Genetics* (2018).

[101] Roy Ronen et al. "Learning natural selection from the site frequency spectrum". In: *Genetics* 195.1 (2013), pp. 181–193.

[102] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[103] Pavlos Pavlidis, Jeffrey D Jensen, and Wolfgang Stephan. "Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations". In: *Genetics* 185.3 (2010), pp. 907–922.

[104] Kao Lin et al. "Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics". In: *Genetics* 187.1 (2011), pp. 229–244.

[105] Peter Bühlmann and Torsten Hothorn. "Boosting algorithms: Regularization, prediction and model fitting". In: *Statistical Science* (2007), pp. 477–505.

[106] Daniel R Schrider and Andrew D Kern. "S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning." In: *PLoS genetics* 12.3 (Mar. 2016), e1005928. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005928. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4792382%5C&tool=pmcentrez%5C&rendertype=abstract.

[107] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine learning* 63.1 (2006), pp. 3–42.

[108] Benjamin K Rosenzweig et al. "Powerful methods for detecting introgressed regions from population genomic data". In: *Molecular ecology* 25.11 (2016), pp. 2387–2397.

[109] R C Griffiths and P Marjoram. "Ancestral Inference from Samples of DNA with Recombination". In: 3.4 (1996), pp. 479–502.

[110] Paul Fearnhead and Peter Donnelly. "Estimating Recombination Rates From Population Genetic Data". In: (2001).

[111] Na Li and Matthew Stephens. "Modeling Linkage Disequilbrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data". In: *Genetics* 2233.December (2003), pp. 2213–2233.

[112] Gilean a T McVean and Niall J Cardin. "Approximating the coalescent with recombination." In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360.1459 (July 2005), pp. 1387–93. ISSN: 0962-8436. DOI: 10.1098/rstb.2005.1673. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1569517%5C&tool=pmcentrez%5C&rendertype=abstract.

[113] Paul Marjoram and Jeff D Wall. "Fast "coalescent" simulation." In: *BMC genetics* 7 (Mar. 2006), p. 16. ISSN: 1471-2156. DOI: 10.1186/1471-2156-7-16. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1458357%5C&tool=pmcentrez%5C&rendertype=abstract.

[114] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), pp. 493–496. ISSN: 0028-0836. DOI: 10.1038/nature10231. URL: http://dx.doi.org/10.1038/nature10231.

[115] Matthew D Rasmussen et al. "Genome-wide inference of ancestral recombination graphs." In: *PLoS genetics* 10.5 (Jan. 2014), e1004342. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004342. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4022496%5C&tool=pmcentrez%5C&rendertype=abstract.

[116] Aaron J Stern, Peter R Wilton, and Rasmus Nielsen. "An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data". In: *PLoS genetics* 15.9 (2019), e1008384.

[117] G. A. Watterson. "Testing Selection at a Single Locus". In: *Biometrics* 38.2 (1982), pp. 323–331. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2530446.

[118] Iain Mathieson and Gil McVean. "Estimating selection coefficients in spatially structured populations from time series data of allele frequencies". In: *Genetics* 193.3 (2013), pp. 973–984.

[119] Ellen G Williamson and Montgomery Slatkin. "Using Maximum Likelihood to Estimate Population Size From Temporal Changes in Allele Frequencies". In: (1999).

[120] Norman L Kaplan, Thomas Darden, and Richard R Hudson. "The Coalescent Process in Models With Selection". In: 829.2 (1988), pp. 819–829.

[121] Melissa A Ilardo et al. "Physiological and genetic adaptations to diving in sea nomads". In: *Cell* 173.3 (2018), pp. 569–580.

[122] Ammon Corl et al. "The genetic basis of adaptation following plastic changes in coloration in a novel environment". In: *Current Biology* 28.18 (2018), pp. 2970–2977.

[123] Michael D Edge and Graham Coop. "Reconstructing the history of polygenic scores using coalescent trees". In: *Genetics* 211.1 (2019), pp. 235–262.

[124] Simon Tavaré. "Line-of-descent and genealogical processes, and their applications in population genetics models". In: *Theoretical population biology* 26.2 (1984), pp. 119–164.

[125] RC Griffiths. "Asymptotic line-of-descent distributions". In: *Journal of Mathematical Biology* 21.1 (1984), pp. 67–75.

[126] Ethan M Jewett and Noah A Rosenberg. "Theory and applications of a deterministic approximation to the coalescent model". In: *Theoretical population biology* 93 (2014), pp. 14–29.

[127] Matthias Steinrücken, Ethan M Jewett, and Yun S Song. "Spectraltdf: transition densities of diffusion processes with time-varying selection parameters, mutation rates and effective population sizes". In: *Bioinformatics* 32.5 (2015), pp. 795–797.

[128] Peter R Wilton et al. "A population phylogenetic view of mitochondrial heteroplasmy". In: *Genetics* 208.3 (2018), pp. 1261–1274.

[129] Jacob A Tennessen et al. "Evolution and functional impact of rare coding variation from deep sequencing of human exomes". In: *science* 337.6090 (2012), pp. 64–69.

[130] Andrew D Kern and Daniel R Schrider. "Discoal: flexible coalescent simulations with selection". In: *Bioinformatics* 32.24 (2016), p. 3839.

[131] Molly Przeworski. "The signature of positive selection at randomly chosen loci". In: *Genetics* 160.3 (2002), pp. 1179–1189.

[132] Montgomery Slatkin. "Simulating genealogies of selected alleles in a population of variable size". In: *Genetics Research* 78.1 (2001), pp. 49–57.

[133] Matthias Meyer et al. "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104 (2012), pp. 222–226.

[134] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), p. 43.

[135] Kay Prüfer et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia". In: *Science* 358.6363 (2017), pp. 655–658.

[136] Sara Mathieson and Iain Mathieson. "FADS1 and the Timing of Human Adaptation to Agriculture". In: *Molecular Biology and Evolution* 35.12 (Oct. 2018), pp. 2957–2970. ISSN: 0737-4038. DOI: 10.1093/molbev/msy180. eprint: http://oup.prod.sis.lan/mbe/article-pdf/35/12/2957/26997983/msy180.pdf. URL: https://dx.doi.org/10.1093/molbev/msy180.

[137] Joseph H Marcus and John Novembre. "Visualizing the geography of genetic variants". In: *Bioinformatics* 33.4 (2017), pp. 594–595.

[138] Nicholas Eriksson et al. "Web-based, participant-driven studies yield novel genetic associations for common traits". In: *PLoS genetics* 6.6 (2010), e1000993.

[139] Jiali Han et al. "A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation". In: *PLoS genetics* 4.5 (2008), e1000074.

[140] Patrick Sulem et al. "Genetic determinants of hair, eye and skin pigmentation in Europeans". In: *Nature genetics* 39.12 (2007), p. 1443.

[141] Simon Gravel et al. "Demographic history and rare allele sharing among human populations." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.29 (July 2011), pp. 11983–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1019276108. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3142009%5C&tool=pmcentrez%5C&rendertype=abstract.

[142] Richard A Sturm et al. "A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color". In: *The American Journal of Human Genetics* 82.2 (2008), pp. 424–431.

[143] Emilia Huerta-Sánchez et al. "Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA". In: *Nature* 512.7513 (2014), p. 194.

[144] Sandra Wilde et al. "Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y". In: *Proceedings of the National Academy of Sciences* 111.13 (2014), pp. 4832–4837.

[145] Tony Frudakis et al. "Sequences associated with human iris pigmentation". In: *Genetics* 165.4 (2003), pp. 2071–2083.

[146] Patrick Sulem et al. "Two newly identified genetic determinants of pigmentation in Europeans". In: *Nature genetics* 40.7 (2008), p. 835.

[147] Fan Liu et al. "Digital quantification of human eye color highlights genetic association of three new loci". In: *PLoS genetics* 6.5 (2010), e1000934.

[148] Eimear E Kenny et al. "Melanesian blond hair is caused by an amino acid change in TYRP1". In: *Science* 336.6081 (2012), pp. 554–554.

[149] Sajad Mirzaei and Yufeng Wu. "RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination". In: *Bioinformatics* 33.7 (2016), pp. 1021–1030.

[150] Jerome Kelleher et al. "Inferring the ancestry of everyone". In: *BioRxiv* (2018), p. 458067.

[151] Vladimir Shchur, Liliia Ziganurova, and Richard Durbin. "Fast and scalable genome-wide inference of local tree topologies from large number of haplotypes based on tree consistent PBWT data structure". In: *bioRxiv* (2019), p. 542035.

[152] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *BioRxiv* (2019), p. 550558.

[153] Pier Francesco Palamara et al. "High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability". In: *bioRxiv* (2018), p. 276931.

[154] Patrick K. Albers and Gil McVean. "Dating genomic variants and shared ancestry in population-scale sequencing data". In: *bioRxiv* (2018). DOI: 10.1101/416610. eprint: https://www.biorxiv.org/content/early/2018/09/13/416610.full.pdf. URL: https://www.biorxiv.org/content/early/2018/09/13/416610.

[155] Aaron J Stern et al. "Disentangling selection on genetically correlated polygenic traits using whole-genome genealogies". In: *bioRxiv* (2020).

[156] Po-Ru Loh et al. "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis". en. In: *Nat. Genet.* 47.12 (Dec. 2015), pp. 1385–1392.

[157] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. "Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data". en. In: *Am. J. Hum. Genet.* 99.1 (July 2016), pp. 139–153.

[158] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. *An Expanded View of Complex Traits: From Polygenic to Omnigenic*. 2017.

[159] Yuval B Simons et al. "A population genetic interpretation of GWAS findings for human quantitative traits". en. In: *PLoS Biol.* 16.3 (Mar. 2018), e2002985.

[160] Luke J O'Connor et al. "Extreme Polygenicity of Complex Traits Is Explained by Negative Selection". en. In: *Am. J. Hum. Genet.* 105.3 (Sept. 2019), pp. 456–476.

[161] Armin P Schoech et al. "Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection". en. In: *Nat. Commun.* 10.1 (Feb. 2019), p. 790.

[162] Jaleal S Sanjak et al. "Evidence of directional and stabilizing selection in contemporary humans". In: *Proceedings of the National Academy of Sciences* 115.1 (2018), pp. 151–156.

[163] Bruce Walsh and Michael Lynch. *Evolution and Selection of Quantitative Traits*. en. Oxford University Press, June 2018.

[164] Laura K Hayward and Guy Sella. *Polygenic adaptation after a sudden change in environment*.

[165] Jeremy J Berg and Graham Coop. "A population genetic signal of polygenic adaptation". en. In: *PLoS Genet.* 10.8 (Aug. 2014), e1004412.

[166] Fernando Racimo, Jeremy J Berg, and Joseph K Pickrell. "Detecting Polygenic Adaptation in Admixture Graphs". en. In: *Genetics* 208.4 (Apr. 2018), pp. 1565–1584.

[167] Yair Field et al. "Detection of human adaptation during the past 2000 years". en. In: *Science* 354.6313 (Nov. 2016), pp. 760–764.

[168] Lawrence H Uricchio et al. "An evolutionary compass for detecting signals of polygenic selection and mutational bias". en. In: *Evol Lett* 3.1 (Feb. 2019), pp. 69–79.

[169] R C Griffiths and P Marjoram. "Ancestral inference from samples of DNA sequences with recombination". en. In: *J. Comput. Biol.* 3.4 (1996), pp. 479–502.

[170] Matthew D Rasmussen et al. "Genome-wide inference of ancestral recombination graphs". en. In: *PLoS Genet.* 10.5 (May 2014), e1004342.

[171] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". en. In: *Nat. Genet.* 51.9 (Sept. 2019), pp. 1321–1329.

[172] Michael D Edge and Graham Coop. "Reconstructing the History of Polygenic Scores Using Coalescent Trees". en. In: *Genetics* 211.1 (Jan. 2019), pp. 235–262.

[173] Michael C Turchin et al. "Evidence of widespread selection on standing variation in Europe at height-associated SNPs". en. In: *Nat. Genet.* 44.9 (Sept. 2012), pp. 1015–1019.

[174] Matthew R Robinson et al. "Population genetic differentiation of height and body mass index across Europe". en. In: *Nat. Genet.* 47.11 (Nov. 2015), pp. 1357–1362.

[175] J J Berg, X Zhang, and G Coop. "Polygenic adaptation has impacted multiple anthropometric traits". In: *BioRxiv* (2017).

[176] Luis-Miguel Chevin, Sylvain Billiard, and Frederic Hospital. "Population and evolutionary genetics-Hitchhiking both ways: Effect of two interfering selective sweeps on linked neutral variation". In: *Genetics* 180.1 (2008), p. 301.

[177] Brendan Bulik-Sullivan et al. "An atlas of genetic correlations across human diseases and traits". en. In: *Nat. Genet.* 47.11 (Nov. 2015), pp. 1236–1241.

[178] Brendan K Bulik-Sullivan et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". en. In: *Nat. Genet.* 47.3 (Mar. 2015), pp. 291–295.

[179] Jeremy J Berg et al. "Reduced signal for polygenic adaptation of height in UK Biobank". en. In: *Elife* 8 (Mar. 2019).

[180] Mashaal Sohail et al. "Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies". en. In: *Elife* 8 (Mar. 2019).

[181] Kyoko Watanabe et al. "A global overview of pleiotropy and genetic architecture in complex traits". en. In: *Nat. Genet.* 51.9 (Sept. 2019), pp. 1339–1348.

[182] Po-Ru Loh et al. "Mixed-model association for biobank-scale datasets". en. In: *Nat. Genet.* 50.7 (July 2018), pp. 906–908.

[183] Claire Churchhouse et al. "Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK biobank". In: *Neale Lab* (2017).

[184] Margaux L A Hujoel et al. "Combining case-control status and family history of disease increases association power". en. Dec. 2019.

[185] Tomaz Berisa and Joseph K Pickrell. "Approximately independent linkage disequilibrium blocks in human populations". en. In: *Bioinformatics* 32.2 (Jan. 2016), pp. 283–285.

[186] Nayanah Siva. "1000 Genomes project". en. In: *Nat. Biotechnol.* 26.3 (Mar. 2008), p. 256.

[187] N Sinnott-Armstrong et al. "Genetics of 38 blood and urine biomarkers in the UK Biobank". In: *BioRxiv* (2019).

[188] Kevin R Thornton. "Polygenic Adaptation to an Environmental Shift: Temporal Dynamics of Variation Under Gaussian Stabilizing Selection and Additive Effects on a Single Trait". en. In: *Genetics* 213.4 (Dec. 2019), pp. 1513–1530.

[189] Joseph K Pickrell et al. "Detection and interpretation of shared genetic influences on 42 human traits". en. In: *Nat. Genet.* 48.7 (July 2016), pp. 709–717.

[190] Andy Dahl et al. "A Robust Method Uncovers Significant Context-Specific Heritability in Diverse Complex Traits". en. In: *Am. J. Hum. Genet.* 106.1 (Jan. 2020), pp. 71–91.

[191] Scott H Williamson et al. "Localizing recent adaptive evolution in the human genome". en. In: *PLoS Genet.* 3.6 (June 2007), e90.

[192] Luigi Luca Cavalli-Sforza et al. *The History and Geography of Human Genes*. en. Princeton University Press, 1994.

[193] Peter Frost. "European hair and eye color: A case of frequency-dependent sexual selection?" In: *Evol. Hum. Behav.* 27.2 (Mar. 2006), pp. 85–103.

[194] Zhaohui Yang et al. "Darwinian Positive Selection on the Pleiotropic Effects of KITLG Explain Skin Pigmentation and Winter Temperature Adaptation in Eurasians". en. In: *Mol. Biol. Evol.* 35.9 (Sept. 2018), pp. 2272–2283.

[195] Richard C Lewontin. *Race and Intelligence*. 1970.

[196] Arthur Jensen. "How much can we boost IQ and scholastic achievement". In: *Harv. Educ. Rev.* 39.1 (1969), pp. 1–123.

[197] Rosa Cheesman et al. "Comparison of Adopted and Nonadopted Individuals Reveals Gene-Environment Interplay for Education in the UK Biobank". en. In: *Psychol. Sci.* (Apr. 2020), p. 956797620904450.

[198] J V Neel. "Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"?" en. In: *Am. J. Hum. Genet.* 14 (Dec. 1962), pp. 353–362.

[199] Timothy J Lyons and Arpita Basu. "Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers". en. In: *Transl. Res.* 159.4 (Apr. 2012), pp. 303–312.

[200] Julie K Bower et al. "Glycated hemoglobin and risk of hypertension in the atherosclerosis risk in communities study". en. In: *Diabetes Care* 35.5 (May 2012), pp. 1031–1037.

[201] E S Lander and N J Schork. "Genetic dissection of complex traits". en. In: *Science* 265.5181 (Sept. 1994), pp. 2037–2048.

[202] B Devlin and K Roeder. "Genomic control for association studies". en. In: *Biometrics* 55.4 (Dec. 1999), pp. 997–1004.

[203] J K Pritchard et al. "Association mapping in structured populations". en. In: *Am. J. Hum. Genet.* 67.1 (July 2000), pp. 170–181.

[204] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". en. In: *Nat. Genet.* 38.8 (Aug. 2006), pp. 904–909.

[205] Matthew L Freedman et al. "Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 103.38 (Sept. 2006), pp. 14068–14073.

[206] Gregory M Marcus et al. "European ancestry as a risk factor for atrial fibrillation in African Americans". en. In: *Circulation* 122.20 (Nov. 2010), pp. 2009–2015.

[207] Lauren A Wise et al. "African ancestry and genetic risk for uterine leiomyomata". en. In: *Am. J. Epidemiol.* 176.12 (Dec. 2012), pp. 1159–1168.

[208] Yakir A Reshef et al. "Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk". en. In: *Nat. Genet.* 50.10 (Oct. 2018), pp. 1483–1493.

[209] Xuanyao Liu et al. "Quantification of genetic components of population differentiation in UK Biobank traits reveals signals of polygenic selection". en. June 2018.

Appendix A

Supplementary Materials to Ch. 2

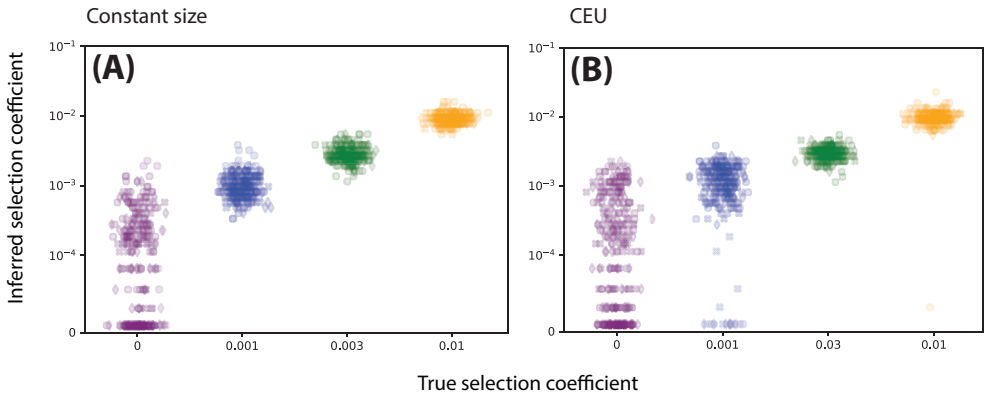


Figure A.1: **Selection coefficients inferred directly from the true local trees.** Left: constant population size ($N_e = 10^4$). Right: Tennessen CEU demographic model. Shape of marker denotes the terminal frequency conditioned upon in the simulation: ∘, 25%; ◇, 50%; ×, 75%.

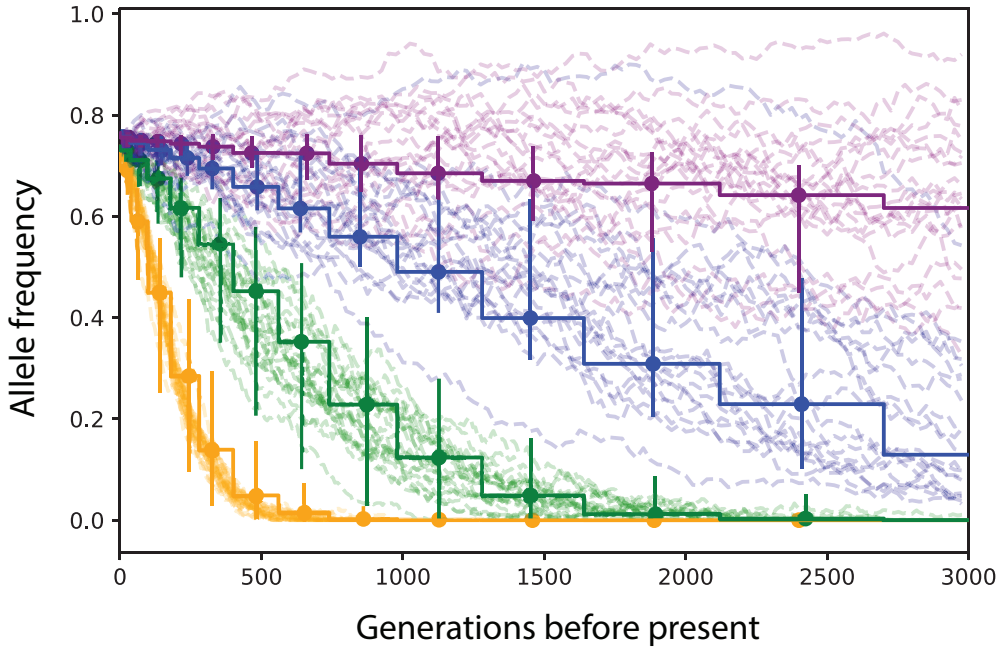


Figure A.2: **Allele frequency trajectories inferred from ARGweaver local trees when $\mu = r$.** We set $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 2.5 \times 10^{-8}$ recombinations/bp/gen and fix the present day allele frequency to $X_0 = 50\%$ Stepwise trajectories are inferred, dashed trajectories are the ground truth. Vertical bars denote the 25-75th percentile range of estimates. For each condition we show 20 randomly selected simulations and their corresponding inferences. All data simulated under a demographic model with constant size $N = 10^4$.

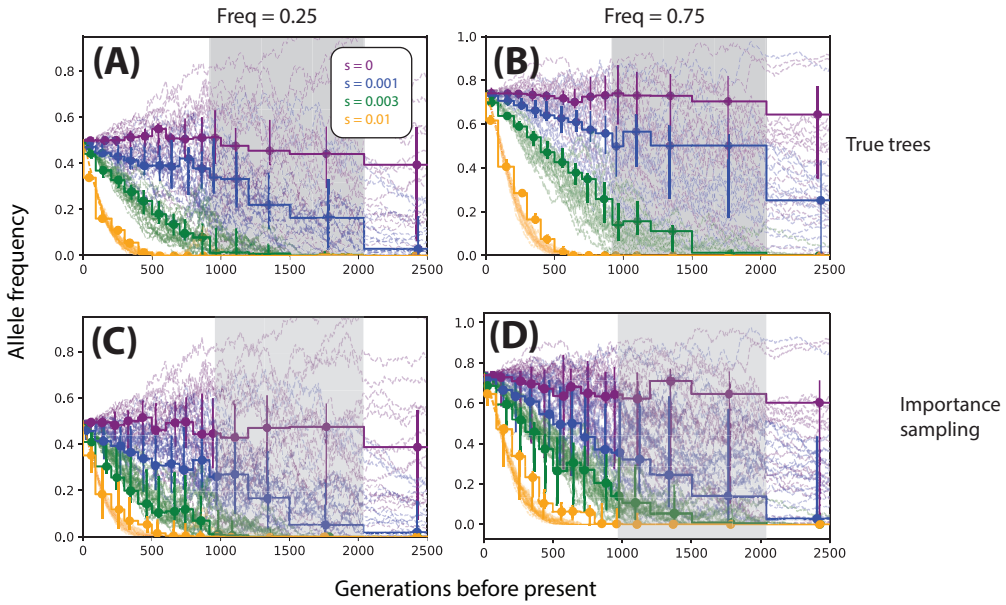


Figure A.3: **Inferring allele frequency trajectories under CEU demography.** Trajectories were inferred from true local trees (top row) and importance sampling on ARGweaver local trees (bottom row). Columns correspond to different present-day allele frequencies (left: 50%, right: 75%). For each condition we show 20 randomly selected simulations (dashed, translucent lines) and their corresponding inferences (piecewise constant curves; dots and vertical bars indicate the median and 25-75 percentiles of estimates, respectively). The gray box indicates the timing of the bottleneck, occurring approximately 920-2040 generations ago. Simulations were done under the European demographic model described in Methods and Materials using a locus of 200kb, $n = 25$ diploid individuals and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ recombinations/bp/gen.

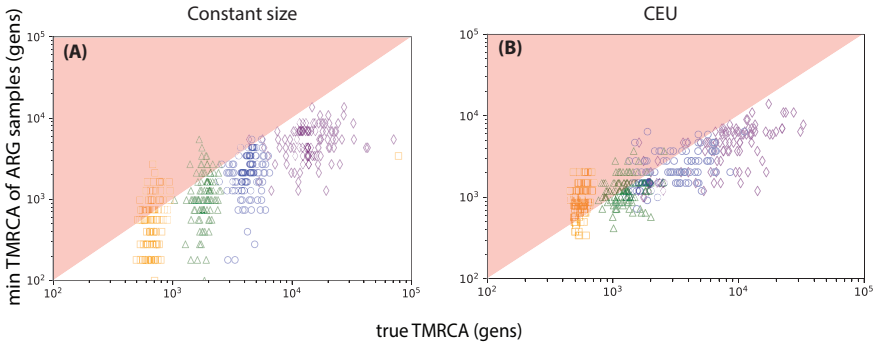


Figure A.4: **ARGweaver proposes less accurate trees under non-equilibrium demography.** Left: constant $N_e = 10^4$. Right: Tennessen CEU demographic model. We found in Fig. Fig. A.2 that importance sampling using ARGweaver tends to underestimate the selection coefficient under a model of CEU demography. To demonstrate that the proposal distribution for sampling the local tree is the source of this bias, we use TMRCA of the local tree as a heuristic the locus's selection coefficient. For the sake of argument, we postulate that as one decreases the TMRCA of a local tree, the maximum-likelihood estimate of the selection coefficient strictly descreases. If so, then if the minimum value of the sampled TMRCAs is greater than the true TMRCA, then this instance of the importance sampling estimate will underestimate the selection coefficient. Hence, one can measure importance sampling efficiency by looking at the probability that the minimum value of the sampled TMRCAs is less than the true TMRCA. This is shown graphically by the proportion of points that fall in the red upper triangle. The selection coefficients $s = 0, 0.001, 0.003, 0.01$ are indicated by purple, blue, green, and orange, respectively. Simulations were done under the European demographic model described in Methods and Materials using a locus of 200kb, $n = 25$ diploid individuals and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ recombinations/bp/gen. We fixed the present-day allele frequency to be 75%.

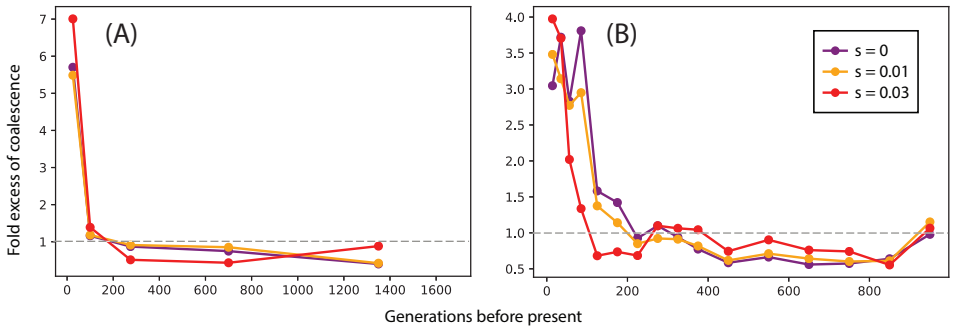


Figure A.5: **ARGweaver infers an excess of recent coalescences.** As a diagnostic for the trees outputted by ARGweaver, we compared the amount of coalescence in the sample trees vs. the local trees. Let $a_{i,s}$ be the number of coalescent events during epoch i in the sth replicate of the simulation. We calculate $e_i = \frac{\frac{1}{M}\sum_{s=1}^{S}\sum_{m=1}^{M} a_{i,s}^{(m)}}{\sum_{s=1}^{S} a_{i,s}}$ as an estimate of the fold excess of coalescence, where $a_{i,s}^{(m)}$ denotes the mth ARGweaver sample of $a_{i,s}$. Notice that if the sample trees closely approximate the true tree, then $e_i \approx 0$. The dashed line indicates no excess, i.e., no bias in the estimates. We find that ARGweaver can have a nearly 4× excess of inferred coalescence events in the most recent epochs (e.g. [0,100] generations ago). Here we simulated recent selection starting 100 generations ago under a model of European demography with $n = 50$ diploid individuals, a physical length of 200kb, and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ mut/bp/gen. We condition on the variant having a present-day frequency of 0.50.

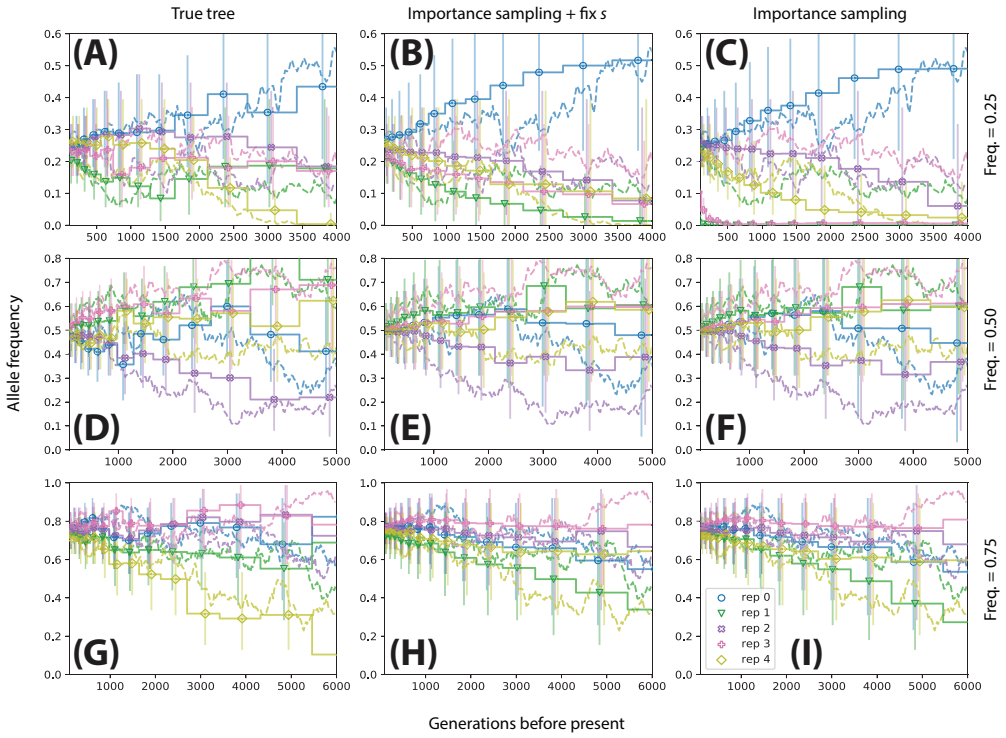


Figure A.6: **Performance of trajectory inference across replicates.** We illustrated inferred vs. true trajectories, holding selection coefficient constant across the replicates. In this case, we chose $s = 0$ in order to maximize the amount of variability in the trajectories of the replicates. Rows correspond to simulations conditioned on different present-day allele frequencies (0.25: A-C,0.50: D-F,0.75: G-I). Left column (A,D,G): trajectories inferred using the true local tree. Middle column (B,E,H): trajectories inferred using importance sampling, fixing the selection coefficient to the ground truth ($s = 0$). Right column (C,F,I): trajectories inferred under importance sampling and estimating s. Simulations were done under the constant size model described in Methods and Materials using a locus of 100kb, $n = 25$ diploid individuals and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ recombinations/bp/gen.

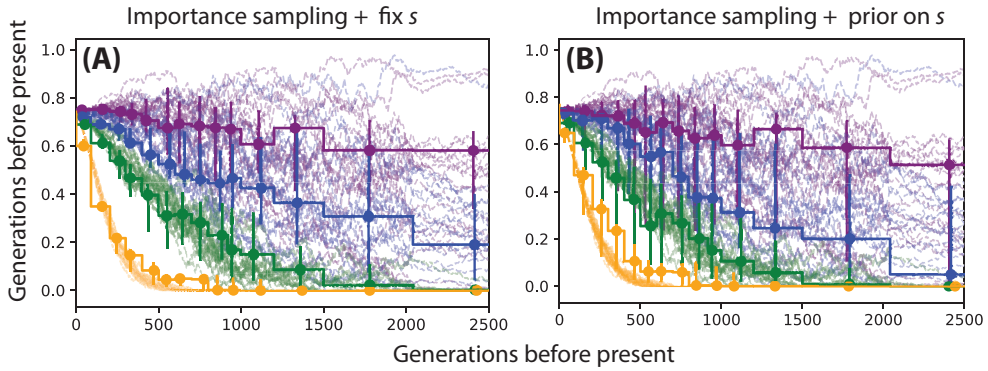


Figure A.7: **Effect of uncertainty in s on trajectory inference.** A: trajectories inferred using importance sampling (i.e. Eq. 2.34), fixing $\hat{s} = s$. B: trajectories inferred using importance sampling and integrating over a uniform prior on s (i.e. Eq. 2.37). Simulations were done under the European demographic model described in Methods and Materials using a locus of 200kb, $n = 25$ diploid individuals and $\mu = 2.5 \times 10^{-8}$ mut/bp/gen, $r = 1.25 \times 10^{-8}$ recombination-s/bp/gen.

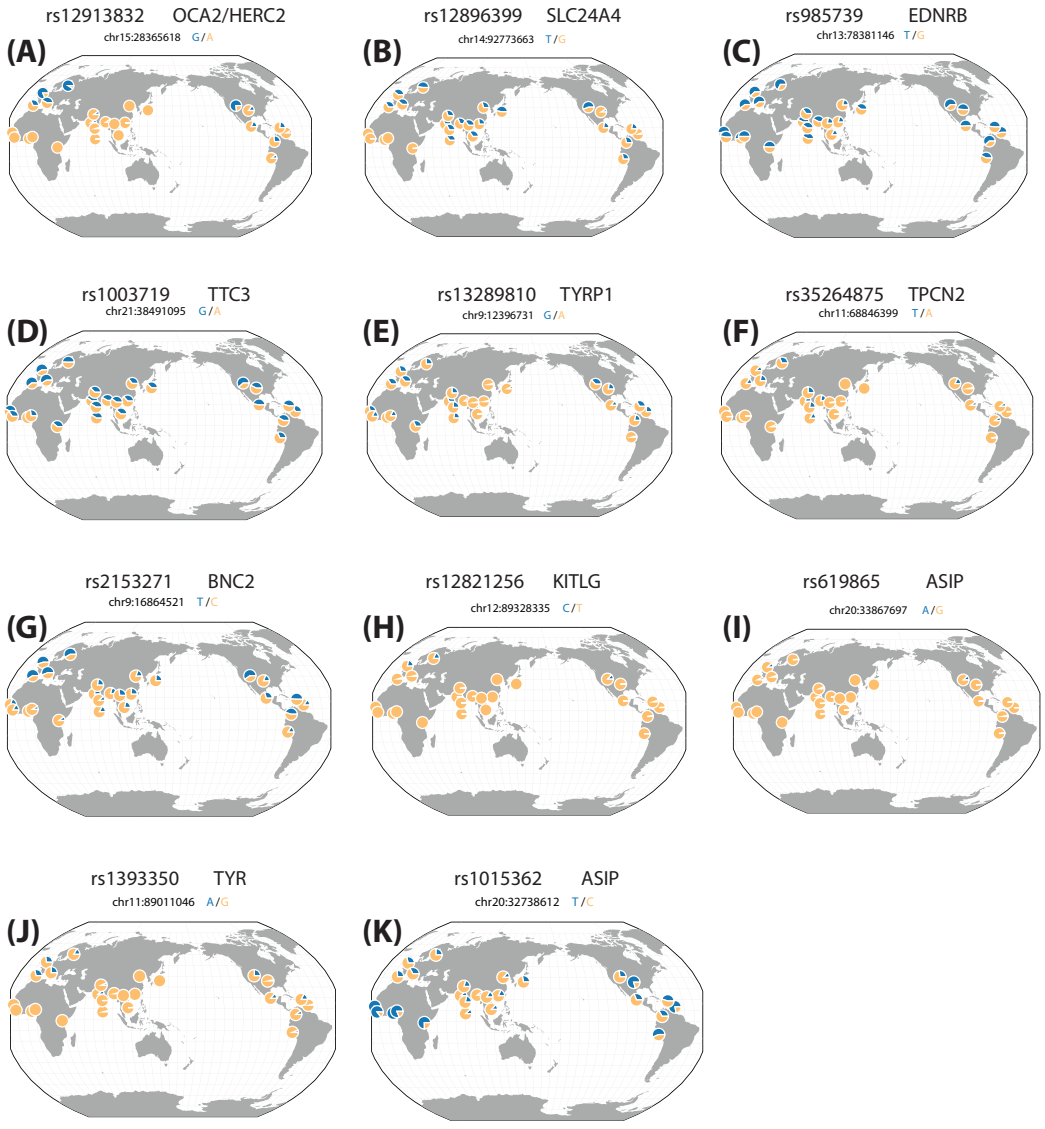


Figure A.8: **Geographical distribution of pigmentation SNPs.** Population-wide allele frequencies of pigmentation SNPs from Fig. 2.9 plotted geographically using GGV.

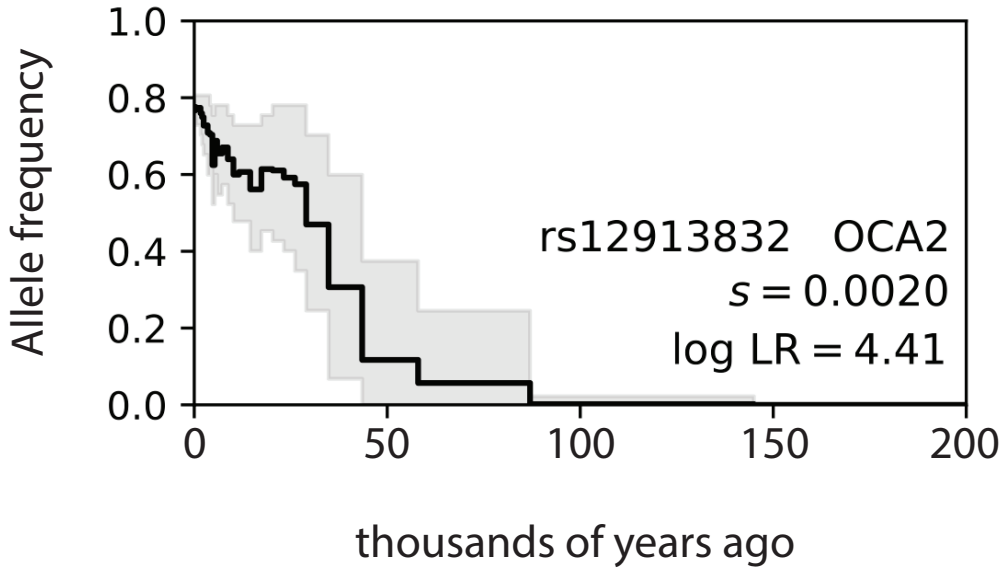


Figure A.9: **Allele frequency trajectory estimate of rs12913832 (OCA2/HERC2).** The same trajectory estimate as in Fig. 2.9A with x-axis limits extended to illustrate earlier history of the allele.

Appendix B

Supplementary materials to Ch. 3

B.1 Inference of selection gradient

Importance sampling estiamtion of the likelihood function of selection

Our likelihood model builds heavily on our previous work, which developed importance sampling approaches to estimating the likelihood function of the selection coefficient acting on a SNP, $L^{SNP}(s)$ [116]. Here, we briefly explain the importance sampling method used to estimate $L(\omega)$, the likelihood of the multivariate selection gradient:

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}^T \omega\right) \tag{B.1}$$

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

$$L^{SNP}(s) = E_p\left[P(D \mid G, s)\right] = E_q\left[P(D \mid G, s)\frac{p(G \mid s)}{q(G)}\right] \tag{B.2}$$

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

$$\hat{L}^{SNP}(s) = \frac{1}{m}\sum_{l=1}^{m} E_q\left[P(D \mid G, s)\frac{P(G \mid s)}{q(G)}\right] \tag{B.3}$$

Where the convergence is almost surely as $m \to \infty$.

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

$$\widehat{LR}^{SNP}(s) = \frac{1}{m}\sum_{l=1}^{m}\frac{p\left(G_i{}^{(l)} \mid s\right)}{p\left(G_i{}^{(l)} \mid s = 0\right)} \tag{B.4}$$

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

$$\widehat{LR}^{SNP}(s) = \sum_{l=1}^{m}\frac{p(G_i{}^{(l)}|s)}{\pi(G_i{}^{(l)})}\Big/\sum_{l=1}^{m}\frac{p(G_i{}^{(l)}|s=0)}{\pi(G_i{}^{(l)})} \tag{B.5}$$

where $\pi(\cdot)$ is a neutral prior on coalescence trees. While $p(\cdot)$ is calculated using the structured coalescent, with lineages subtending the same allele with frequency $X(t)$ coalescing at rate $\lambda(t) = N(0)/[N(t)X(t)]$, the prior $\pi(\cdot)$ is calculated using the unstructured coalescent with rate $\lambda(t) = N(0)/N(t)$. Note that we do not explicitly model population structure (e.g. gene flow). We also note that we have made several additional modifications to the importance sampling approximation of the likelihood ratio: first, we assume that the allele frequency trajectory is a deterministic, logistic function of time, when previously we modeled stochasticity in the allele frequency trajectory (see the next section for more details). Because we focus on applying our method to detecting adaptation in the recent past, this approximation is appropriate when drift has had little opportunity to distort allele frequencies. Second, we make a functional approximation to $\log\widehat{LR}^{SNP}(s)$. We do a grid search for the optimal value of s^*, and then we fit a quadratic function to points $(s, \log\widehat{LR}^{SNP}(s)) : \quad |s - s^*| < \delta$. Optimizing $\log\widehat{LR}(\omega)$ then becomes a simple process of solving a linear system of equations:

$$\log\widehat{LR}(\omega) = \sum_i \left(a_i\left(\beta_{(i)}^T\omega\right)^2 + b_i\left(\beta_{(i)}^T\omega\right) + c_i\right) \tag{B.6}$$

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

$$\hat{\omega} = \left[2\sum_i a_i\beta_{(i)}\beta_{(i)}^T\right]^{-1}\left(\sum_i b_i\beta_{(i)}\right). \tag{B.7}$$

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

Accounting for multiple SNPs in LD

In our analyses we assume independence of local LD blocks (see e.g. Berisa & Pickrell, 2016). Generally we choose to ascertain a single SNP for each LD block and include its SNP likelihood in the product (Eq. B1). However, in joint analyses it may be necessary to ascertain multiple SNPs per LD block, each corresponding to a GWAS hit for a different trait. Let $B(i)$ denote the set of ascertained SNPs in the same LD block as i. If only 1 SNP from each LD block is included, then $B(i) = 1$ for each ascertained SNP i. If multiple SNPs from the same LD block are included, we exponentiate each of these SNPs' likelihoods by a factor $1/|B_i|$:

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}^T \omega\right)^{1/|B_i|} \tag{B.8}$$

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

$$L(\omega) = \prod_{i=1}^{M} L_i^{SNP}\left(\beta_{(i)}^T \omega\right)^{1/|B_i|} \tag{B.9}$$

$$= K(\omega) \prod_{i \in S:\ i \neq i_1, i_2} L_i^{SNP}\left(\beta_{(i)}^T \omega\right) \tag{B.10}$$

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

Selection gradient and correlated response standard errors

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

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

$$R = \omega^{(T \cup j)} - \omega^{(j)} \tag{B.11}$$

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

B.2 Coalescent likelihood models

Relate prior

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

$$\pi(T) = \prod_{i=1}^{n-1} p(U_i = u_i \mid U_{i-1} = u_{i-1}) \tag{B.12}$$

$$p(U_i = u_i \mid U_{i-1} = u_{i-1}) = \frac{n-i+1}{2} \cdot N(0)/N(u_i) \cdot \exp(-\frac{n-i+1}{2}(\Lambda(u_i) - \Lambda(u_{i-1}))) \tag{B.13}$$

$$\Lambda(u) = \int_0^u N(0)/N(t)\ \cdot dtN(0)/N(t)\ \cdot dt \tag{B.14}$$

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

$$\Lambda_i = \sum_{k=1}^{b(u_i)} N_0\tau_k/N_k + N_0(u_i - \tau_{b(u_i)})/N_{b(u_i)} \tag{B.15}$$

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

Coalescent selection likelihood under deterministic model

Unlike in our previous work [116], in which we treated the allele frequency as a latent random variable, here we use a deterministic approximation of the allele frequency trajectory. Under the standard 'hard sweep' model, an appropriate approximation would be $X(t \mid s) = (1 + (1 - x_0)/x_0 \cdot e^{st})^{-1}$. Technically, if we want to express the trajectory conditional on the present-day derived allele frequency (DAF) x_0, it would be more appropriate to use a closer approximation of the backwards Wright-Fisher diffusion with selection. However, since we are mostly interested in modeling the recent past for common alleles ascertained in a GWAS (usually DAF > 1%), this approximation is appropriate, especially in populations of large recent N_e such as humans, where drift is negligible on short timescales.

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

$$X(t, s, x_0) = x_0, \qquad t < a \tag{B.16}$$

$$= (1 + (1 - x_0)/x_0 \cdot e^{s(t-a)})^{-1}, \qquad a \leq t < b \tag{B.17}$$

$$= (1 + (1 - x_0)/x_0 \cdot e^{sb})^{-1,} \qquad t > b \tag{B.18}$$

To calculate $p(T \mid s)$, we split the tree into two subtrees (imagine 'deleting' the branch on which the mutant allele arose). Note that we implicitly assume the site is biallelic, such as under the infinite sites assumption. Let us label these alleles A1 and A2; these labels must be consistent with the polarization of the GWAS summary statistics; we assume that those are polarized w.r.t the A1 allele. Within each of these subtrees, we find the coalescent times U^{A1} and U^{A2}. Then

$$p(T \mid s) = \prod_{i=1}^{n_1-1} p(U_{n-i} = u_i^{A1} \mid U_{n-i+1} = u_{i-1}^{A1}, s, x_0) \tag{B.19}$$

$$\times \prod_{i=1}^{n_2-1} p(U_{n-i} = u_i^{A2} \mid U_{n-i+1} = u_{i-1}^{A2}, -s, 1 - x_0) \tag{B.20}$$

$$p(U_{k-1} = t \mid U_k = t', s, f) = \frac{k}{2} \cdot \frac{N(0)}{N(t)X(t)} \cdot \exp(-\frac{k}{2}(\Lambda(t, s, f) - \Lambda(t', s, f))) \tag{B.21}$$

$$\Lambda(t, s, f) = N(0)/[N(\tau)X(\tau, s, f)] \cdot d\tau \tag{B.22}$$

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

B.3 Supplementary Figures

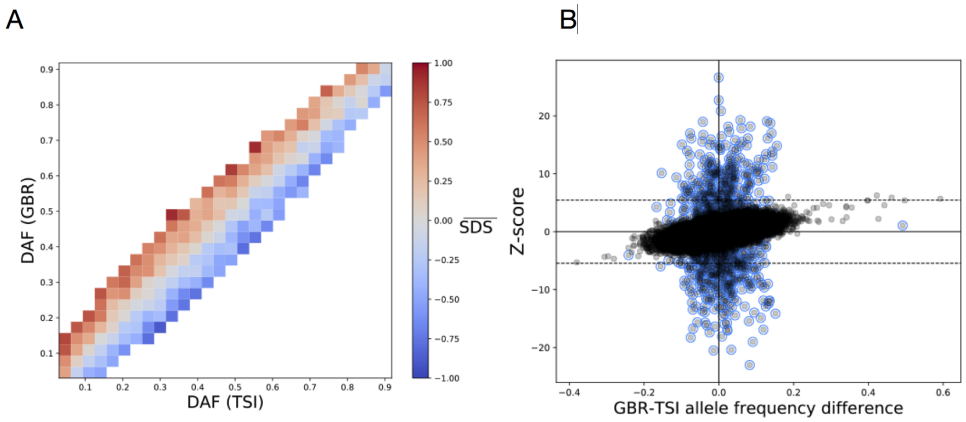


Figure B.1: **Distribution of frequencies and SDS in 1000 Genomes SNP set.** (A) mean SDS score with respect to joint derived allele frequency (DAF) in TSI vs GBR. We selected a set of 40,320 autosomal SNPs with MAF >0.5% in the UK Biobank and $MAC_{GBR+TSI} \geq 4$, HWE $P > 10^{-6}$ in the 1000 Genomes Phase 3 data, which we used in our simulations of uncorrected stratification. We found that this SNP set recapitulates the pattern demonstrated in Sohail, et al. (2019); namely, that SNPs with higher frequency in GBR tend to have higher SDS, and vise versa for TSI. DAF was calculated from 1000 Genomes phase 3 data for all autosomes and SDS was obtained from previous analysis of the UK10K cohort. To limit noise, we show DAF bins with ≥30 SNPs. (B) An example simulation of uncorrected stratification in Z-scores of the aforementioned SNP set. Here we set $h^2 = 50\%$, $M = 10^3$, $N = 10^5$, $N_{TSI}/N_{GBR} = 5\%$, $\sigma_S = 0.1$. Causal SNPs are circled in blue. Dashed lines indicate genome-wide significance thresholds ($P < 5 \times 10^{-8}$).

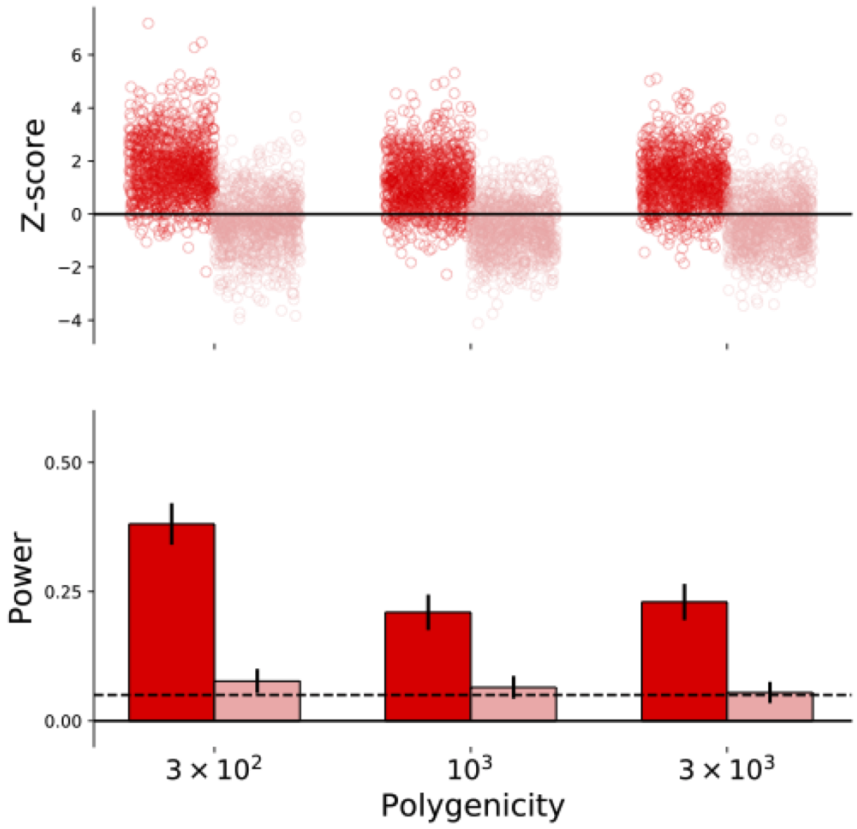


Figure B.2: **Calibration and power under GBR demography** We simulated polygenic adaptation under a model of changing population size based on GBR 1000 Genomes individuals with $\omega = 0.05$, and a pulse of recent selection over the last 35 generations. Red denotes simulations with selection, pink denotes neutral simulations. Dashed lines indicate nominal FPR (5%) and black lines denote 95% Bonferroni-corrected CIs. We simulate a trait with $h^2 = 50\%$ and $N = 10^5$. To estimate trees we used a sample size of $n = 400$ haplotypes of length 1Mb and assume mutation and recombination rates of $\mu = r = 10^{-8}$/bp/gen. We simulate 1000 replicates in each case.

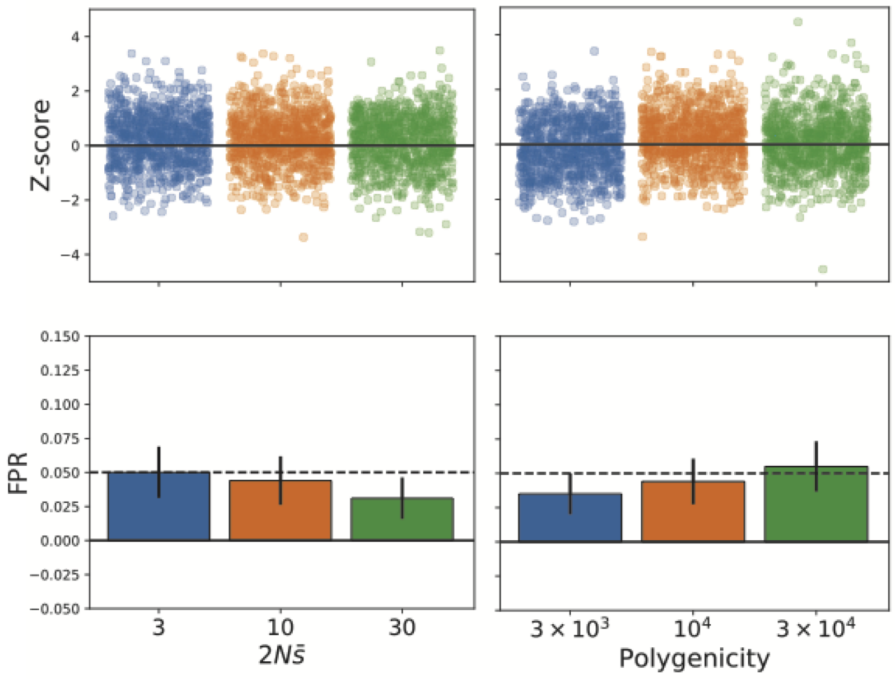


Figure B.3: **Robustness to purifying selection.** We simulated traits under purifying selection using the model of Schoech, et al. (2019). That is, we model the joint distribution of allele frequency and selection coefficient of a causal SNP. We assume that SNPs are always deleterious and the magnitude of the selection coefficient has an exponential distribution, with mean s. Given s for a particular SNP, the allele frequency is drawn randomly from its stationary distribution. We assume that $|\beta| = c \cdot s^{1/2}$ – the constant of proportionality is chosen post-hoc to normalize SNP heritabillity to 50% – so our results also approximate for the dynamics of Gaussian stabilizing selection, modulo underdominance and epistatic effects. Dashed lines indicate nominal FPR (5%) and black lines denote 95% Bonferroni-corrected CIs. We simulate GWAS with $N = 10^5$. To estimate trees we used a sample size of $n = 400$ haplotypes of length 1Mb and assume mutation and recombination rates of $\mu = r = 10^{-8}$/bp/gen. We simulate 1000 replicates in each case.

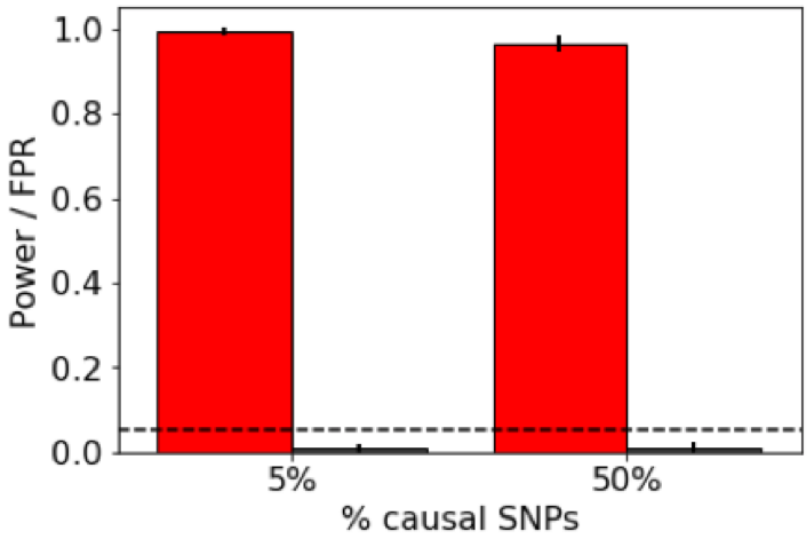


Figure B.4: **Calibration and power under allelic heterogeneity** We simulated polygenic adaptation of a trait with multiple linked causal SNPs in LD (i.e., allelic heterogeneity) in SLiM. We assume $h^2 = 50\%$ and a mutational target of $100 \times 100kb$ regions with $\mu = r = 10^{-8}$. We assume a constant population size of $N_e = 10^3$ and burn in simulations for 9990 generations, and then for the next 10 generations, we simulate a directional selection gradient of $\omega = 0.2$ (we chose this value because it resulted in ~1SD increase in the mean phenotype). We simulated two levels of allelic heterogeneity; (left) 5% and (right) 50% of mutations in the target are causal. Red bar indicate power to detect simulations with selection; pink bars (which are too short to see this color) indicate FPR under the null, $\omega = 0$. SNPs were ascertained by taking one at each independent regions with the maximum value of $2pq\beta^2$ within the region. We tested for selection using the true local trees at these sites. Dashed lines indicate nominal FPR (5%) and black lines denote 95% Bonferroni-corrected CIs. We simulated 1000 replicates in each case.

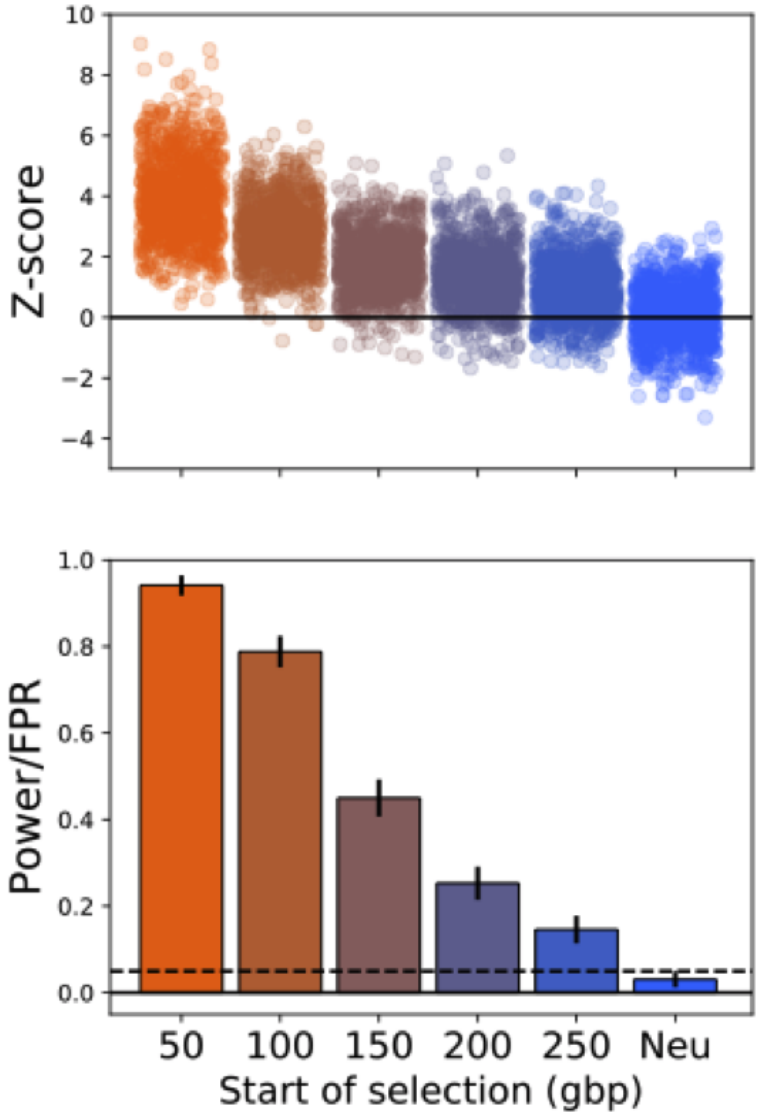


Figure B.5: **Time specificity of test for recent selection** We simulated under models of different timings of selection, assuming a 50-generation pulse with start time ranging from 50 to 250 generations ago, as well as neutral simulations, and run PALM under a nominal model of selection in the last 50 generations. Dashed lines indicate nominal FPR (5%) and black lines denote 95% Bonferroni-corrected CIs. We simulate GWAS with $N = 10^5$. To estimate trees we used a sample size of $n = 400$ haplotypes of length 1Mb and assume mutation and recombination rates of $\mu = r = 10^{-8}$/bp/gen. We simulate 1000 replicates in each case.

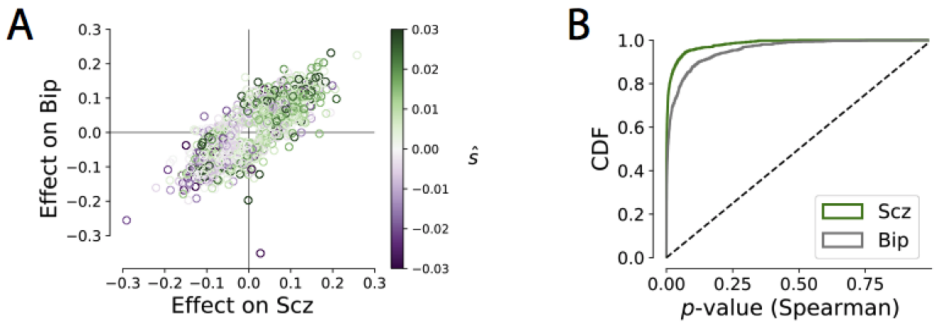


Figure B.6: **Pleiotropy causes bias in tests for polygenic adaptation** We simulated a bivariate trait (modeled after h^2 and r_g of schizophrenia [Scz], and bipolar disorder [Bip]). We simulated a pulse of selection acting to increase Scz prevalence over the last 50 generations. (A) Estimates of directional selection on SNPs ($\hat{s}$) are positively correlated with both SNP effects for Scz and Bip. We estimated selection using our importance sampling method and included SNPs ascertained in our simulated GWAS. (B) Testing for selection on a neutral correlated trait (here, Bip) yields massive inflation of the false positive rate. We evaluated p-values by computing significance of the Spearman correlation of $\hat{\beta}$ and $\hat{s}$.

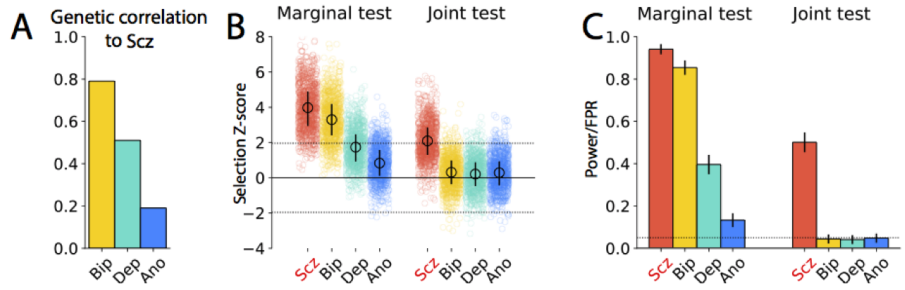


Figure B.7: **Marginal vs. joint test comparison, lower pleiotropy** Here we recreate Figure 3 (main text) but setting the level of pleiotropy used to simulate the polygenic trait architectures to be lower (ϱ = 60%). See Fig. 4.3 for all simulation details.

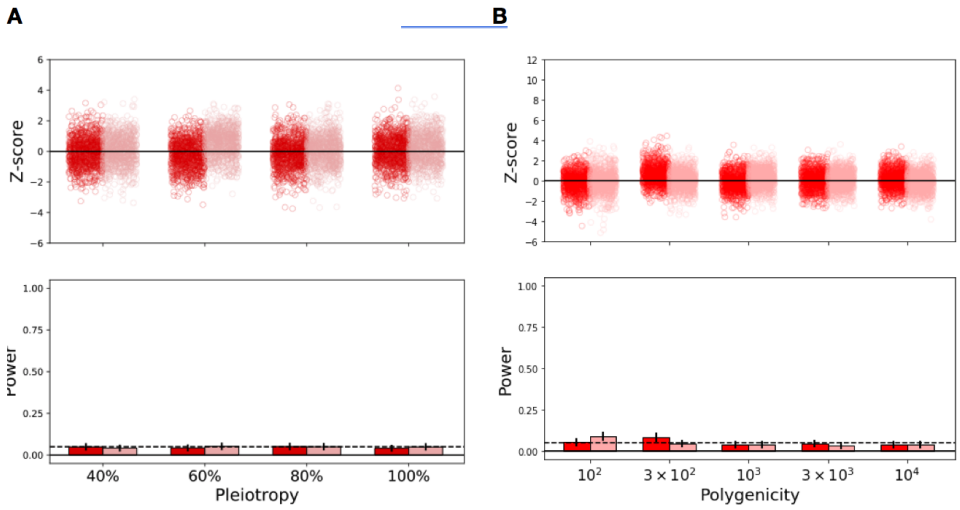


Figure B.8: **Calibration of joint test** Simulations under neutrality, testing Trait I and Trait III jointly. Here we show calibration of J-PALM when neither of the two traits is under selection (i.e., both are neutral). We consider the effects of degree of pleiotropy (A) and polygenicity (B). Other simulation details are identical to Fig. 4.3.

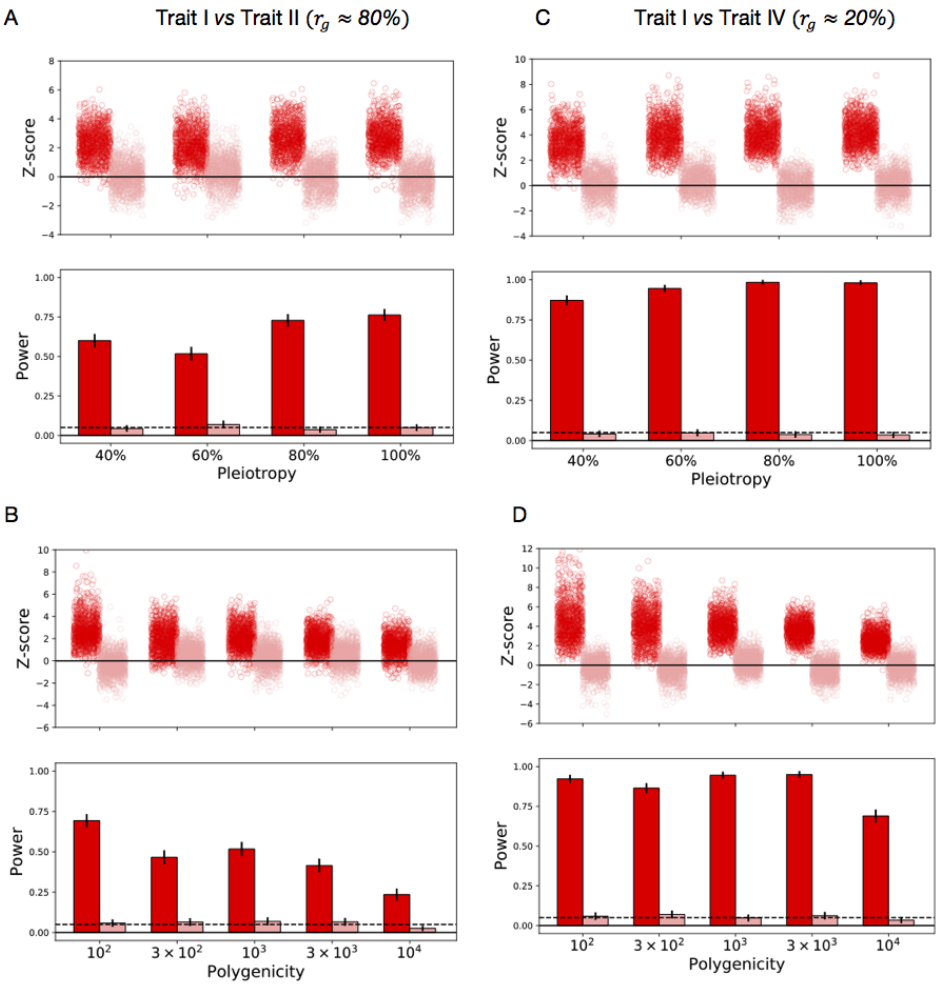


Figure B.9: **Joint test power and calibration for other trait pairs** Simulations under neutrality, using J-PALM to test Trait I vs Trait II (A,B) and Trait I vs Trait IV (C,D) jointly. We varied degree of pleiotropy ϱ (A,C) and polygenicity M (B,D). In all simulations we simulated selection to increase Trait I, with all other traits neutral. Other simulation details are identical to Fig. 4.3.

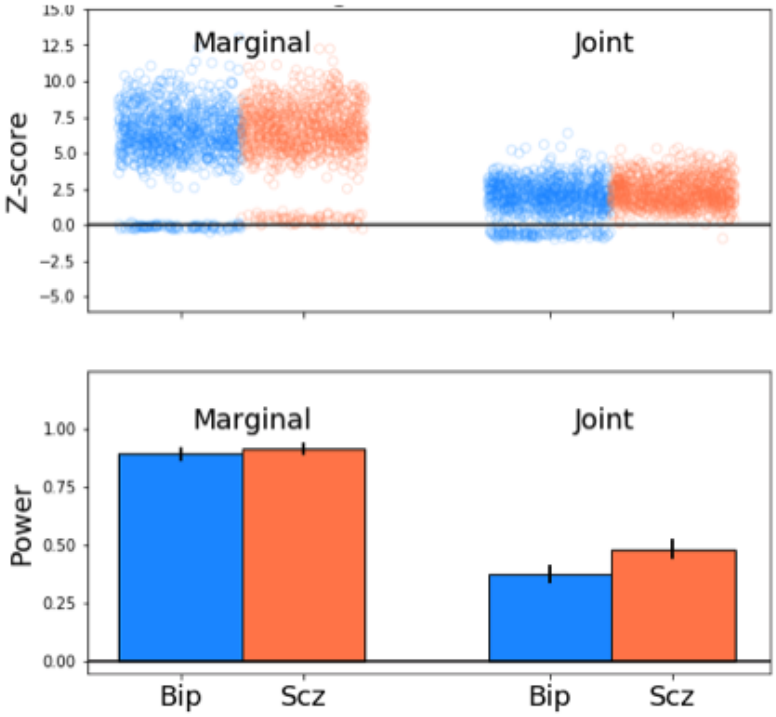


Figure B.10: **Joint estimates under complementary selection** Positive selection (ω = 0.05) simulated on both of two traits, Trait I & II, modeled after SNP heritability and genetic correlation of bipolar (Bip) and schizophrenia (Scz). Selection was estimated marginally (left) and jointly (right). Simulations follow the approach demonstrated in Fig. 4.3.

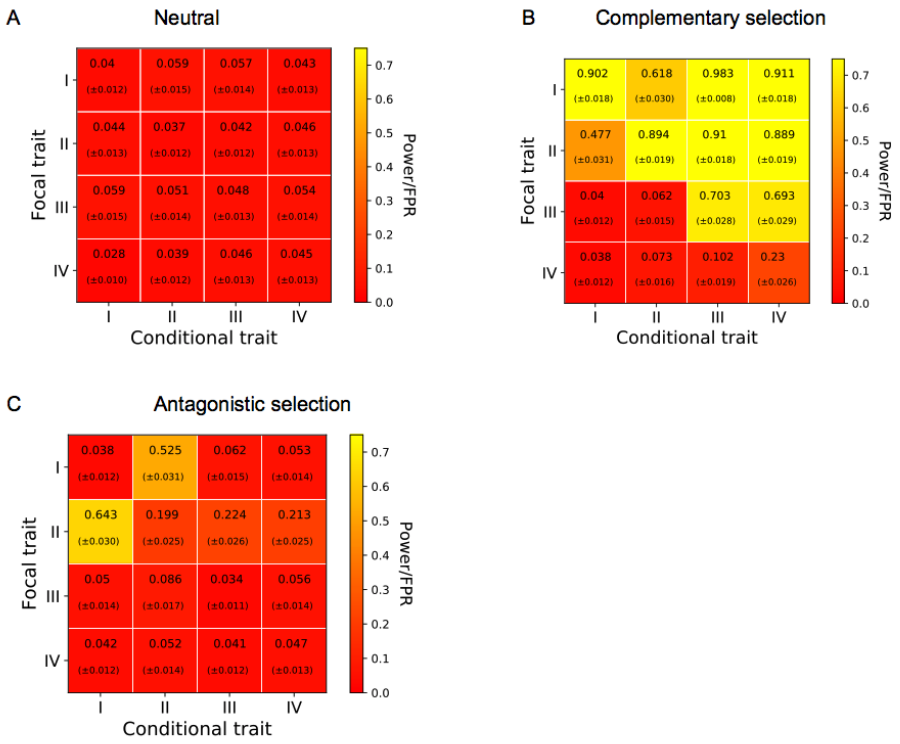


Figure B.11: **Joint test, including/excluding the causal trait** We conducted joint tests over all pairs of Traits I/II/III/IV under 3 scenarios: (A) neutral (all traits neutral), (B) Traits I & II under positive selection, (C) Trait I under positive selection, Trait II under negative selection. Text and color show positive rate. Parentheses indicate standard errors.

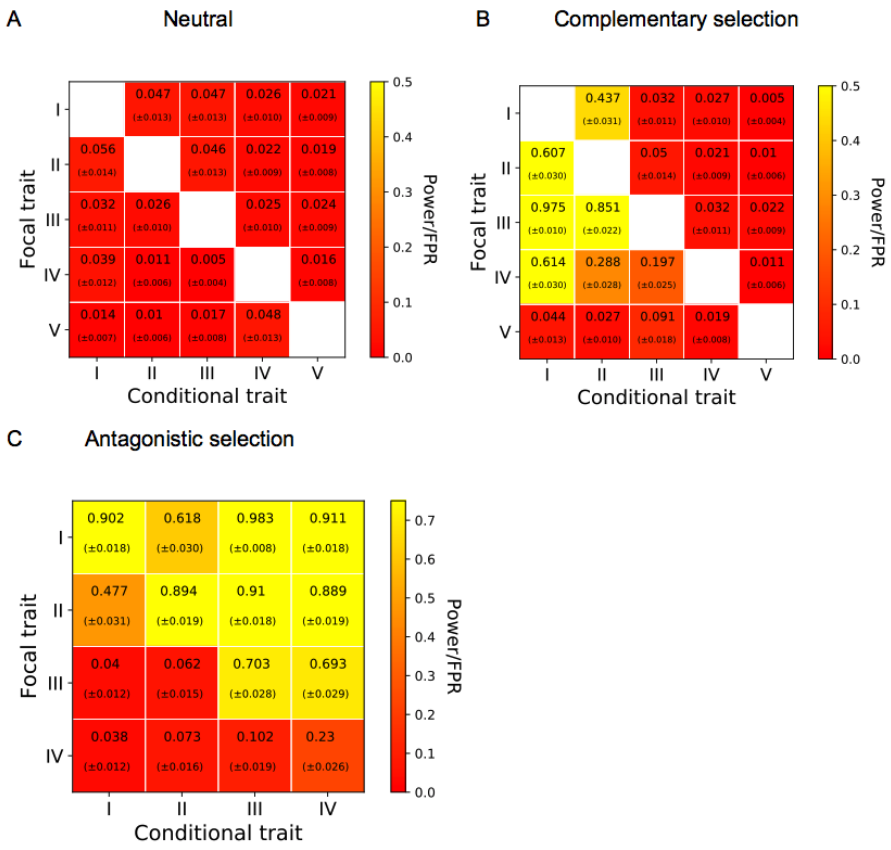


Figure B.12: **Correlated response test, including/excluding the causal trait** We conducted tests for correlated response over all pairs of Traits I/II/III/IV under 3 simulation scenarios: (A) neutral (all traits neutral), (B) Traits I & II under positive selection, (C) Trait I under positive selection, Trait II under negative selection. Text and color show positive rate. Parentheses indicate standard errors.

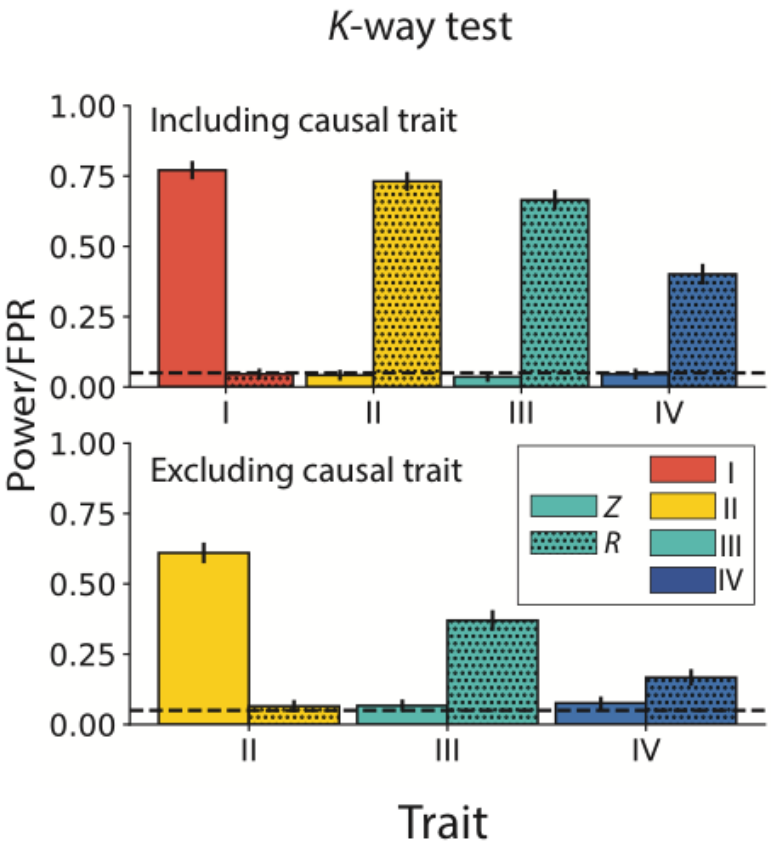


Figure B.13: **K-way tests for selection and correlated response.** Top: including all traits. Bottom: excluding trait I. Solid bars: test positive rate for selection test. Dotted bars: test positive rate for correlated response test. Trait I is under positive selection, with all other traits neutral.