Statistical Inference Using Identity-by-Descent Segments: Perspectives on Recent Positive Selection

Seth David Temple

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Sharon R. Browning, Chair
Elizabeth A. Thompson
Kelley Harris

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical Inference Using Identity-by-Descent Segments:
Perspectives on Recent Positive Selection

Seth David Temple

Chair of the Supervisory Committee:
Sharon R. Browning
Department of Biostatistics

Positive selection is suggested to be the primary mechanism of phenotypic adaptation. Selective sweeps are one model of positive selection in which beneficial mutations increase in frequency. Many existing methods to detect positive selection do not adjust for multiple hypothesis tests. Additionally, many approaches to estimate the selection coefficient, a parameter that influences the rate of allele frequency change, lack uncertainty quantification.

Here we develop theory and methodology to study recent positive selection with genetic data from the present day. Our methods use long identity-by-descent segments which should be unusually abundant in strong and recent selective sweeps. In our first project, we prove that the rate of detectable identity-by-descent segments around a locus is normally distributed for large sample size and large scaled population size. In our second project, we propose an estimator of the selection coefficient, with confidence intervals, that is an easy-to-interpret one-to-one non-decreasing function of the identity-by-descent rate. Furthermore, we provide methods to analyze selective sweeps regardless of whether the selected allele is known or genotyped. In our third project, we derive a multiple testing correction to control family-wise error rate when scanning for excess identity-by-descent rates. We apply our suite of methods to detect and model selective sweeps in European, African, and South Asian human populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

In general, upper case variables denote random variables and lower case variables denote observations of random variables. **Bold** face font is used to denote vectors, sums of vectors, or averages of vectors. The main nomenclature used is listed below.

- t : coalescent time in generations from present-day
- T_k : coalescent time until any pair of k haplotypes reaches a common ancestor
- $T^+_{n:k}$: coalescent time from present-day to the $(n-k+1)^{\text{th}}$ coalescent event
- n : sample size
- N : population size
- $N(t)$: population size at time t
- s : selection coefficient
- $p(t)$: allele frequency at time t
- R_a : sample haplotype a's recombination endpoint to the right of a focal point
- L_a : sample haplotype a's recombination endpoint to the left of a focal point
- W_a : sample haplotype a's segment overlapping a focal point
- $R_{a,b}$: the IBD segment of a and b to the right of a focal point
- $L_{a.b}$: the IBD segment of a and b to the left of a focal point
- $W_{a,b}$: the IBD segment of a and b overlapping a focal point
- w : Morgan length threshold
- $X_{a,b}$: the binary indicator that $R_{a,b} \geq w$
- $Y_{a,b}$: the binary indicator that $W_{a,b} \geq w$
- $Z_{a,b}$: mean-centered $X_{a,b}$
- $\tilde{Z}_{a,b}$: mean-centered $Y_{a,b}$

The following list includes common abbreviations for genetic terms:

- AFR: African ancestry
- ARG: ancestral recombination graph
- EHH: extended haplotype homozygosity
- EUR: European ancestry
- GWAS: genome-wide association study
- IBD: identity-by-descent
- LD: linkage disequilibrium

The following list includes abbreviations for the demographic models we study:

- BN: population bottleneck
- C##: constant population of size ##
- G3: three phases of exponential growth

The following list includes common abbreviations for statistical terms:

- FDR: false discovery rate
- FWER: family-wise error rate
- MLE: maximum likelihood estimate/estimator/estimation

ACKNOWLEDGMENTS

I greatly appreciate the advice and support these last five years given to me by some of the best and brightest in the fields of statistics and population genetics, in particular my research advisors Dr. Sharon Browning and Dr. Elizabeth Thompson. Their guidance has been essential to my growth as a rigorous, practical, and creative researcher and statistician. I am also grateful to my other committee members (Dr. Kelley Harris and Dr. Amy Willis), other mentors and collaborators (Dr. Ellen Wijsman, Dr. Elizabeth Blue, Dr. Timothy Thornton, Dr. Erick Matsen, Dr. Kimberly Kaufeld, Dr. Hugh Haddox, Dr. Morgan Gorris, and Dr. Christopher Sinclair), and my academic advisors (Dr. Daniela Witten and Ellen Reynolds). Additionally, I am thankful to the members of the Browning Lab, especially Dr. Ryan Waples and Ruoyi Cai whose discussions and help have been important to my research development, and Nicholas Irons and Vydhourie Thiyageswaran, whose discussion of and reference to the Chandrasekhar et al. [33] paper have been important to the main result of this thesis.

Throughout my time at University of Washinton, I have been fortunate to receive funding from several sources, including the Department of Statistics, Department of Biostatistics, the NIA Alzheimer's Disease Sequencing Project, the Browning Lab IBD grant (No. NHGRI HG005701), a Pre-Doctoral Training Grant in Statistical Genetics (No. NIGMS T32GM081062), and the National Defense Science and Engineering Graduate Fellowship. I am also grateful to the NHLBI Trans-Omics for Precision Medicine Project (TOPMed) and the United Kingdom British Biobank (UKBB) for providing access to the data analyzed in this dissertation. In particular, I would like to thank the investigators, staff, and participants of these studies for their contributions and for making my work possible. Any opinions, find-

ings, ideas, interpretations, conclusions, or recommendations contained in this dissertation are mine and do not necessarily reflect the views of these studies or investigators.

In the first year of my program, I was challenged to learn remotely during pandemic lockdowns. I couldn't have made it through the degree program without the support and friendship of my amazing PhD cohort and other students, in particular Anupreet Porwal, Dr. Antonio Olivas, Leah Andrews, and Nina Galanter. I am indebted to David Frank and Andy Dang for their companionship and for letting me rent their basement unit during those times. I also appreciate my lifelong friends from Oregon (Go Ducks), especially Colin Lipps, for their support. And last but not least, I am thankful to Karl Anderson for spell-checking the captions of the more than seventy figures in this thesis, listening to my practice talks, cooking me tasty vegetarian meals, showing me around Bainbridge Island and the Olympic Peninsula, and keeping me smiling.

Finally, I cannot put into words how thankful I am for the unconditional love and support of my parents Dave and Joy and my brother Matt these past twenty-plus years. We're a very close-knit family. They have made a home for me to grow and flourish in as a student, professional, and overall person. Their kindness, hospitality, and work ethic are traits I hope to learn from and model. While they didn't teach me to do math, they did teach me to windsurf, ski, and play cards, among other things, which is much more important to my sanity than math is. I am also grateful to my extended Miller and Temple families, whose regular reunions give me a sense of connection and belonging. Love you all!

DEDICATION

To my family, especially my parents, for a lifetime of love and support

Chapter 1

INTRODUCTION

1.1 Theories of molecular evolution

Natural selection is one of the oldest and most discussed topics in population genetics. Based on observations of different phenotypes among closely related species separated by islands, Darwin [40] proposed a theory of evolution via natural selection. Many theories have been developed since to explain some of the observed genotypic and phenotypic variation within and between species and populations. Among his many contributions to the theory of molecular evolution, Kimura [86] proposes a neutral theory in which most molecular changes have no fitness effect on survival or reproduction, and hence most genetic differences can be explained by random changes in allele frequency (referred to as genetic drift). Ohta [107] instead suggests that most molecular changes are slightly deleterious, which may result in slight adjustments to the expectations of neutral theory. (See Kreitman and Akashi [90] for a review of these theories and molecular evidence for or against them in observed data.) Modern research has pivoted towards quantifying the impacts of genetic drift alongside selective forces targeted at certain regions of genomes and at certain times of a species' history [154]. Identifying genes under selection may complement genome-wide association studies, as such deviations from neutral theory are likely to follow from functionally important molecular effects [154].

There are many models for selection from the simple example of a single beneficial mutation to considerably more complex scenarios. Background selection involves most mutations being strongly deleterious, and thus this force depresses genetic diversity from what is expected under neutral theory. Conversely, positive selection is when alleles increase in frequency because they are advantageous for survival or reproduction. The distinctions between

these models of directional selection concern which allele is being studied, the magnitude of its effect, and the frequency of selection events. Balancing selection concerns multiple beneficial alleles being maintained at appreciable frequencies in a population. Two mechanisms of balancing selection are when the heterozygote has a positive fitness effect relative to homozygotes or when fitness effects depend on the relative frequencies of alleles. Crow and Kimura [38] and Felsenstein [48] provide lengthier discussion of these topics in population genetics.

Selective sweeps are a precise example of positive selection in which a single beneficial allele increases in frequency. This selection event is posited to be the primary mechanism of phenotypic adaptation [154], can have a large effect on the surrounding genome, and occurs rarely. Selective sweeps can be classified further in terms of the origins of the sweeping alleles, the current frequencies of sweeping alleles, the number of sweeping alleles at a given locus, and the fitness effects of sweeping alleles over time. In this work, we focus on mathematical modeling of this evolutionary scenario.

1.2 Mathematical modeling of selective sweeps

Vitti et al. [154] categorize methods to detect selective sweeps at a microevolutionary level into three types: population differentiation-based, frequency-based, and linkage disequilibrium-based (LD). Population differentiation-based methods scan for alleles that are abundant in one population and not in others. Frequency-based methods like Tajima's D [145] and Fay and Wu's H [46] measure if there are unusually many high-frequency alleles or a surplus of rare alleles. LD-based methods like extended haplotype homozygosity (EHH) [124] and integrated haplotype score (iHS) [155] examine the persistence in a population of unusually long haplotypes. These approaches are especially useful in studying selective sweeps that are ongoing [154]. The methods we propose in this thesis fall under LD-based methods.

In the mathematical model for a selective sweep, a model parameter, the selection coefficient, influences the rate of change of allele frequency. The sweep is referred to as hard when only one haplotype with an adaptive allele increases in frequency, whereas in a soft sweep

multiple haplotypes carrying the adaptive allele increase in frequency [57, 76, 114, 115, 77]. Many methods to detect and study selective sweeps do not have a direct mathematical connection to the hard selective sweep model. Without the connection between summary statistics to the mathematical model, it is difficult to develop hypothesis tests or estimate model parameters.

Commonly in population genetics, a statistic is proposed that intuitively captures the expected effects of positive selection on haplotypes or frequencies, and then a battery of neutral simulations are conducted to "normalize" the statistic. This approach is the case for many LD-based methods to detect selection. Voight et al. [155] say of their integrated haplotype score (iHS) method:

"The iHS statistic is constructed to provide a tool for identifying SNPs, or genomic regions, that are unusual relative to the genome as a whole, and not to provide formal significance testing relative to a theoretical model. We will show that in all populations there is an excess of extreme iHS signals relative to simulated models. However, since there is considerable uncertainty in simulated models, we prefer not to assign formal p-values to the signals that we find."

While these methods and others not mentioned have been useful in identifying putatively non-neutral regions for follow-up research, there remains opportunity to develop methods for formal hypothesis testing.

Beyond scanning for signals of positive selection, a recent trend in population genetics is to estimate the selection coefficient in the hard selective sweep model. If data on allele frequency changes is not available, which is often the case in genetic analysis, estimation becomes very challenging. We highlight two particular approaches for estimating the selection coefficient with *modern sequence data only*. One approach is to optimize a (composite) likelihood over an inferred ancestral tree [142, 152]. Another approach is to simulate data under different selection coefficients and train a neural network [75, 103, 151]. These approaches have been evaluated for the accuracy of point estimates for selection coefficients

close to zero (no selection). Neither approach provides confidence intervals for the selection coefficient. We also consider it difficult to interpret the results of these approaches or to audit their performance, given the complexity of ancestral trees and neural networks, respectively. Overall, there remains opportunity to develop selection coefficient estimators that provide some statistical guarantees.

We aim to develop a formal framework for statistical inference of recent positive selection. Our data consists of long segments of DNA shared pairwise from a common ancestor (referred to as identical-by-descent). We draw a direct connection between a model for identity-by-descent segment lengths and the hard selective sweep model. We study a simple-to-interpret test statistic: the number of identity-by-descent segments longer than some threshold. We show how to use this test statistic in hypothesis testing and selection coefficient estimation. The main questions we investigate about statistical inference are:

1. Does our hypothesis test control the family-wise error rate? [1]

2. Is our estimator unbiased, sufficient, and/or consistent? [2]

3. Is the probability that our ninety-five percent confidence interval contains the true parameter equal to ninety-five percent?

To some extent, the answer to all of these questions is, technically, "no". The primary reason why we lack clean theoretical results is due to the correlations between identity-by-descent segments. We show under some circumstances that these correlations are minimal, and thus estimators based on long identity-by-descent segments might behave as if we had independent, identically distributed data. We show in exhaustive simulations that the answer to all of these questions is "almost".

---

[1] Family-wise error rate is the probability of rejecting the null model one or more times when the null model is true. This quantity concerns the significance level in the case of multiple tests.

[2] See Casella and Berger [31] for definitions of these statistical properties. Generally speaking, unbiasedness means that the estimator is equal to the true parameter on average, sufficiency means that the estimator uses all relevant aspects of observed data, and consistency means that with high probability the estimator gets closer to the true parameter as sample size increases.

1.3 Organization and summary of chapters

The outline of this thesis is as follows. We start with theoretical contributions to population genetics, then we present some methods for the analysis of genetic data, and we finish with a data analysis of numerous populations in the Trans-Omic for Precision Medicine (TOPMed) project [146] and the United Kingdom Biobank (UKBB) [24]. In most cases, we introduce content and methods in the order that they would appear in a real data analysis. First, we scan the genome for signals of non-neutral evolution. Second, we identify a causal selected allele or we estimate the frequency and location of a sweep. Third, we estimate a selection coefficient. With minimal modifications, the content of Temple and Thompson [147] is presented in Chapters 2 and 3 and Appendix A. With minimal modifications but some reorganization, the content of Temple et al. [148] is presented in Chapters 4 and 5, and in the data analysis of Europeans in Chapter 7. Chapter 6 and the data analyses of African and South Asian ancestry samples in Chapter 7 concern ongoing research. Our preprints Temple et al. [148] and Temple and Thompson [147] are in peer-review for publication. Each chapter contains an introduction, subsections on theory or methodology, subsections on simulation studies or real data analysis, and a discussion.

In Chapter 2, we define our mathematical notation and introduce the model for identity-by-descent segment lengths overlapping a fixed location. We also give a fast algorithm to simulate identity-by-descent segment lengths overlapping a fixed location. The algorithm helps build intuition for the data generating process. Its efficient runtime is also vital in proceeding chapters where we conduct enormous simulation studies. Demographic scenarios are given as well in this chapter, which are referred back to in all of our simulation experiments.

In Chapter 3, we present and derive our main theoretical result. Identity-by-descent segment lengths are correlated via unobserved tree and recombination processes, which commonly presents challenges to the derivation of theoretical results in population genetics. Under interpretable regularity conditions, we show that the proportion of detectable identity-by-descent segments around a locus is normally distributed for large sample size and large

scaled population size. We give the schematic of the proof in the main text and precise derivations in Appendix A. We use efficient and exact simulations to study the distributional behavior of the detectable identity-by-descent rate in finite samples. In finite samples, we reject the null hypothesis of normality more often than the nominal significance level, indicating that hypothesis tests that assume identity-by-descent rates overlapping a focal location are normally distributed may be anti-conservative.

In Chapter 4, we propose a selection coefficient estimator that is an easy-to-interpret one-to-one non-decreasing function of the identity-by-descent rate overlapping a focal location. We believe that many of the statistical properties of the identity-by-descent rate overlapping a focal location may also apply to our selection coefficient estimator. We use the parametric bootstrap to make valid confidence intervals under the assumption that the selection coefficient estimator is normally distributed. We characterize the behavior of our selection coefficient estimator in extensive simulation studies. Appendix B contains additional simulation studies concerning our selection coefficient estimator.

In Chapter 5, we develop a suite of methods to analyze selective sweeps in genetic data. For instance, to use our selection coefficient estimator, we require the frequency of the sweep and accurately detected identity-by-descent segments at the selected locus. Our analysis workflow includes methods to identify possible sweeping alleles and estimate the approximate location and frequency of a sweeping allele, even if it is not genotyped. We evaluate the accuracy of our sweep frequency and selection coefficient estimators in simulated sequence data. We also compare our methods alongside state-of-the-art methods to rank candidate sweep alleles and estimate selection coefficients.

In Chapter 6, we derive a multiple testing adjusted significance threshold in genome-wide scans for excess identity-by-descent rates. Our correction for multiple testing is based on normally distributed random variables. Its motivation comes from our asymptotic theory in Chapter 3, and its anti-conservativeness in simulation studies comes from the finite-sample behavior we describe in Chapter 3. We show that our hypothesis test has high power to reject false null models when the selection coefficient is large enough to be reliably estimated

in Chapters 4 and 5. We discuss how to use this anti-conservative hypothesis test and how it compares to existing approaches to determine statistical significance.

In Chapter 7, we use the methods developed in Chapters 4, 5, and 6 to study positive selection in multiple human populations. We scan for non-neutrally evolving regions of the genome in African, European, and South Asian ancestry samples. Certain loci have excess identity-by-descent rates in all three continental ancestry groups, which may indicate balancing selection or some other evolutionary process, not a selective sweep. We model a few selective sweeps in Europeans where there is a preponderance of evidence in our data and agreement with the existing literature. This data analysis demonstrates how to examine and quantify the effects of strong and recent positive selection using our entire methodology.

In Chapter 8, we give concluding remarks about the key insights made in this dissertation. We also indicate directions for future research modeling strong and recent positive selection using identity-by-descent segments.

Chapter 2

SHARED HAPLOTYPES OVERLAPPING A FOCAL POINT

2.1 Introduction

Two individuals share a haplotype segment identical-by-descent (IBD) if they inherit it from the same common ancestor. Throughout this dissertation, we focus on the lengths of IBD segments overlapping a focal location. Ignoring gene conversion, IBD segments are randomly cut by crossover recombination in each future generation. The lengths of IBD segments are thus shorter with higher probability the more removed its common ancestor is from the present-day. We are interested in the insights that IBD segment length distributions may provide about the recent genetic history of a population.

Models for IBD segment lengths integrate over two waiting time distributions: the time until a common ancestor and the genetic length until a crossover. In this chapter, we formally define our model for IBD segment lengths overlapping a specific locus. Section 2.2 reviews two ancestral tree processes: the discrete-time Wright-Fisher model and the Kingman coalescent [87, 88]. Section 2.3 defines IBD segment lengths as random variables conditional on coalescent times. From these two intermediate models, we develop an exact and efficient algorithm to simulate long IBD segments overlapping a fixed location for large samples within a population. In Section 2.4, we offer probabilistic arguments that elaborate on our simulation algorithm's runtime. We proceed to benchmark the compute time of our simulation algorithm as the sample size increases exponentially. We make concluding remarks about the importance of the algorithm's efficiency to our simulation studies in Chapters 3 and 4.

2.2 The time until a common ancestor

Let n be the haploid sample size and $k \leq n$ be the size of a subsample. Define $N(t)$ to be the population size t generations ago. Unless otherwise specified, time $t \geq 0$ always refers to time backward from the present day. For constant population size, note that $N = N(t)$ for all t. In the discrete-time Wright-Fisher process, each haploid has a haploid ancestor in the previous generation, and, if haploids have the same haploid ancestor, their lineages join.

We now define a probability distribution on the time until a common ancestor. Let the random variable T_k denote the time until a common ancestor is reached for any two of k haploids. The random variable $T^+_{n:k} := \sum_{l=k}^{n} T_l$ is the time until $n - k + 1$ coalescent events. The time to the most recent common ancestor (TMRCA) of the sample is $T^+_{n:2}$. The probability that the time until the most recent common ancestor of two specific haploids is

$$P(T_2 = t) = \prod_{\tau=1}^{t-1} \left(1 - \frac{1}{N(\tau)}\right) \frac{1}{N(t)}, \tag{2.1}$$

where $1/N(\tau)$ is the probability that a haploid has the same haploid parent as the other haploid at generation τ. The approximate probability that the time until a common ancestor is reached for any two of k haploids is

$$P(T_k = t \,|\, T^+_{n:k+1} = t_0) = \prod_{\tau=t_0+1}^{t-1} \left(1 - \frac{\binom{k}{2}}{N(\tau)}\right) \frac{\binom{k}{2}}{N(t)} \tag{2.2}$$

when k is much smaller than $\min_t N(t)$ [74]. The geometric model assumes that multiple coalescent events in a single generation are improbable. Its rate $\binom{k}{2}/N(\tau)$ is the probability that any two of k haploids have the same haploid parent at generation τ. This model approximation can be violated in analyses of large samples, *e.g.*, some human biobanks, which we discuss in Section 2.4.1.

The n-coalescent comes from the continuous time limit of Equations 2.1 and 2.2 for constant population size N [1]. Specifically, T_k converges weakly to Exponential($\binom{k}{2}$) for $k \ll N$, $N \to \infty$, and time is scaled in units of N generations [87, 88]. For our proofs in

[1] Varying population sizes are implemented through changes to the coalescent process [74].

Section 2.4 and Chapter 3, we assume this weak convergence for all $k \leq n$. Henceforth, we consider the positive real-valued T_k in units of N generations.

2.3 The distance until crossover recombination

The genetic distance between two points is the expected number of crossovers between them in an offspring gamete. This unit of haplotype segment length is the Morgan. Assuming no interference in double-stranded breaks and that crossovers occur randomly and independently, Haldane [67] derives that the genetic distance until crossover recombination is exponentially distributed, with the Poisson process modeling the crossover points along the genome. The number of crossovers between two points is then Poisson distributed with mean equal to the genetic distance between the two points, which leads to the Haldane map function connecting Morgans to the recombination frequency[2].

From a fixed location, the Morgan distance until a crossover in one gamete offspring is distributed as Exponential(1). An important property of the exponential random variable is that the minimum of independent exponential random variables is an exponential random variable with a rate that is the sum of the rates of the independent random variables. Since meioses are independent after t meioses the haplotype segment length to the right of a focal location is distributed as Exponential(t)[3].

Figure 2.1 illustrates the coalescent and recombination processes. Let a, b, c, d be sample haplotypes in the current generation. Define $L_a, R_a \,|\, t \sim \text{Exponential}(t)$ to be sample haplotypes a's recombination endpoints to the left and right of a focal location. Since crossovers to the left and right are independent, the extant width derived from the ancestor at time t is $W_a := L_a + R_a \,|\, t \sim \text{Gamma}(2, t)$. The segment length overlapping a focal location is necessarily longer than the segment length in one direction[4]. Because recombination

[2]The Haldane map function is $\rho = 0.5(1 - \exp(-2d))$, where ρ is the recombination frequency and d is the genetic distance.

[3]We use the rate parameterization of the exponential random variable.

[4]The segments to the right and left W_a and $W_{a,b}$ are stochastically dominant to and have larger expected values and variances than those of the segments to the right R_a and $R_{a,b}$, respectively.

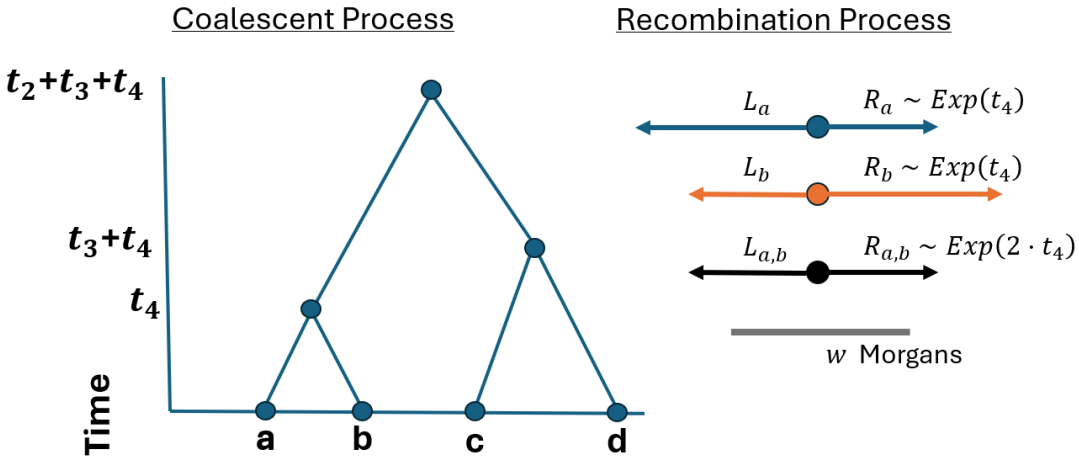

Figure 2.1: Conceptual framework for IBD segment lengths. (Left) Sample haplotypes a, b, c, d trace their lineages back to common ancestors at times $t_4, t_4+t_3, t_4+t_3+t_2$. (Right) Relative to a focal point, the haplotype segments lengths R_a, R_b, L_a, L_b are independent, identically distributed Exponential(t_4). The lengths shared IBD are $R_{a,b} := \min(R_a, R_b)$ and $L_{a,b} := \min(L_a, L_b)$. The IBD segment length $W_{a,b} := L_{a,b} + R_{a,b} \sim \text{Gamma}(2, 2 \cdot t_4)$ exceeds w Morgans, so the IBD segment indicator $Y_{a,b} = 1$.

events are independent in the t meioses descending to a and b from their common ancestor, the IBD segments that are shared by a and b are $L_{a,b}, R_{a,b}|t \sim \text{Exponential}(2t)$ and $W_{a,b} \sim \text{Gamma}(2, 2t)$. The shared segment length $R_{a,b}$ is necessarily no greater than the individual segment lengths R_a and R_b[5].

2.4 *An efficient algorithm to generate identity-by-descent segment lengths overlapping a focal point*

Simulation is incredibly useful in population genetics when analytical results are unavailable. Based on Sections 2.2 and 2.3, the blueprint to simulate IBD segment lengths around a

[5]The haplotype segment length R_a is stochastically dominant to and has larger expected value and variance than that of the IBD segment length $R_{a,b}$.

locus is as follows: 1) simulate a coalescent tree for a sample from a population, 2) draw recombination endpoints to the left and right of a focal point at each coalescent event, and 3) derive from the recombination endpoints the haplotype segment lengths that are shared IBD. The third step involves calculating the minimum lengths to the right and left of a focal point for every pair of haplotypes. Compute times in simulating IBD segment lengths should thus scale quadratically as the sample size grows. In large populations, the occurrence of an IBD segment longer than a couple of centiMorgans (cM) is rare, and consequently large sample sizes are necessary to observe IBD segments around a specific locus.

In Algorithm 1, we formally state the method to simulate long IBD segments around a single locus. We make four modifications to the naive simulation algorithm that are designed to reduce compute times when the primary goal is the generate IBD segments longer than some detection threshold. These implementations are supposed to reduce compute times due to the mathematical properties of the coalescent time and recombination endpoint distributions, not because of clever computational tricks. The intuition is that the majority of recombination endpoint comparisons happen at the oldest coalescent events (Appendix C) and that IBD segments descendant from ancestors at the oldest coalescent events are unlikely to be long.

First, in the discrete-time Wright-Fisher model, we approximate the sampling of haploid parents as a binomial random variable whenever there is likely to be more than one coalescent event in a generation (Section 2.4.1). Second, we exchange the Kingman coalescent for the discrete-time Wright-Fisher model once the number of non-coalesced haploids is much smaller than the population sizes. This implementation is similar to the hybrid simulation approach in Bhaskar et al. [14]. Third, "pruning" is when a sample haplotype is not considered for IBD segment calculation at future coalescent events once its haplotype segment length is less than the specified detection threshold. In Section 2.4.2, we elaborate on the rare probability of long haplotype segments in large populations. Fourth, "merging" is when two sample haplotypes are combined for IBD segment calculation at future coalescent events if they share the same left and right recombination endpoints. In Section 2.4.3, we derive results

concerning the probability of merging.

2.4.1 Approximating the Wright-Fisher process in large samples

Simulating the Kingman coalescent is much faster than simulating the discrete-time Wright-Fisher process. The accuracy of the Kingman coalescent requires that the sample size is much smaller than the population size. This requirement is so that there being more than one coalescent event in a generation is improbable. It can be violated in analyses of human biobanks. Under this violation, the coalescent approximation can deviate significantly from the exact discrete-time Wright-Fisher model [14, 109, 156].

In the following approximations for the sampling of haploid parents at each generation, we suppress the dependence on the generation time t. Let $k := k(t-1)$ be the number of lineages at generation $t-1$. Let $k' := k(t)$ and $N' := N(t)$ be the number of lineages and the population size in the previous generation t. The probability that a parent among $\{1, \ldots, N'\}$ has no children is $(1 - 1/N')^k$. The probability that a parent has at least one child is $1 - (1 - 1/N')^k$. The Taylor series expansion in $1/N'$ about zero is

$$1 - \left(1 - k/N' + \frac{k(k-1)}{2N'^2} - \frac{k(k-1)(k-2)}{6N'^3} \pm \ldots\right). \tag{2.3}$$

The second order approximation $k/N' - \binom{k}{2} \times N'^{-2}$ is accurate if $k^3 = o(N'^3)$. The expected number of parents in the previous generation t with a child in generation $t-1$ is then

$$\mathbb{E}[k'] \approx N'\left(k/N' - \binom{k}{2} \times N'^{-2}\right) = k - \binom{k}{2} \times N'^{-1}. \tag{2.4}$$

As an example, consider a sample of twenty thousand haploids whose ancestral population sizes in the recent ten generations are more than two hundred thousand haploids. The second order approximation is accurate for the first ten generations because $k^3 \cdot N^{-3} = 10^{-3}$ when $k = 2 \cdot 10^4$ and $N = 2 \cdot 10^5$. For this choice of k and N', the expected number of coalescent events is approximately five hundred.

Compared to drawing a parent for each child and then scanning a vector of size k for siblings, simulating the number of coalescent events in one generation from Binomial($\binom{k}{2}, N'^{-1}$)

Algorithm 1 Efficient simulation of IBD segment lengths

Input: sample size n, population sizes $N(t)$, Morgan threshold w

Output: Detectable IBD segment lengths $\ell_{a,b} \geq w$ for $a, b \in \{1, \ldots, n\}$ and the coalescent times of their common ancestors

1. Initialize recombination endpoints $l_a, r_a = \infty$ for all $a \in \{1, \ldots, n\}$, $k = n$, $t = 1$
2. **Simulate a coalescent tree**

   while $k > 2$

   (a) **if not** $k^3 \ll N(t)^3$ **then**

      i. Draw from Binomial($\binom{k}{2}, N(t)^{-1}$) (or Poisson in the limit)

      ii. Choose $a, b \in \{1, \ldots, n\}$ to coalesce

      iii. Iterate k down by 1

   (b) **else**

      i. Draw a common ancestor from $\{1, \ldots, N(t)\}$ for each $a \in \{1, \ldots, n\}$

      ii. **if** a, b have the same common ancestor, they coalesce, and iterate k down

   (c) Iterate t up by 1
3. **Simulate recombination endpoints**

   Initialize $\tau = 1$

   while $\tau <= t$

   (a) **for** coalescent event at time τ

      i. For each sample j under the subtree, draw $l'_j, r'_j \sim$ Exponential(v_j).

      if w_j is the time of its last interior node, **then** $v_j = \tau - w_j$

      - Update $l_j = \min(l_j, l'_j), r_j = \min(r_j, r'_j)$, and $w_j = \tau$
      - **if** $l_j + r_j < c$, ignore all future updates for j

      ii. For each pair i, j, **if** $l_i = l_j$ and $r_i = r_j$, **then** merge nodes together

   (b) Iterate τ up by 1

   (**Option:** Use the Kingman coalescent if $k \ll N(t)$ for all remaining t.)

can be an efficient approximation. The last term being subtracted in Equation 2.4 is equal to the expected value of a Binomial random variable of $\binom{k}{2}$ trials with success probability N'^{-1}. Next, let A_1 and A_2 be the number of children from two haploid parents in the previous generation. If A_1 and A_2 are independent, then $P(A_1 = a_1, A_2 = a_2) = P(A_1 = a_1) \times P(A_2 = a_2)$. A_1 and A_2 are not independent because a haploid child can only have one haploid parent, but the difference between the left and right terms can be close when N' is large.

$$\begin{aligned}
&P(A_1 = a_1, A_2 = a_2) - P(A_1 = a_1) \times P(A_2 = a_2) \\
&= \binom{k}{a_1} \frac{1}{N'} \cdots \frac{1}{N' - a_1} \times \binom{k - a_1}{a_2} \frac{1}{N' - a_1 - 1} \cdots \frac{1}{N' - a_1 - a_2} \\
&- \binom{k}{a_1} \frac{1}{N'} \cdots \frac{1}{N' - a_1} \times \binom{k}{a_2} \frac{1}{N'} \cdots \frac{1}{N' - a_2} \\
&\leq \binom{k}{a_1} \frac{1}{N'} \cdots \frac{1}{N' - a_1} \times \binom{k}{a_2} \frac{1}{N' - a_1 - 1} \cdots \frac{1}{N' - a_1 - a_2}.
\end{aligned} \tag{2.5}$$

Equation 2.5 is bounded by $O(k^{a_1+a_2} \cdot N'^{-(a_1+a_2)})$. If both A_1 and A_2 have finite two or more children, then Equation 2.5 is $o(1)$ when the second order approximation is accurate.

In Algorithm 1, we assume that all simultaneous coalescent events are the result of only two children having the same parent. Bhaskar et al. [14] have shown that the majority of simultaneous coalescent events in a generation are of this type.

2.4.2 The probability of detectable haplotype segment lengths

Within tens of generations, most haplotype segment lengths are shrunk by crossovers to measure less than detection thresholds that are used in IBD-based analyses[6]. The probabilities of a detectable haplotype segment to the right of and overlapping a focal location, R_a and W_a, respectively, conditional on coalescent time Nt (in generations), are

$$1 - F_{R_a|t}(w) = \exp(-Ntw), \tag{2.6}$$

$$1 - F_{W_a|t}(w) = \exp(-Ntw) + Ntw \cdot \exp(-Ntw). \tag{2.7}$$

[6] A Morgan length threshold at least greater than 0.01 is typical in applied research [19, 20, 26, 148].

Figure S1 shows that the survival functions of R_a and W_a are decreasing exponentially over Nt generations. The probabilities of haplotype segment lengths greater than 0.01 can be far from zero when the haplotype is descendant from an ancestor within the last one hundred generations. The probabilities of haplotype segments lengths greater than 0.02 are nearly zero when they are descendant from an ancestor more than three hundred generations ago[7].

For large populations, the coalescent times of ancestral lineages can be much greater than 500 generations. The expected time of the $(n-k+1)^{\text{th}}$ coalescent event can be derived with a telescoping sum argument:

$$\begin{aligned}\mathbb{E}[T^{+}_{n:k}] &= 2 \times \sum_{l=k}^{n} \frac{1}{l(l-1)} \\ &= 2 \times \sum_{l=k}^{n} \Big(\frac{1}{l-1} - \frac{1}{l}\Big) \\ &= 2 \times ((k-1)^{-1} - n^{-1}).\end{aligned} \tag{2.8}$$

For $N = 10,000$ and $n \to \infty$, the expected coalescent time $\mathbb{E}[T^{+}_{n:40}]$ is 512.82 generations. For $N = 100,000$ and $n \to \infty$, the expected coalescent time $\mathbb{E}[T_{n:400}]$ is 501.25. If the majority of recombination endpoint comparisons happen at common ancestors older than five hundred generations, many haplotypes can be pruned ahead of time[8].

2.4.3 The probability that recombination endpoints are shared between haplotypes

At some point in the past, two sample haplotypes may share the same recombination endpoints to the left and right of a fixed location. Without loss of generality, let haplotypes a and b coalesce to their common ancestor d at time u, and let haplotypes c and d coalesce to their common ancestor e at time $u + v$. Observe that the recombination endpoints to

[7]But exponential random variables have heavy, non-negligible upper tail probabilities, so, in large samples, we may still detect some long IBD segments descendant from ancestors older than three hundred generations.

[8]Here we use expected values to give an analytically tractable Equation 2.8. The exponentially distributed $\{T_k\}$ are right-skewed in so much as times less than their means $\{\mathbb{E}[T_k]\}$ are more probable than those greater than.

the right $R_{a,d}, R_{b,d} \sim \text{Exponential}(u)$ and $R_{d,e} \sim \text{Exponential}(v)$. Figure S2 illustrates the coalescent tree in this scenario.

The merging step in Algorithm 1 serves to avoid comparing the endpoints of a and b with c when a and b have the same endpoints at time $u+v$. Specifically, if a and b's shared recombination endpoint $R_{d,e}$ is smaller than their separate endpoints $R_{a,d}$ and $R_{b,d}$, we can henceforth consider them "merged": they can be the treated as the same haplotype without loss of information. The probability of merging is

$$\begin{aligned} P(\min(R_{a,d}, R_{b,d}, R_{d,e}) = R_{d,e}) &= \int_{R_{d,e}=r} P(R_{d,e} = r) \times P(\min(R_{a,d}, R_{b,d}, r) = r) \\ &= \int_{R_{d,e}=r} P(R_{d,e} = r) \times P(R_{a,d} \geq r)^2 \\ &= \int_r v \exp(-(2u+v)r) dr \\ &= \frac{v}{2u+v}. \end{aligned} \tag{2.9}$$

For fixed u, the limit of Equation 2.9 is 1 as v gets large. This limit says that at some point haplotypes a and b will merge.

The limit argument is unsatisfying because the scaled TMRCA has a finite expected value $2 \times (1 - n^{-1})$ [74]. We derive a result that replaces arbitrary coalescent times u and $u+v$ with the expected times after the $(n-k)^{\text{th}}$ and $(n-j)^{\text{th}}$ coalescent events, respectively. Figure S2 illustrates the difference between the limit of Equation 2.9 and the following proposition in the context of branch lengths.

Proposition 2.4.1. *Let* $u/2 = \mathbb{E}[T^+_{n:(k+1)}] = 1/k - 1/n$ *and* $v/2 = \mathbb{E}[T^+_{n:(j+1)}] - \mathbb{E}[T^+_{n:(k+1)}] = 1/j - 1/k$ *(Equation 2.8). For* $j = o(k)$,

$$P(\min(R_{a,d}, R_{b,d}, R_{d,e}) = R_{d,e}) \to 1.$$

Proof. Note that $j = o(n)$ as well because $k \leq n$.

$$P(\min(R_{a,d}, R_{b,d}, R_{d,e}) = R_{d,e}) = \frac{1/j - 1/k}{1/j + 1/k - 2/n}$$

$$
\begin{aligned}
&= \frac{(k-j)n}{(nk+nj-2kj)} \\
&= \frac{1-j/k}{(1+j/k-2j/n)} \\
&\to 1.
\end{aligned}
$$

□

The implication of Proposition 2.4.1 is that haplotypes that share a recent common ancestor should have the same endpoints at the most distant common ancestors, and hence comparing haplotypes with the same endpoints versus newly coalesced haplotypes is redundant. Since recombinations to the right and left of a focal location are independent, the result of Proposition 2.4.1 extends to simulating IBD segments overlapping a focal location.

2.5 Simulation studies

In Chapters 3 and 4, we conduct enormous simulation studies involving sample sizes as large as ten thousand individuals and population sizes as large as ten million individuals. The sample and population individuals are "diploids", which we implement as a haploid model with the number of haploids equal to the number of individuals times a factor of 2. These empirical studies are feasible because of Algorithm 1, whose runtime we benchmark in this section.

2.5.1 Demographic scenarios

We consider constant population sizes of as little as two thousand individuals to as many as ten million individuals and two complex demographic scenarios. We refer to the complex demographic scenarios as examples of three phases of exponential growth and a population bottleneck. Figure 2.2 shows the demographic scenarios graphically. The three phases of exponential growth scenario involves an ancestral population of five thousand individuals that grew exponentially at different rates in three different time periods. This demographic model is similar to the "UK-like" model in Cai et al. [26]. The population bottleneck scenario

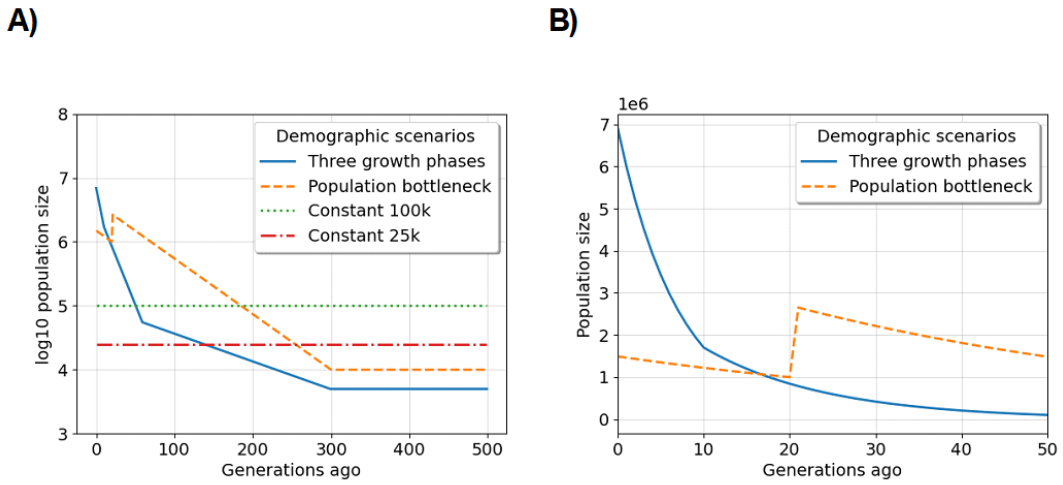

Figure 2.2: Demographic scenarios we consider in simulation studies: A) coalescent time in generations ago by the log 10 population size, and B) the most recent fifty generations by population size for examples of exponential growth. The legends specify the color and line style for each scenario. As opposed to coalescent time used in the main text, we describe here the scenarios forward in time. The three phases of exponential growth model is as follows: a population of ancestral size five thousand diploids increases exponentially each generation at rates one, seven, and fifteen percent starting three hundred, sixty, and ten generations ago. This demographic model is similar to the "UK-like" model in Cai et al. [26]. The population bottleneck model is as follows: a population of ancestral size ten thousand diploids increases exponentially each generation at a rate of two percent starting three hundred generations ago, but, twenty generations before the present day, the population experiences an instantaneous reduction in size to one million diploids. Otherwise, the demographic scenarios we explore here are populations of constant size twenty-five and one hundred thousand diploids.

involves an ancestral population of ten thousand individuals that grew exponentially at a fixed rate but experienced an instantaneous reduction in size twenty generations before the present day.

We also consider a genetic model for positive selection that is described in Chapter 4. Briefly, the allele frequency $p_s(t)$ decreases backward in time as a function of a nonnegative selection coefficient s. The selection coefficient reflects the advantage the allele has relative to alternative alleles. The larger the selection coefficient is, the faster the allele frequency increased. Also, the larger the selection coefficient is, the more detectable IBD segments there are on average. Positive selection around a locus is implemented via a coalescent with two subpopulations, one has the sweeping allele and one does not have the sweeping allele. The population sizes are $N_e(t) \cdot p(t)$ and $N_e(t) \cdot (1 - p(t))$. Until the coalescent reaches the sweeping alleles time of *de novo* mutation, IBD segments are not possible between individuals in separate subpopulations.

2.5.2 Compute time to simulate identity-by-descent segment lengths around a locus

To assess the effect of the pruning and merging rules, we evaluate four implementation strategies: merging and pruning (Algorithm 1), pruning only, merging only, and neither pruning nor merging (the naive approach). For each implementation, we run five simulations for sample sizes increasing by a factor of 2, recording the average wall clock compute time.

Figure 2.3 shows the average runtime per sample size between the methods. Simulating IBD segment lengths without pruning and merging takes more than one minute for eight thousand samples. Simulating IBD segment lengths with either pruning or merging can take less than one minute for sixty-four samples. Pruning appears to give a larger reduction in compute time than merging. Merging can further reduce runtime for sample sizes greater than one hundred thousand. The difference in five to ten seconds can be important when the number of simulations is enormous, as is the case in this study.

Figure S3 shows the algorithm's average runtime per sample size for different demographic scenarios and varying selection coefficients. Simulating IBD segment lengths takes more time for the population bottleneck and three phases of exponential growth scenarios compared to constant-size population scenarios. Runtime increases with the selection coefficient. The highest average measurement is more than four minutes for sixty-four thousand samples, the

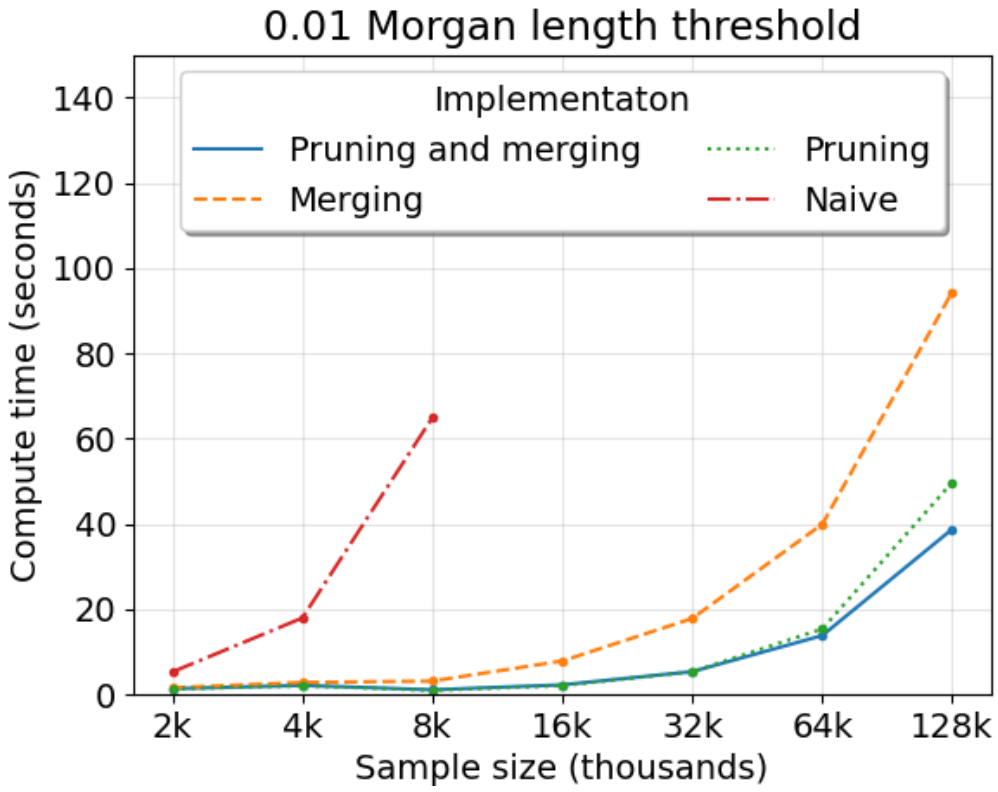

Figure 2.3: Compute time to simulate IBD segment lengths around a locus depending on algorithm implementation. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legend denotes colored line styles for implementations using Algorithm 1 as is (blue), merging only (orange), pruning only (green), and neither pruning nor merging (red). The main text describes "merging" and "pruning" techniques. The demography is the population bottleneck. The Morgan length threshold is 0.01.

population bottleneck scenario, and $s = 0.04$.

Overall, we benchmark that our simulation algorithm can be run tens of thousands of times within a day on one core processing unit of an Intel 2.2 GHz compute node. Despite performance savings, we observe that our simulation algorithm maintains quadratic behavior

in sample size (Figure 2.3 and Figure S3). One explanation for this finding is that a sizeable fraction of all lineages coalesce in the first few generations when the sample size exceeds ten thousand [14].

Coalescent time conditional on identity-by-descent segment detection

Pruning and merging should be most effective at decreasing the runtime of Algorithm 1 when coalescent times are many generations removed from the present day. The joint distribution of coalescent times that detectable IBD segments derive from is not known analytically. One notion of a typical conditional coalescent time of sample haplotypes a and b is to solve for t such that $\text{Mode}[W_{a,b}] = t^{-1}/2 = w$. Another notion of a typical conditional coalescent time of sample haplotypes a and b is to solve for t such that $\mathbb{E}[W_{a,b}] = t^{-1} = w$. This method of moments calculation does not account for the heavy tail of $W_{a,b}$. Moreover, the joint coalescent times come from an ancestral tree which can depend on complex demography and selection.

Figure S4 shows simulations of the coalescent times, conditional on IBD segments longer the 0.02 Morgans, for different constant population sizes. We observe that the densities of the coalescent times are roughly the same for population size N between ten thousand and one million diploids. This histogram and the following histograms concern distributions of conditional coalescent times, not the numbers of detectable IBD segments around a locus, which vary for different demographic scenarios. The distributions appear similar to the Gamma family distributions (which is not surprising[9]).

Figure S5 shows simulations of the coalescent times, conditional on IBD segments longer the 0.02 Morgans, for the three phases of exponential growth and population bottleneck demographic scenarios. Compared to constant-size populations, we observe larger densities

[9]For the first event time of sample haplotypes a and b, the conditional coalescent time $N \cdot T_n \mid W_{a,b} \geq w$ can be shown via convolution to be $\text{Gamma}(2, 2w + \binom{n}{2}N^{-1})$. We cannot easily study the conditional $\{T^+_{n:k}\}$ for $k < n$ because the unconditional $\{T_{n:k^+}\}$ are sums of independent Gamma random variables with different rates. If the rates were the same, via convolution, we would be able to analytically derive the conditional $\{T_{n:k^+}\}$ as a Gamma family distribution.

of coalescent times greater than fifty generations ago. These scenarios involve very large populations in the present day that expanded from ancestral population sizes of five and ten thousand individuals, respectively, indicating that detectable IBD segments in a rapidly increasing population can shed insight into demography hundreds of generations ago.

Figure 2.4 shows simulations of coalescent times, conditional on IBD segments longer the 0.02 Morgans, in selective sweeps of different magnitudes. For selection coefficients $s \leq 0.01$, we observe distributions similar to the neutral model in a constant population size. For selection coefficients $s = 0.02$ and $s = 0.04$, we see spikes in the observed coalescent times at one hundred and two hundred generations ago. These spikes correspond to the times at which the selected allele was introduced, demonstrating time depths where detectable IBD segments can provide information about strong selection.

In Figures S4, S5, and 2.4, we notice that most simulated coalescent times, conditional on detectable IBD segments around a locus, descend from common ancestors within the past three hundred generations. Meanwhile, our expected value calculations in Section 2.4.2 indicate that the oldest coalescent events should be older than three hundred generations. At these time depths, many haplotypes should be pruned in Algorithm 1 before their recombination endpoints are compared. Our calculations in Appendix C indicate that the expected number of recombination endpoint comparisons that occur at these oldest coalescent events is quadratic and could comprise a large proportion of the total $\binom{n}{2}$ comparisons. These calculations are consistent with the approximately linear runtime we observe in Figure 2.3 when pruning is implemented.

2.5.3 Compute time to simulate identity-by-descent segment lengths from the ancestral recombination graph

Some existing simulation frameworks can report IBD segment lengths. Baumdicker et al. [9] and Palamara [109] simulate coalescent trees with recombination along a chromosome, referred to as the ancestral recombination graph (ARG), which contains all the necessary data to calculate IBD segment lengths. The purpose of these simulation frameworks is not

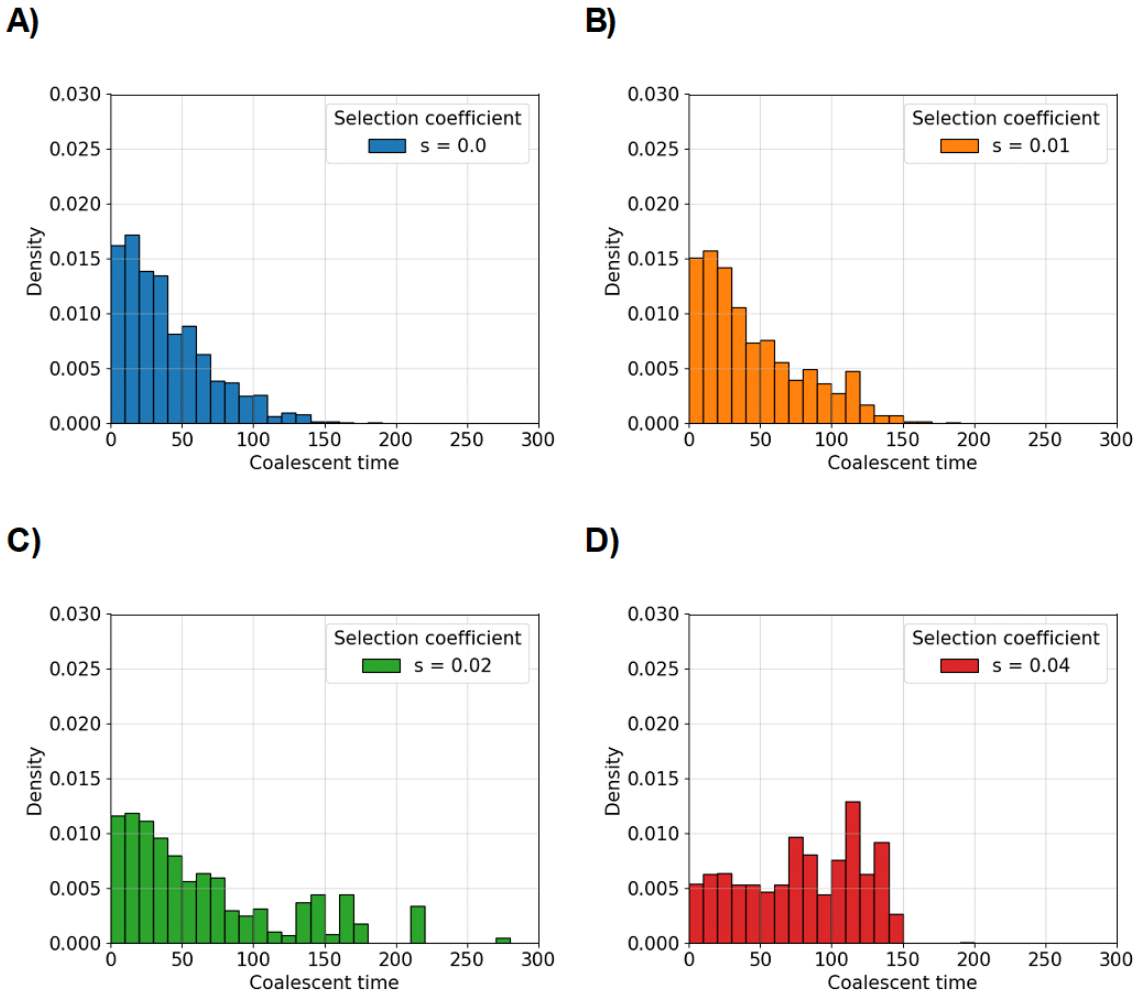

Figure 2.4: Empirical distributions of coalescent times conditional on detectable IBD segments around a locus for different selection coefficients. Histograms show the density of coalescent times split into fifty bins. Each panel represents one simulation. The selection coefficient is set to A) s=0.00, B) s=0.01, C) s=0.02, and D) s=0.04. The present-day allele frequency is twenty-five percent. The Morgan length threshold is 0.02. The population size is twenty-five thousand diploids. The sample size is five thousand diploids.

strictly to calculate IBD segment lengths, but Palamara [109] and Guo et al. [65] do offer utilities to output most IBD segments longer than a specified detection threshold.

We measure the times it takes `ARGON` (Palamara [109]) and `tskibd` (Guo et al. [65])[10] to simulate IBD segments $\geq$ 2.0 cM in a 7.0 cM region. For our study on IBD length distributions overlapping a focal point, these simulations are comparable to running Algorithm 1 with a 2.0 cM detection threshold. Both programs visit nodes in the ARG in small, non-overlapping sliding windows. We consider window sizes of 0.01 and 0.001 cM in benchmarking runtimes. Some true IBD segments will not be detected if the window size is too large, but decreasing the window size increases runtime.

Table 2.1 reports the average runtimes of each method for increasing sample size in the population bottleneck demographic scenario. The time spent to simulate the ARG is dominated by the time spent to calculate IBD segment lengths. `ARGON` takes nearly an hour to simulate the IBD segment length distribution of two thousand diploids. We do not run it for more than two thousand diploids due to concerns surrounding quadratic runtimes. With 0.01 cM windows, `tskibd` takes less than twenty minutes to simulate the IBD length distribution of four thousand diploids and a little over an hour to simulate the IBD length distribution of eight thousand diploids. To get more precise IBD segment endpoints with `tskibd`, we use 0.001 cM windows, which can increase runtime eightfold or more. In comparison, for eight thousand diploids, our improved approach simulates IBD segments $\geq$ 1.0 cM around a locus in less than two seconds (Figure 2.3). Even our naive approach completes the same scope of simulations in less than two minutes.

2.6 Discussion

This chapter formally introduces coalescent time and IBD segment lengths as random variables, and it uses simulation to build intuition for the distributional properties of these random variables. IBD segment lengths around a locus, conditional on coalescent times, are

[10] This measurement does not include the time to simulate an ARG with `msprime` (Baumdicker et al. [9]).

Method	Sample size	Window size (cM)	Hours	Minutes	Seconds
`ARGON`	500	0.01	0.0	3.2	19.8
	1000		0.0	10.3	34.7
	2000		0.1	49.6	21.9
`tskibd`	500	0.01	0	0	6.8
	1000		0	0	28.8
	2000		0	2	31.8
	4000		0	15	21.6
	8000		1	13	29.4
`tskibd`	500	0.001	0	0.9	16.0
	1000		0	6.0	10.2
	2000		0	23.8	32.4
	4000		2	55.7	36.8
	8000		8.9	23.5	23.9

Table 2.1: Average runtime to simulate detectable IBD segments with `ARGON` and `tskibd`. Using `ARGON` or `tskibd`, we report the mean number of hours, minutes, and seconds to simulate IBD segments ≥ 2.0 cM in a 7.0 cM region. Sample sizes range from 500 to 8000 diploids. Sliding non-overlapping window sizes are either 0.001 or 0.01 cM. Averages runtimes are taken over ten simulations.

modeled as Gamma family random variables. The rate parameters are often large for IBD segment lengths, resulting in a heavy left skew towards small values. We exploit this fact in developing an efficient algorithm to generate long IBD segments. The algorithm's compute time scales approximately linear as the number of haplotype pairs increases quadratically. For example, we benchmark that simulating a joint distribution of detectable IBD segment lengths around a locus can take as little as a couple of seconds or tens of seconds for sample sizes of order 10^4 or 10^5, respectively. Some human biobanks have sample sizes of these

orders, which serve as the current upper bounds on sample size in genetic data.

Existing methods `ARGON` and `tskibd` simulate IBD segment lengths for tens of cM regions and thousands of samples within hours to days. Those methods simulate IBD segment lengths as a feature of a much broader simulation framework, whereas our algorithm is narrowly designed to study the IBD segment lengths around a locus but runs orders of magnitude faster. Another feature of our method, versus those methods that we compare to, is modeling the effects of positive selection on IBD segment lengths around a locus. In Chapter 4, we demonstrate one important use case of our algorithm: quantifying the uncertainty of an estimator in selective sweeps via the parametric bootstrap. For sample sizes greater than a few thousand individuals, with `ARGON` or `tskibd` making bootstrap confidence intervals for summary statistics of IBD segment lengths around a locus is impractical.

In the following Chapters 3 and 4, we use our simulation algorithm to study distributional behaviors of IBD segment lengths around a locus and statistical guarantees of confidence intervals. In both use cases, we simulate hundreds to thousands of times to generate empirical distributions on IBD segment lengths around a locus, and then we replicate that procedure for thousands to tens of thousands of hypothesis tests or estimates of confidence interval coverage probabilities, respectively. Altogether, these empirical studies amount to billions of simulations. Despite the speed of our method for a single locus, the scope of our simulations takes hundreds of days of compute time, which we spread across core processing units. Without our simulation algorithm, we would not be able to assess certain aspects of statistical inference in the expansive investigations of this dissertation.

Chapter 3

IDENTITY-BY-DESCENT IN LARGE SAMPLES

3.1 Introduction

In population genetics, complex correlation structures make it challenging to show theoretical results for statistical properties like asymptotic consistency and normality. For independent, identically distributed data, maximizing likelihood estimators are asymptotically consistent, efficient, and normally distributed under regularity conditions [31]. Such properties are appealing in statistical inference. A few IBD-based methods that estimate population genetics parameters solve composite likelihoods [110, 148, 150]. To what extent these properties extend to composite likelihood estimators is generally unknown [94]. Studying these estimators can be especially challenging if the maximum does not have a closed form [110, 150].

Throughout this dissertation, we will study a closed-form sample mean and functions of it. Consider the indicator function that an IBD segment is longer than some Morgan detection threshold. All the IBD segment indicators are a sample of binary random variables. Obviously, IBD segments are correlated (the same individual is compared against every other individual). Meanwhile, we contrast the sum of IBD segment indicators to the sum of *independent* Bernoulli random variables with the same expectation. Figure 3.1 illustrates that the sum of IBD segment indicators has more variance than the sum of independent Bernoulli random variables with the same mean. The two distributions are more similar when the Morgan threshold is 0.03 instead of 0.02, and they are both visually similar to the normal distribution. The central limit theorem applies to the sample mean of the independent Bernoulli random variables, so the latter observation is not surprising. At the same time, Figure 3.1 suggests the possibility of a central limit theorem for the IBD rate that depends on the detection threshold.

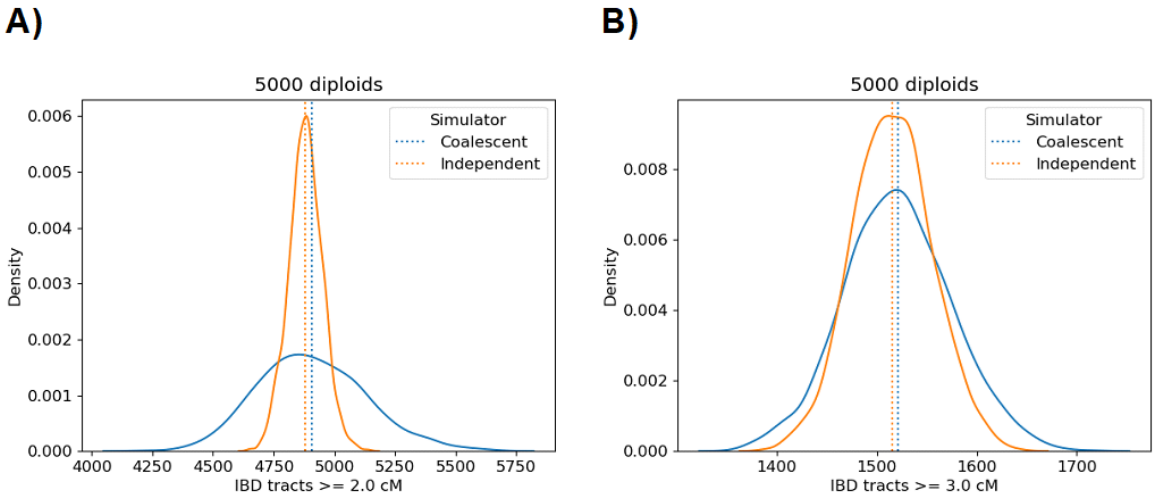

Figure 3.1: The identity-by-descent rate has more variance than the binomial sample mean. Kernel density estimates show the number of IBD segments A) $\geq$ 2.0 and B) $\geq$ 3.0 cM over 2,500 replicate simulations of five thousand diploids in the population bottleneck demographic scenario. Blue denotes data from Algorithm 1 and orange denotes data from sampling a Binomial distribution with success probability equal to the probability of a detectable IBD segment $P(W_{a,b} \geq w)$ and the number of trials equal to the number of haplotype pairs. Vertical dashed lines denote averages.

In this chapter, we derive sufficient conditions under which the proportion of detectable IBD segments around a locus[1] is asymptotically normally distributed. Our central limit theorem concerns a mean of correlated binary variables, where the correlation structure comes from an unobserved and random ancestral tree. The proof is to show that the variance of detectable IBD segments[2] dominates the covariance between detectable IBD segments[3]. Our conditions involve a minimum length of detectable IBD segments and the population size

[1]This statistic maximizes the binomial composite likelihood with parameters $\binom{n}{2}$ and the probability of an IBD segment longer than a Morgan threshold.

[2]The diagonal entries of the covariance matrix

[3]The off-diagonal entries of the covariance matrix

that a large sample is drawn from. The population size requirement, in particular, indicates that most of the branch lengths in the ancestral tree must be long for the result to hold. The overall contribution of this work is to support IBD-based statistical inference with rigorous theory and extensive simulation studies.

The outline of this chapter is as follows. In Section 3.2, we formally define IBD segment indicators. In Section 3.3, we present and prove our main result for the asymptotic normality of the detectable IBD rate. In Section 3.4, we use simulation to investigate the statistical properties of the detectable IBD rate and IBD graphs around a locus. Many calculations of covariance terms are left to the Appendix A.

3.2 The presence of detectable IBD segments

Relative to a focal point, we focus on the detection of long IBD segments in a sample. Recall that $R_{a,b}, L_{a,b}$, and $W_{a,b}$ are the IBD segment lengths to the right of, to the left of, and overlapping a focal point, respectively. Let $X_{a,b} := X_{a,b}(w) = I(R_{a,b} \geq w)$ indicate if the IBD segment to the right that is shared by sample haplotypes a and b is longer than a Morgan threshold w. The binary random variables $\{X_{a,b}\}$ are identically distributed with the same mean $\mathbb{E}_2[X_{a,b}]$, and they are correlated through the unobserved coalescent tree. We use $\mathbb{E}_2, \mathbb{E}_3$, and $\mathbb{E}_4$ and $\mathrm{Cov}_2, \mathrm{Cov}_3$, and Cov_4 to denote expected values and covariances with respect to coalescent trees of two, three, and four sample haplotypes, respectively.

The random variables important to our central limit theorem are defined as follows, borrowing notation from Chandrasekhar et al. [33]. The detectable IBD rate is

$$\bar{\mathbf{X}}_{\binom{n}{2}} := \binom{n}{2}^{-1} \sum_{(a,b)} X_{a,b}. \tag{3.1}$$

The number of detectable IBD segments $\sum_{(a,b)} X_{a,b}$ is the statistic shown in Figure 3.1. Let $Z_{a,b} := X_{a,b} - \mathbb{E}_2[X_{a,b}]$ be the mean-centered binary random variable, and let the sum of all binary random variables minus one be $\mathbf{Z}_{-a,b} := \sum_{(c,d)} Z_{c,d} - Z_{a,b}$. The sum of variances of

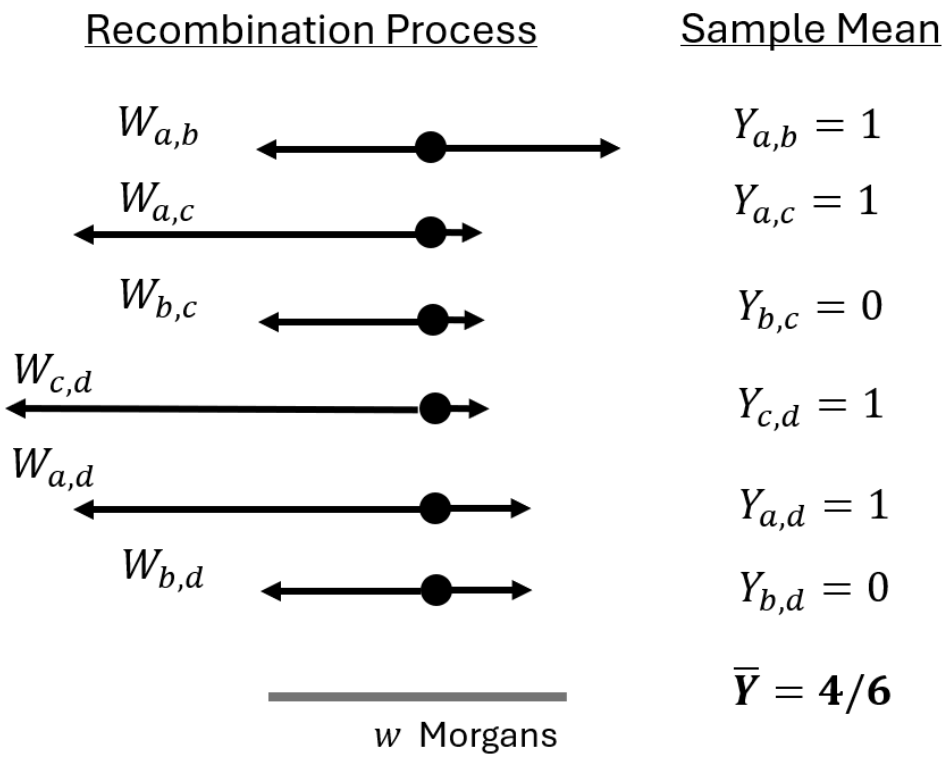

Figure 3.2: Example calculation of the detectable IBD rate. IBD segment lengths overlapping a focal point for sample haplotypes a, b, c, d are shown. The IBD segment indicators ($Y_{i,j}$'s) are 1 if their IBD segment lengths ($W_{i,j}$'s) exceed w Morgans and otherwise 0. The detectable IBD rate $\bar{\mathbf{Y}}$ around a locus is the mean of these correlated binary random variables. The detectable IBD rate to the right of the focal point, $\bar{\mathbf{X}}$, is calculated similarly.

all IBD segment indicators is

$$\Omega_{\binom{n}{2}} := \sum_{(a,b)} \mathrm{Var}(X_{a,b}) = \binom{n}{2} \times \mathbb{E}_2[X_{a,b}] \times (1 - \mathbb{E}_2[X_{a,b}]) \tag{3.2}$$

Finally, the mean-centered and suitably scaled IBD rate is

$$\bar{\mathbf{Z}}_{\binom{n}{2}} := \Omega_{\binom{n}{2}}^{-1/2} \times (\bar{\mathbf{X}}_{\binom{n}{2}} - \mathbb{E}_2[X_{a,b}]). \tag{3.3}$$

For IBD segments overlapping a focal location, let $Y_{a,b} := I(L_{a,b} + R_{a,b} \geq w)$ and $\tilde{Z}_{a,b} := Y_{a,b} - \mathbb{E}_2[Y_{a,b}]$. The terms $\bar{\mathbf{Y}}_{\binom{n}{2}}$ $\tilde{\mathbf{Z}}_{-a,b}$, $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2}}$, and $\tilde{\Omega}_{\binom{n}{2}}$, are defined analogously to $\bar{\mathbf{X}}_{\binom{n}{2}}$, $\mathbf{Z}_{-a,b}$,

$\bar{\mathbf{Z}}_{\binom{n}{2}}$, and $\Omega_{\binom{n}{2}}$, respectively. Figure 3.2 provides a conceptual example calculating $\bar{\mathbf{Y}}$ for four sample haplotypes.

3.3 The asymptotic normality of the identity-by-descent rate

The standard central limit theorem does not apply in our case because the IBD segment indicators $\{X_{a,b}\}$ to the right of a focal point are not independent of each other[4]. We start by focusing on the mean-centered and suitably scaled IBD rate $\bar{\mathbf{Z}}_{\binom{n}{2},N}$ to the right of a focal location, where N denotes the constant population size that the haplotypes are sampled from. The intuition for our weak law is that the covariance between IBD segment indicators $\sum_{(a,b)\neq(c,d)} \mathrm{Cov}(X_{a,b}, X_{c,d})$ is small relative to the sum of the variances of the individual IBD segment indicators $\Omega_{\binom{n}{2}}$. This idea of the covariance between random variables being much smaller than the variance of the random variables themselves is the basis of the general central limit theorem for dependent data that is given in Chandrasekhar and Jackson [32] and Chandrasekhar et al. [33]. The following central limit theorems concern large sample size n and large population size N scaled by Morgan length threshold w [5].

Theorem 3.3.1. *The mean-centered and suitably scaled IBD rate statistic $\bar{\mathbf{Z}}_{\binom{n}{2},N}$ converges in distribution to the standard normal distribution for n and Nw tending to infinity when the following are true:*

1. *$Nw = o(n^2)$, scaled population size is small relative to the number of pairs;*

2. *$n = o(Nw)$, sample size is small relative to scaled population size;*

3. *$\mathbb{E}[Z_{a,b} \times \mathbf{Z}_{-a,b} | \mathbf{Z}_{-a,b}] \geq 0$ for all $\mathbf{Z}_{-a,b}$.*

Proof. We show that our three conditions are sufficient to apply Corollary 1 in Chandrasekhar et al. [33]. Without loss of generality, we derive integrals over a tree with two sample

[4]Similarly, IBD segment indicators $\{Y_{a,b}\}$ overlapping a focal location are not independent of each other.

[5]Fixed detection thresholds $0.01 \leq w \leq 0.04$ are commonly used to control for false positives and negatives in inference from genetic data [54, 134, 163, 105, 104].

haplotypes a and b, a tree with three sample haplotypes a, b, and c, and a tree with four sample haplotypes a, b, c, and d.

$$\mathbb{E}_2[X_{a,b}] = \int \exp(-2Nt_2w)\exp(-t_2)\,dt_2 = (2Nw+1)^{-1} = O((Nw)^{-1}). \tag{3.4}$$

It is easy to show that $\mathbb{E}_2[X_{a,b}] \to 0$ uniformly for large scaled population size. The second condition implies that $\Omega_{\binom{n}{2}} \to \infty$. The assumption in Chandrasekhar et al. [33] that $\mathbb{E}[|Z_{a,b}|^3]/\mathbb{E}[|Z_{a,b}|^2]^{3/2}$ is bounded above is true for non-degenerate Bernoulli random variables [32]. Lastly, given $n = o(Nw)$, we show that

$$\sum_{(a,b)\neq(c,d)} \mathrm{Cov}(\, X_{a,b}, X_{c,d}\,) = o(\Omega_{\binom{n}{2}}). \tag{3.5}$$

In Appendix A, we derive bounds on the integrals $\mathrm{Cov}_3(X_{a,b}, X_{a,c}) = O((Nw)^{-2})$ and $\mathrm{Cov}_4(X_{a,b}, X_{c,d}) = O((Nw)^{-3})$. Next, there are $n(n-1)(n-2) \sim n^3$ combinations of three haplotypes a, b, and c, and there are $n(n-1)(n-2)(n-3)/4 \sim n^4$ combinations of four haplotypes a, b, c, and d. In asymptotic arguments, the notation $\sim$ means asymptotic equivalence, not distributed as.

$$\Omega_{\binom{n}{2}} \sim n^2 \cdot O((Nw)^{-1}) = o((Nw)^2) \cdot O((Nw)^{-1}) = o(Nw); \tag{3.6}$$

$$\sum_{a,b,c} \mathrm{Cov}_3(X_{a,b}, X_{a,c}) \sim n^3 \cdot O((Nw)^{-2}) = o((Nw)^3) \cdot O((Nw)^{-2}) = o(Nw); \tag{3.7}$$

$$\sum_{a,b,c,d} \mathrm{Cov}_4(X_{a,b}, X_{c,d}) \sim n^4 \cdot O((Nw)^{-3}) = o((Nw)^4) \cdot O((Nw)^{-3}) = o(Nw). \tag{3.8}$$

Therefore, the sum of covariances between IBD segment indicators (Equations 3.7 and 3.8) is controlled by the sum of variances of the individual IBD segment indicators (Equation 3.6). □

The first two conditions have appealing interpretations. First, $Nw = o(n^2)$ says that the sample size is large enough relative to the scaled population size such that we observe many IBD segments to the right of a focal location that are longer than the Morgan threshold w. Second, $n = o(Nw)$ says that the sample size is not too large relative to the scaled population

size such that we do not observe many large clusters of haplotypes with IBD segments to the right of a focal location that are longer than the Morgan threshold w.

The third condition also has an interpretation in the context of population genetics. It says that if the number of detectable IBD segments to the right of a focal location, except for $X_{a,b}$, is less than the expectation $\mathbb{E}[X_{a,b}] \times (\binom{n}{2} - 1)$, then the IBD segment to the right of a focal location that is shared by a and b is shorter than w on average, and vice versa if $\mathbf{X}_{-a,b}$ is greater than its expected value. This assumption seems plausible if IBD segments to the right of a focal point have non-negative covariance, which we show in Appendix A.

One can show that the small sample size $n = 3$ is a pathological example where the third condition breaks down. We demonstrate this issue in Appendix A. We do not otherwise calculate $\mathbb{E}[Z_{a,b} \times \mathbf{Z}_{-a,b} | \mathbf{Z}_{-a,b}]$ for all $\mathbf{Z}_{-a,b}$, which involves integration over the space of all coalescent trees and the $2^{\binom{n}{2}-1}$ hypercube of 0's and 1's. In a simulation study, we evaluate the third condition via the Monte Carlo method (Appendix A), concluding that this condition might hold in large samples.

The asymptotic normality of $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ follows from the same arguments as those of the proof in Theorem 3.3.1. We show in Appendix A that $\mathrm{Cov}_2(Y_{a,b}, Y_{a,b})$, $\mathrm{Cov}_3(Y_{a,b}, Y_{a,c})$, and $\mathrm{Cov}_4(Y_{a,b}, Y_{c,d})$ are $O((Nw)^{-1})$, $O((Nw)^{-2})$, and $O((Nw)^{-3})$, respectively.

Theorem 3.3.2. *The mean-centered and suitably scaled IBD rate statistic $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ converges in distribution to the standard normal distribution for n and Nw tending to infinity when the following are true:*

1. $Nw = o(n^2)$;
2. $n = o(Nw)$;
3. $\mathbb{E}[\tilde{Z}_{a,b} \times \tilde{\mathbf{Z}}_{-a,b} | \tilde{\mathbf{Z}}_{-a,b}] \geq 0$ *for all* $\tilde{\mathbf{Z}}_{-a,b}$.

Proof. The argument is the same as in Theorem 3.3.1. The difference is that we use the sum of recombination distances to the right and left of a focal location. In Lemma A.5, we derive

the same bounds $O((Nw)^{-1}), O((Nw)^{-2})$, and $O((Nw)^{-3})$ for $\mathrm{Cov}_2(Y_{a,b}), \mathrm{Cov}_3(Y_{a,b}, Y_{a,c})$, and $\mathrm{Cov}_4(Y_{a,c}, Y_{b,d})$, respectively. □

We can derive a similar result for varying population sizes. Let $N_1 = \max_t N(t)$ and $N_2 = \min_t N(t)$. Compared to varying population sizes $N(t)$, the indicator of a detectable IBD segment around a focal location has larger expected value and variance when sample haplotypes come from a constant population of size N_2. Conversely, compared to varying population sizes $N(t)$, the indicator of a detectable IBD segment around a focal location has smaller expected value and variance when sample haplotypes come from a constant population of size N_1. We use these facts to establish covariance bounds for complex demography.

Theorem 3.3.3. *The mean-centered and suitably scaled IBD rate statistic* $\bar{\mathbf{Z}}_{\binom{n}{2},N(t)}$ *converges in distribution to the standard normal distribution for* n, $N_1 w$, *and* $N_2 w$ *tending to infinity when the following are true:*

1. $N_1 w = o(n^2)$;
2. $n = o(N_2 w)$;
3. $\mathbb{E}[Z_{a,b} \times \mathbf{Z}_{-a,b} | \mathbf{Z}_{-a,b}] \geq 0$ *for all* $\mathbf{Z}_{-a,b}$.

The same conditions imply weak convergence for $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N(t)}$.

Proof. The argument is the same as in Theorem 3.3.1, except we use N_1 and N_2 to upper and lower bound covariance terms.

$$\Omega_{\binom{n}{2}} \sim n^2 \cdot O((N_2 w)^{-1}) = o(N_2 w); \tag{3.9}$$

$$\sum_{a,b,c} \mathrm{Cov}_3(X_{a,b}, X_{a,c}) \sim n^3 \cdot O((N_2 w)^{-2}) = o(N_2 w); \tag{3.10}$$

$$\sum_{a,b,c,d} \mathrm{Cov}_4(X_{a,b}, X_{c,d}) \sim n^4 \cdot O((N_2 w)^{-3}) = o(N_2 w). \tag{3.11}$$

□

Theorem 3.3.1 is a special case of Theorem 3.3.3 when $N_1 = N_2$. The first two conditions in Theorem 3.3.3 on sample size and scaled population size are unlikely to hold in real data examples, and they are more difficult to interpret. Note that the proof of Theorem 3.3.3 does not make use of the entire curve $N(t)$. The population sizes at the most recent coalescent times impact the covariance of and between IBD segments around a focal location the most. As is the case in Theorem 3.3.2, we can extend Theorem 3.3.3 to address IBD segments overlapping a focal location.

3.4 Simulations studies

In comparing the standardized detectable IBD rate $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ overlapping a focal point to the standard normal distribution, we require massive simulations to form tens of thousands of empirical distributions. To investigate the asymptotic conditions indicated by Theorems 3.3.1, 3.3.2, and 3.3.3, we increase sample sizes up to ten thousand, which could involve 10^8 recombination endpoint comparisons in the worst case. The speed of Algorithm 1 is exceptionally important to simulation studies of this magnitude and scope.

3.4.1 The identity-by-descent rate in finite samples

Hypothesis testing for normality

Recall that $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ is shown to be asymptotically normally distributed under regularity conditions on sample size *and* scaled population size. Using the Shapiro-Wilk test, we investigate if empirical distributions of $\sum_{a,b} Y_{a,b}$ resemble normal distributions as sample size n, population size N, and the Morgan length threshold w increase. We partition simulated data into five hundred empirical distributions based on one thousand observations. The null hypothesis of this test is that the data is normally distributed. Rejecting the null hypothesis means that there is enough evidence indicating that the empirical distribution is not normally distributed. We report the proportion of times we reject the null hypothesis at the 0.05 confidence level.

We consider sample sizes between one thousand and ten thousand individuals and population sizes between ten thousand and ten million individuals. We simulate the count of IBD segments $\sum_{a,b} Y_{a,b}$ longer than Morgan length thresholds 0.01, 0.02, 0.03, and 0.04. Due to heavy compute times, for constant population size and all sample sizes, we implement the n-coalescent instead of the discrete-time Wright-Fisher process[6], despite the issues mentioned in Section 2.4.1.

Figure 3.3 shows the proportion of rejected tests for increasing population size and Morgan length threshold with sample size fixed at five and ten thousand individuals. The trend is that the proportion of rejected tests decreases for increasing population size and Morgan length threshold. Figure S6 shows that this trend does not depend on the confidence level. These observations align with the condition $n = o(Nw)$ in Theorem 3.3.1 and Theorem 3.3.2. The setting for which the proportion is closest to 0.05 is $n = 10^4$, $N = 10^6$, and $w = 0.04$. Interestingly, for the same sample size and Morgan length threshold, we observe more rejected tests for $N = 10^7$ than for $N = 10^6$. This observation aligns with the condition $Nw = o(n^2)$ in Theorem 3.3.1 and Theorem 3.3.2 (there are too few observed IBD segments).

Figure S7 shows the proportion of rejected tests for increasing sample size and Morgan length threshold with population size fixed at fifty and one hundred thousand individuals. The proportion of rejected tests decreases slightly with increasing sample size. This trend may be explained by the fact that sample size does not affect the correlations of IBD segment indicators (Lemmas A.3, A.4, and A.5).

Figure S8 shows the proportion of rejected tests for increasing sample size and Morgan length threshold in the three phases of exponential growth and population bottleneck demographic scenarios (Section 2.5.1). For Morgan length threshold greater than or equal to 0.03, the proportions of rejected tests are less than 0.3 and 0.1 in the three phases of exponential growth and population bottleneck scenarios, respectively. Consistent with our central limit theorems, we observe a decreasing trend as we increase the Morgan length threshold, even

[6] In these simulations only

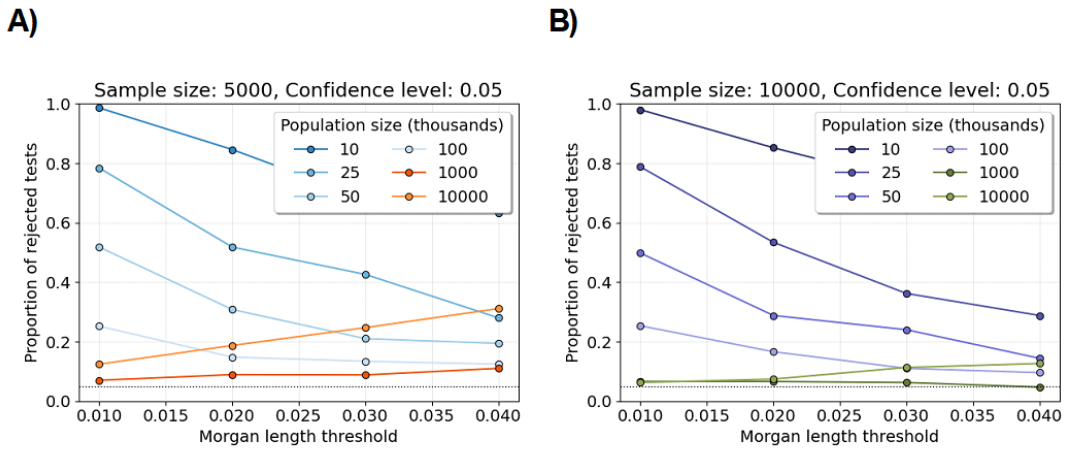

Figure 3.3: Shapiro-Wilk tests for varying population sizes. Line plots show the proportions of Shapiro-Wilk tests rejected at a confidence level of 0.05 (y-axis) for varying population size and fixed sample size. Each proportion is computed over five hundred tests. Each test is based on an empirical distribution from one thousand simulations of the number of identity-by-descent segments longer than a specified Morgan length threshold (x-axis). A) The sample size is five thousand diploids. B) The sample size is ten thousand diploids. The legends give colors assigned to different population sizes. The horizontal dotted line is at 0.05.

though the proportions of rejected tests around 0.3 and 0.1 are not close to the nominal significance level of 0.05. Additionally, these proportions are less than their corresponding proportions in the population of twenty-five thousand individuals (Figure 3.3). In both demographic scenarios, the population sizes exceed twenty-five thousand in the most recent one hundred generations but three hundred generations ago their population sizes were five and ten thousand (Figure 2.2). While the conditions in Theorem 3.3.3 on sample size and scaled population size are very strict, in demographic scenarios with large recent population sizes, the detectable IBD rate may have normal-like distributional behavior, regardless of their more ancient population sizes (Figure S8).

Overall, we reject normality at rates greater than the nominal 0.05 significance level with the sample sizes and population sizes explored here. These magnitudes are already quite large relative to existing sample sizes and inferred effective population sizes[7]. Nevertheless, the trends of increasing sample size and scaled population size suggest the fidelity of our central limit theorems.

The upper tail of the detectable identity-by-descent rate distribution

Excess IBD rate overlapping a focal location may indicate genomic regions under strong and recent positive selection [20, 148]. Let $\bar{\mathbf{Y}}_m$ be the detectable IBD rate overlapping the m^{th} focal location out of M locations along a genome. The genome-wide sample mean and standard deviations are

$$\hat{\mu}_{1:M} := M^{-1} \sum_{m=1}^{M} \bar{\mathbf{Y}}_m \,; \tag{3.12}$$

$$\hat{\sigma}_{1:M} := \sqrt{(M-1)^{-1} \sum_{m=1}^{M} (\bar{\mathbf{Y}}_m - \hat{\mu}_{1:M})^2}. \tag{3.13}$$

Browning and Browning [20] and Temple et al. [148] report regions where the detectable IBD rate overlapping a focal location exceeds $\hat{\mu}_{1:M} + 4 \times \hat{\sigma}_{1:M}$ as putative loci under strong and recent selection.

When Theorem 3.3.3 holds, we can interpret each such marginal hypothesis test as a one-sample one-sided $z-$test. Under this null hypothesis, we can compute p-values for each $\bar{\mathbf{Y}}_m$. We evaluate the control of FWER in finite samples from finite populations. Using the same simulated data as in Section 3.4.1, Figure 3.4 shows the average upper bound $\hat{\mu}_{1:M} + 4 \times \hat{\sigma}_{1:M}$ divided by the 99.99683 percentile over two million simulations, where the Gaussian cumulative distribution function $\Phi(4) = 0.9999683$. Figure S9 shows the same as in Figure 3.4 except the Gaussian quantile is three and $\Phi(3) = 0.9986501$. The scan threshold is less than the simulated percentile threshold for all sample sizes, population sizes, Morgan length thresholds, and significance levels considered. The scan threshold is proportionally

[7] At least for humans [19, 26]

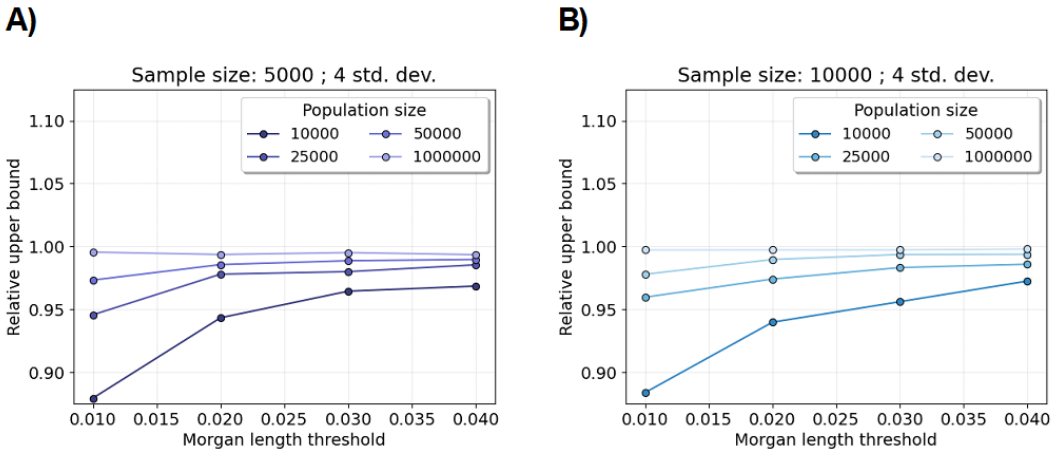

Figure 3.4: Relative upper bound for excess IBD scan. Line plots show the average mean plus four standard deviations divided by the 99.99683 percentile over two million simulations (y-axis). (The standard normal survival function of four is 0.9999683.) Each average relative upper bound is computed over one thousand tests. Each test is based on two thousand simulations of the number of identity-by-descent segments longer than a specified Morgan length threshold (x-axis). A) The sample size is five thousand diploids. B) The sample size is ten thousand diploids. The legends give colors assigned to different constant population sizes.

closer to the percentile threshold as population size and Morgan length threshold increase, which is a result consistent with Section 3.4.1 and our central limit theorems. Figure S8 shows that the scan threshold is also less than the simulated percentile threshold for all sample sizes and Morgan length thresholds in the three phases of exponential growth and population bottleneck demographic scenarios. The scan threshold is proportionally closer to the percentile threshold for the population bottleneck scenario compared to the three phases of exponential growth scenario.

We observe that the simulated distributions of $\bar{\mathbf{Y}}$ can be right skewed and display upper tails that are not sub-normal. For the 0.02 Morgan length threshold and a sample size of five

thousand individuals, we calculate the percentage of simulations where the detectable IBD rate overlapping a focal location is above the three and four standard deviation upper bounds. Figure S10 shows these percentages multiplied by three thousand, which we refer to as the expected number of false positives. Three thousand corresponds to a marginal hypothesis test every 0.01 Morgans along a 30 Morgans genome[8,9]. The expected number of false positives decreases as the population size increases. We measure less than two expected false positives for the four standard deviations test and as much as fifteen expected false positives for the three standard deviations test. Fifteen expected false positives is eleven more false positives than there would be if FWER were controlled. This experiment demonstrates that deviations from normality in finite samples from finite populations can lead to more false positives than expected under normality.

3.4.2 Identity-by-descent graphs around a location

Clusters of detectable IBD haplotypes overlapping a focal point indicate non-negligible covariance between segments and could thus explain the observed non-normality in finite samples. In place of theoretical work, we leverage our simulation framework to describe the attributes of detectable IBD graphs around a locus. We form detectable IBD graphs about a locus by drawing an edge between haplotypes if they share a detectable IBD segment overlapping a focal point. We define detectable IBD clusters to be the connected components in the detectable IBD graph.

We analyze five features of graphs. The number of edges is equivalent to the number of IBD segments longer than the length threshold. A tree of order m is a connected component that has m nodes and $m-1$ edges. An order m complete connected component has m nodes and edges between every pair of nodes. We count the number of trees of order 2 and 3, the number of complete connected components of order 3 or more, and the number of nodes

[8]Ignoring spatial correlation of tests

[9]The human genome is of this scale 3300 cM. Temple et al. [148] suggest studying loci that have excess IBD rates spanning 0.01 Morgans / 1 cM as targets of putative selection.

in the largest connected component. For each feature, we calculate the average, variance, minimum, and maximum over replicate simulations. We also conduct Shapiro-Wilk tests by splitting the simulated data as described in Section 3.4.1.

Comparing to sparse Erdős-Rényi graphs

The Erdős-Rényi graph is a simple network model in which independent edges between nodes occur with a uniform success probability [43]. We denote a sparse Erdős-Rényi network as one in which the success probability is vanishingly small. We compare the features of connected components between detectable IBD and Erdős-Rényi graphs, setting the uniform success probability to be the approximate probability of an IBD segment longer than a Morgan length threshold [110]. This contrast analyzes the evolution of independent edges versus weakly correlated edges of a specific nature.

For sparse Erdős-Rényi graphs, there are theoretical properties associated with the graph features that we consider in our simulation study. When the success probability is small, the number of trees of order m weakly converges to a Gaussian distribution in large networks [44]. Trees of order m_1 have faster convergence than trees of order m_2 when $m_1 < m_2$. Another asymptotic property of sparse Erdős-Rényi graphs says that almost all nodes are in trees of small order or in a single "giant" component [44].

In a sample size of five thousand from a population of one hundred thousand individuals, Figure 3.5 shows that some empirical distributions of graph features resemble normal distributions. Table 3.1 compares our summary statistics between these simulated detectable IBD and sparse Erdős-Rényi graphs. The variance, minimum, and maximum number of edges are larger for detectable IBD graphs compared to sparse Erdős-Rényi graphs, which is a direct consequence of the non-zero covariance of IBD edges[10]. The proportions of rejected hypothesis tests for numbers of trees of order 2 and connected components of degree 3 or more are close to 0.05 for both detectable IBD and sparse Erdős-Rényi graphs. While we

[10]The expected number of edges should be the same, if not for the tail approximation in the probability of detectable IBD segments [110].

Type	Structure	Avg	Var	Min	Max	S.W.t.
IBD	Edges	1,283.42	2,690.85	1,072.00	1,530.00	0.14
	Largest	8.09	1.81	5.00	22.00	1.00
	Tree-2	483.62	346.48	402.00	569.00	0.05
	Tree-3	29.40	28.38	9.00	57.00	0.81
	Complete	135.89	112.45	93.00	187.00	0.18
Erdős-Rényi	Edges	1,312.68	1,313.06	1,158.00	1,475.00	0.07
	Largest	27.02	74.07	11.00	137.00	1.00
	Tree-2	353.31	310.32	284.00	434.00	0.08
	Tree-3	120.31	109.73	78.00	173.00	0.14
	Complete	174.94	146.10	123.00	228.00	0.13

Table 3.1: Summary statistics of IBD and Erdős-Rényi graphs. Network structures of interest are the number of edges (Edges), the degree of the largest components (Largest), the number of trees of order 2 and 3 (Tree-2 and Tree-3), and the number of complete components of degree 3 or more (Complete). Summary statistics are aggregated over 125,000 simulations. Shapiro-Wilk tests at a confidence level of 0.05 are performed with 500 replicates for 250 simulations. The proportions of rejected null hypotheses are reported as S.W.t. The population size is one hundred thousand diploids. The sample size is two thousand diploids. The Morgan length threshold is 0.03.

observe that some limiting distributional behaviors of small degree connected components in detectable IBD graphs match those in sparse Erdős-Rényi graphs, these observations go beyond the theory that we have presented.

The impact of complex demography

Figure S11 shows that the apparent normality of some graph features extends to the three phases of exponential growth and population bottleneck demographic scenarios. Table S1

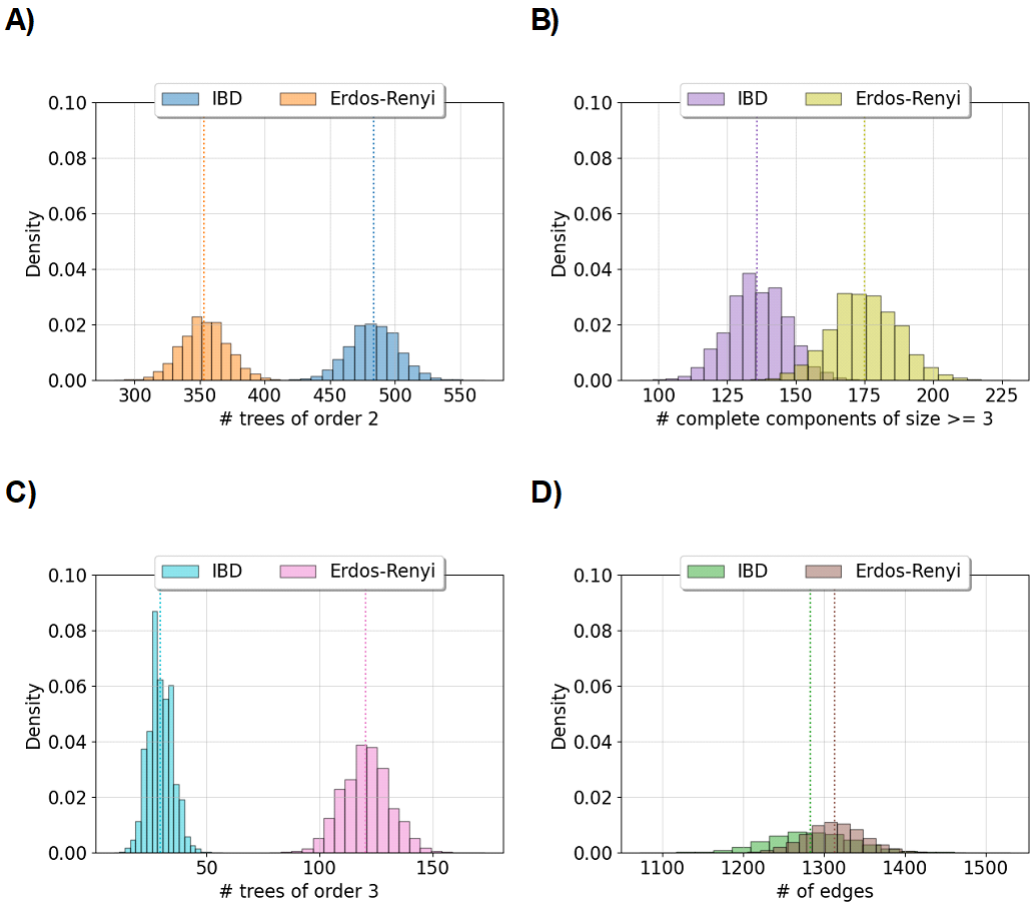

Figure 3.5: Comparing features between IBD and Erdős-Rényi graphs. Histograms compare the density of graph features between IBD and Erdős-Rényi graphs. Each histogram summarizes the results of one hundred and twenty-five thousand simulations. A) and C) show the number of trees of order 2 and 3, respectively. B) shows the number of complete components with more than three nodes. D) shows the total number of edges. The legends give colors assigned to the IBD and Erdős-Rényi graphs. IBD graphs are simulated using the constant one hundred thousand diploids demography and the 0.03 Morgan length threshold. Erdős-Rényi graphs are simulated using the same success probability as in the IBD graph. The sample size is two thousand diploids. Vertical dotted lines show the means.

reports that the proportions of rejected hypothesis tests for numbers of trees of order 2 are close to 0.05 for both demographic scenarios. We also fail to reject normality for the number of trees of order 3 and the number of connected components of degree 3 or more in some simulations of the three phases of exponential growth scenario. These results indicate the limiting distributional behaviors of graph features in detectable IBD graphs around a locus can be similar for large constant-size populations and demographic scenarios with large recent population sizes.

The impact of positive selection

Strong directional selection increases the IBD rate [148], but less is known about how this phenomenon alters the feature distributions of detectable IBD graphs. We conduct more simulations of detectable IBD graphs for selection coefficients between 0.01 and 0.04 and the three phases of exponential growth and population bottleneck scenarios. Tables S2 and S3 demonstrate multiple trends as the selection coefficient increases. The apparent normality of the number of trees of order 2 does not noticeably change as we change the selection coefficient, and we reject normality less often for the number of trees of order 3 and the number of complete components of order 3 or more. It may be that the distributional behaviors of these small degree connected components become clearer under the selection models with more detectable IBD segments. The main effect of strong positive selection appears to be the growth of a largest detectable IBD cluster around a locus that includes haplotypes with a beneficial allele. This idea is a major motivation for the suite of methods we develop in Chapter 5.

3.5 Discussion

In this chapter, we prove a central limit theorem for the detectable IBD rate around a locus whose regularity conditions have intuitive interpretations in population genetics. The sample size squared must be large enough such that there are many IBD segments long enough to be accurately detected by existing methods. The population size must be large enough such

that there are no large IBD clusters around a locus. This central limit theorem is a seminal contribution to this thesis, and its impacts will echo through all chapters.

The conceptual framework for our asymptotic conditions involves envisioning a coalescent tree where the internal branches are long but numerous coalescent events occur near the leaves. The internal branches are long because of the large population size. There are numerous coalescent events near the leaves because of the large sample size. The large Morgan threshold serves to further decrease the probability of a detectable IBD segment and the correlations between IBD segment lengths.

Many authors make observations about statistics in population genetics that appear visually similar to known parametric families. Tajima [145] says that the sampling distribution of their D statistic resembles a Beta distribution. Field et al. [51] and Palamara et al. [111] say that the sampling distributions of their selection scan statistics resemble Gamma distributions. Related to our work, Carmi et al. [29] say that the average amount of the genome that an individual shares identity-by-descent with others in a sample resembles a normal distribution. The difference between these works and our work is that we have an exact mathematical result, not an empirical observation.

We employ simulation to evaluate the assumptions and validity of our central limit theorem. Consistent with our conditions, we reject a null hypothesis of normality less often as sample size and scaled population size increase. In practice, we find that non-normality is typical in finite samples. We indicate that non-negligible covariance may come from the agglomeration of IBD clusters. We also discuss this phenomenon in the context of a similar network model for which certain theoretical properties have been established.

Our regularity conditions concern a balance between sample size and scaled population size that is unlikely to hold in practical settings. We advocate that the collected sample size should always be as large as is feasible and that the smallest Morgan length threshold should be chosen for which IBD segment detection is accurate. While non-normality of the detectable IBD rate may lead to anti-conservative control of FWER in the selection scans of Browning and Browning [20] and Temple et al. [148], we observe no trend between this

anti-conservative behavior and sample size.

Together our theoretical results and simulation studies support ongoing methodological developments based on IBD segments. For instance, the selection coefficient confidence intervals we propose in Chapter 4 should give proper coverage of the selection coefficient if: 1) the detectable IBD rate is approximately normally distributed in scenarios of positive selection, and 2) the Delta method conditions hold for the function that maps the detectable IBD rate to the selection coefficient [31]. Additionally, existing genome-wide scans for excess IBD rates lack formal or exact hypothesis testing frameworks. The IBD-based selection scans in Browning and Browning [20] and Temple et al. [148] use a heuristic threshold of four standard deviations to determine significance. In large samples, the Ornstein-Uhlenbeck process could be a reasonable null model in selection scans and rare variant association testing. Specifically, detectable IBD rates overlapping focal locations may be assumed to be normally distributed, and their correlation is a consequence of the same IBD segments overlapping spatially adjacent locations on a chromosome. In Chapter 6, we expand upon this idea to derive a genome-wide significance threshold for IBD-based selection scans.

Chapter 4

SELECTION COEFFICIENT ESTIMATION

4.1 Introduction

Positive natural selection favors alleles advantageous for survival or reproduction. The process of evolution via positive selection is often modeled via hard or soft selective sweeps. In hard sweeps, an allele increases so rapidly in frequency such that all prevailing haplotypes trace their origin to a single haplotype. In soft sweeps, a beneficial mutation on multiple different genetic backgrounds or multiple independent mutations with similar effects may increase in frequency [76, 114, 115]. In either case, sweeps can come from existing alleles that become newly beneficial in the context of environmental change, referred to as a sweep from standing variation, or they can come from *de novo* mutation. The hard sweep is partial or incomplete if a selected allele is not fixed in the present day. Alleles on the same haplotype as a sweeping allele also increase in frequency until crossover recombination separates them from the sweeping allele [139].

The popular mathematical model to study a hard selective sweep from a single allele frames the problem in terms of a selection coefficient s. This parameter measures the relative advantage that a sweeping allele A has over alternative alleles B. In the haploid model, this ratio is $(1+s):1$. In diploids, the ratios concern genotypes $AA:AB:BB$ for the following models of genic selection:

- Multiplicative $(1+s)^2:(1+s):1$
- Additive $(1+2s):(1+s):1$
- Dominance $(1+2s):(1+2s):1$

- Recessive $(1+2s):1:1$

With the heterozygosity coefficient h, the additive, dominance, and recessive models can be written as $(1+2s):(1+2hs):1$ for h equal to $1/2, 1$, and 0, respectively. Selection coefficients from different genetic models are not comparable. Throughout this text, we restrict to the case of positive selection $s \geq 0$.

Using these ratios, we can express the probability of a sweeping allele frequency at generation $t-1$, denoted $p_s(t-1)$, in terms of the probability of a sweeping allele frequency at generation t and the selection coefficient s. For the haploid model, we have the formula

$$p_s(t-1) := \frac{p_s(t) \times (1+s) \times 1}{(1+s) \times p_s(t) + 1 \times (1-p_s(t))} = p_s(t) \times \frac{1+s}{1+p_s(t) \times s}. \tag{4.1}$$

Equation 4.1 represents the effect of positive selection on the sweeping allele frequency one generation forward in time. The numerator is the frequency of the A allele times its relative fitness and the probability that an A parent passes on an A allele to its offspring. The denominator is the frequencies of the sweeping alleles A and B weighted by $(1+s):s$, referred to as the mean relative fitness. As done in Crow and Kimura [38] and Felsenstein [48], we simplify this formula in terms of $p_s(t)$ multiplied by a fraction. When $p_s(t)$ and s are both small, we approximate that $p_s(t-1) \approx p_s(t) \times (1+s)$, which we interpret as the probability of the sweeping allele increases by a factor of $1+s$ on average. For the multiplicative diploid model, we have the formula

$$\begin{aligned} p_s(t-1) &:= \frac{[p_s(t)^2 \times (1+s)^2 \times 1] + [2 \times p_s(t) \times (1-p_s(t)) \times (1+s) \times 0.5]}{p_s(t)^2 \times (1+s)^2 + 2 \times p_s(t) \times (1-p_s(t)) \times (1+s) + (1-p_s(t))^2} \\ &= p_s(t) \times \frac{(1+s)(1+s \times p_s(t))}{[p_s(t) \times (1+s) + 1 - p_s(t)]^2} \\ &= p_s(t) \times \frac{1+s}{1+p_s(t) \times s}, \end{aligned} \tag{4.2}$$

The numerator is the weighted genotype frequencies AA and AB multiplied by probabilities 1 and 0.5 that they pass an A allele to their offspring. The denominator is the genotype frequencies $AA:AB:BB$ weighted by $(1+s)^2:(1+s):1$. The multiplicative diploid model is equivalent to the haploid model once simplified. Formulas for the haploid, multiplicative,

additive, dominance, and recessive models have been well known in population genetics (see Fisher [52], Haldane [68, 69], and Wright [159]).

Equations 4.1 and 4.2 describe the trajectory of the sweeping allele forward in time. To derive the probability of the sweeping allele at generation t in terms of the probability of the sweeping allele one generation ago $t-1$ and the selection coefficient s, we invert $p_s^{-1}(t-1)$ in $p_s(t)$, holding the selection coefficient s constant. For the haploid model, and equivalently the multiplicative diploid model, the formula for one generation prior is

$$p_s(t) := p_s(t-1) \times \frac{1}{1 + (1 - p_s(t-1)) \times s} \tag{4.3}$$

Inverting the formulas for the additive, dominance, and recessive models leads to a quadratic formula in $p_s(t-1)$. Of the two real-valued solutions to these quadratic formulas, one is a probability value in [0,1], and the other is not. In our work, we use the solution $p_s(t) \in [0,1]$.

In the estimation of the selection coefficient s, we use the deterministic formulas. However, the deterministic formulas $p_s(t)$ in terms of $p_s(t-1)$ and s concern the mean behavior of the Wright-Fisher process (WF) with genetic drift and selection. When quantifying uncertainty due to drift, at each prior generation t, we estimate $\hat{p}_s(t) = n_A(t)/N_e(t)$ with respect to the effective population size $N_e(t)$ and the number of alleles A drawn from $n_A(t) \sim \text{Binomial}(N_e(t), \hat{p}_s(t-1))$.

The magnitude of the selection coefficient s impacts the historical changes in the frequency of the sweeping allele. In Figure 4.1, we illustrate this effect by simulating fifty bootstraps from the WF process with varying present-day sweep frequencies and selection coefficients. After accounting for genetic drift, the differences between WF processes with strong selection $s \geq 0.01$ and $0.15 \leq p(0) \leq 0.85$ are apparent within the past twenty generations. When an allele is near fixation ($p(0) = 0.98$), the allele frequency changes in the past twenty generations are slight, even for s as high as 0.05. When the sweep is very new ($p(0) = 0.05$), the variance in genetic drift dominates the expected value increases in allele frequency, even for large $s = 0.05$. In humans of European ancestry, there is some evidence of selection of these magnitudes at the LCT locus [75, 100, 103, 132, 142, 148, 152].

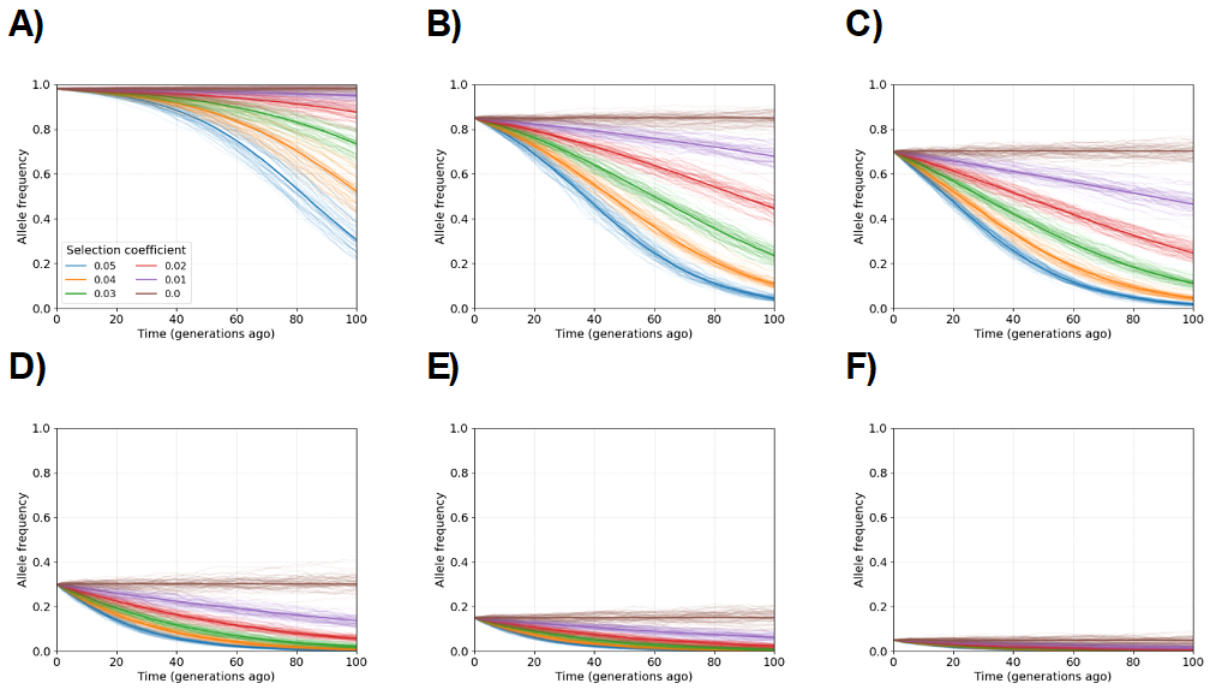

Figure 4.1: Wright-Fisher simulations with strong varying selection coefficeints. Fifty simulations of the Wright-Fisher process for $s \in [0, 0.01, 0.02, 0.03, 0.04, 0.05]$. Allele frequency ($y$-axis) by time ($x$-axis) is shown for the most recent one hundred generations. The allele frequencies in the present day are (A) 0.98, (B) 0.85, (C) 0.7, (D) 0.3, (E) 0.15, and (F) 0.05. The legend in A) defines colors for different selection coefficients and applies to all subplots.

Figure S14 shows fifty bootstraps of the WF process with varying present-day sweep frequencies and $s \leq 0.004$. Within the past one hundred generations, these time series are not distinguishable from the neutral evolution setting of $s = 0.000$. Within the past five hundred generations, WF simulations of $s = 0.004$ are distinguishable from WF simulations of $s = 0.000$, but some WF simulations of $s \leq 0.002$ are still indistinguishable from WF simulations of $s = 0.000$. It is widely believed that selection coefficients of these magnitudes are the predominant examples of selective sweeps in humans.

Here we propose an estimator and confidence intervals for the selection coefficient that

are based on identity-by-descent segment lengths (IBD) among samples within a population. The hard sweep model concerns changes in allele frequency over time, making estimation from only genetic data in the current generation challenging. We consider our approach applicable to recent and ongoing sweeps, whereby "recent" we mean selection within the last three hundred generations, the time period in which long IBD segments can be accurately detected (Figures S4, S5 and 2.4, Tian et al. [150]). In Section 4.2, we formally state our estimator and its confidence interval and discuss the statistical properties of estimators. In Section 4.3, using true IBD segment lengths drawn from Algorithm 1 (Chapter 2, Section 2.4), we explore the accuracy of our inference in simulations of correct model specification and various examples of model misspecification. In Section 4.4, we explore the effects of sample size, IBD segment lengths, and sweep frequency which are pertinent to real data analysis. In Section 4.2.2, we conduct extensive simulations to estimate the coverage of our confidence intervals. Additional simulation studies concerning selection coefficients $s \geq 0.10$ (Section 4.5) and comparison to a method that uses time series data (Section B.3) are provided in Appendix B. We conclude with a discussion on the contribution of our estimation procedure.

4.2 Selection coefficient estimator based on identity-by-descent segments

4.2.1 Point estimation

Three important statistical properties of estimators are unbiasedness, sufficiency, and consistency. Formal definitions of these properties are in Casella and Berger [31]. Let $\hat{s}$ be an estimator of the selection coefficient. Unbiasedness means that the expected value of the estimator $\hat{s}$ is equal to the true value s. Sufficiency means that the estimator $\hat{s}$ captures all the relevant information about s that is available in the data. Consistency means the probability that the estimator $\hat{s}$ gets arbitrarily close to s goes to 1 as we collect more data. In this chapter, we focus on these properties in mathematical exposition and empirical results.

For sample haplotypes a and b, recall that $Y_{a,b} := I(W_{a,b} \geq w)$ is the indicator random variable that their IBD segment $W_{a,b}$ overlapping a focal location exceeds some Morgan length

threshold w (Chapter 3, Section 3.2). The IBD segment indicators $\{Y_{a,b}\}$ for all haplotype pairs are identically distributed Bernoulli random variables with probability $P_s(W_{a,b} \geq w)$ and correlations that depend on unobserved coalescent tree and recombination processes. The notation P_s generalizes our work to selective sweep models around a locus.

Let T be the coalescent time of two specific haplotypes, T_{AA} be the coalescent time of two specific haplotypes carrying a sweeping allele A, and T_{BB} be the coalescent time of two specific haplotypes carrying an alternative allele B. The probability mass functions for these random variables are

$$\begin{aligned} P_s(T_{AA} = t) &= \prod_{\tau=1}^{t-1} \left(1 - \frac{1}{N_e(\tau) \times p_s(\tau)}\right) \times \frac{1}{N_e(t) \times p_s(t)}; \\ P_s(T_{BB} = t) &= \prod_{\tau=1}^{t-1} \left(1 - \frac{1}{N_e(\tau) \times q_s(\tau)}\right) \times \frac{1}{N_e(t) \times q_s(t)}; \\ P(T = t) &= p(0)^2 \times P(T_{AA} = t) + q(0)^2 \times P(T_{BB} = t), \end{aligned} \tag{4.4}$$

where effective sizes $N_e(t)$ can take on a general form, $p_s(t)$ is as defined in Section 4.1, and $q_s(t) := 1 - p_s(t)$. Conditional on the coalescent time, IBD segment lengths to the right and left of a sweeping allele are exponentially distributed with rate $2t$ (Section 2.5.2). To compute the probability $P_s(W_{a,b} \geq w)$ that an IBD segment overlapping a focal location exceeds w Morgans, we integrate over the IBD segment length $W_{a,b}|T$ and the coalescent time T. The empirical tail probability

$$\hat{P}_s(W_{a,b} \geq w) := f(n) \times \sum_{a \neq b} Y_{a,b} = \bar{Y}_{a \neq b} \tag{4.5}$$

is the number of detected IBD segments divided by the total number of haplotype pairs $f(n)$. For n haploids, $f(n) = \binom{n}{2}$; for n diploids, $f(n) = \binom{2n}{2} - 2n$. The sample mean $\bar{Y}_{a \neq b}$ is an unbiased estimator of $P_s(W_{a,b} \geq w)$, and it maximizes the composite likelihood where IBD segment indicators are independent. Under certain conditions, maximum likelihood estimators (MLEs) are sufficient and consistent [31]. Maximum composite likelihood estimators typically give consistent point estimates in statistical genetics [94].

We propose the selection coefficient estimator $\hat{s}$ that solves the equation $P_{\hat{s}}(W_{a,b} \geq w) = \bar{Y}_{a \neq b}$. Because inverting $P_s(W_{a,b} \geq w)$ is analytically intractable, we optimize the expression

$$\hat{s} := \arg\min_s (\bar{Y}_{a \neq b} - P_s(W_{a,b} \geq w)). \tag{4.6}$$

The estimator is conditional on prior knowledge of the sweep frequency $p(0)$ and effective population sizes $N_e(t)$. It is also a function of an unbiased estimator that maximizes the composite likelihood for IBD segment indicators $\{Y_{a,b}\}$[1].

In implementation, we allow the optimization to be over IBD segment indicators discretized into bins based on their lengths. In Appendix B, we explore selection coefficient estimation using the number of detectable IBD segments versus a discretized length distribution of detectable IBD segments (Section B.1). These additional simulation studies evaluate if a discretized length distribution improves estimation compared to merely the number of detectable IBD segments. The number of detectable IBD segments is a simpler statistic to interpret than the length distribution.

If $P_s(W_{a,b} \geq w)$ is monotone in the selection coefficient s, then the inverse function mapping $P_s(W_{a,b} \geq w)$ to s is one-to-one. For general positive-valued $N_e(t)$, we cannot show that $P_s(W_{a,b} \geq w)$ is monotone. However, the probability $P_s(W_{a,b} \geq w)$ of IBD segments being at least w long around a focal point is related to the probability of alleles being IBD, and Albrechtsen et al. [3] have proven that the probability of IBD alleles is nondecreasing in s when positive selection acts on an allele. Showing monotonicity is difficult because of the additive terms in Equation 4.4. Heuristically, it is straightforward to show that the geometric failure probabilities

$$1 - \frac{1}{N_e(t) \times p_{s_1}(t)} \geq 1 - \frac{1}{N_e(t) \times p_{s_2}(t)}$$

for $s_1 \leq s_2$. This inequality leads to the inequality $P_{s_1}(T = t) \leq P_{s_2}(T = t)$, and then we argue that more recent coalescent times for haplotypes carrying the sweeping allele leads to more detected IBD segments. In Figure S15, we verify the monotonicity of $P_s(W_{a,b} \geq w)$ in

[1] If $\hat{\eta}$ is an MLE of parameter η, then functions $f(\hat{\eta})$ are MLEs of $f(\eta)$ [31].

s for some examples of complex demography. Overall, the larger the true selection coefficient s is, the more detected IBD segments there are expected to be.

4.2.2 Confidence intervals

Confidence intervals, and analogous interval estimators in Bayesian statistics, nonparametric statistics, and machine learning, quantify the uncertainty of an estimator in a principled way. Given a significance level α, the ideal property is that $(1-\alpha)\%$ interval estimators contain the true parameter with probability $1-\alpha$. As an example, the interval estimator of the mean μ from a sample of $N(\mu, \sigma)$ random variables is the average plus or minus the sample standard deviation times the $\alpha/2$ quantile of the standard normal distribution.

We propose a parametric bootstrap approach to estimate $(1-\alpha)\%$ confidence intervals. Using Algorithm 1, we simulate detectable IBD segment lengths from a model with selection coefficient estimate $\hat{s}$. For B bootstraps, we use Equation 4.6 to get $\hat{s}_{1:B} := (\hat{s}_1, \ldots, \hat{s}_B)$ estimates. Next, we calculate the sample standard deviation $\hat{\sigma}_{1:B}(\hat{s})$ of the selection coefficient estimates $\hat{s}_{1:B}$. Finally, our interval estimator for the selection coefficient is $\hat{s} \pm z_{\alpha/2} \times \hat{\sigma}_{1:B}(\hat{s})$, where $z_{\alpha/2}$ is the standard normal quantile of $\alpha/2$. In essence, our approach is analogous to the interval estimator for the mean of a sample of normally distributed random variables.

Under certain regularity conditions, MLEs are asymptotically normally distributed [31], which is a useful theoretical result in calculating the asymptotic coverage probability. Recall in Theorems 3.3.1, 3.3.2, and 3.3.3 that the detectable IBD rate overlapping a focal point is normally distributed under certain asymptotic conditions. For functions of asymptotically normally distributed statistics that satisfy the Delta method conditions[2], functions of these statistics themselves are asymptotically normally distributed [31]. The appeal of our parametric bootstrap approach is that $\hat{s}$ is asymptotically normally distributed if the IBD rate is asymptotically normally distributed for models P_s and the Delta method applies to the function mapping $P_s(W_{a,b} \geq w)$ to s. The Delta method condition that there is a

[2]The function has a first derivative at the parameter value that is non-zero.

non-zero first derivative for the function mapping $P_s(W_{a,b} \geq w)$ to s appears to hold in the demographic scenarios explored here (Figure S15). The need for many bootstrap selection coefficient estimates $\hat{s}_{1:B}$ is not onerous because of the speed of Algorithm 1 (Figure 2.3).

Bias correction is a common technique used in bootstrap methods [98]. We do not implement bias correction in our approach because our optimization procedure (Equation 4.6) can be unstable for s close to zero. This instability is due to the probability distribution of the detectable IBD rate around a locus being similar between the neutral model $s = 0.00$ and s close to zero, which is reflected in Figure S14.

4.3 Estimation under correct and incorrect model specifications

In simulation study, we assess the empirical properties of our selection coefficient inference. Over all simulations, we calculate the average selection coefficient estimate, the average width of ninety-five percent confidence intervals[3], and the empirical coverage of ninety-five percent confidence intervals. The average selection coefficient estimate is a proxy measurement for unbiasedness, which is not theoretically guaranteed for our estimator (Jensen's inequality [31]). Width is the (absolute) difference between the left and right bounds of the confidence interval. The empirical coverage is the percent of simulations in which the interval estimate contains the true value[4]. Empirical coverages close to ninety-five percent give evidence to the validity of our confidence intervals. Additionally, for fixed coverage, it is desirable to have smaller confidence interval widths.

For each unique combination of model parameters, we calculate these metrics aggregated over ten thousand simulations. Default simulation settings are $s = 0.030, p(0) = 0.50$, multiplicative selection, and the population bottleneck (BN) demographic model (Section 2.5.1). The sweep is incomplete and from *de novo* mutation. The sample size is five thousand diploids. Unless otherwise specified, selection coefficient estimates are based on IBD segments greater than 3.0 cM simulated from Algorithm 1.

[3]By default, confidence intervals are computed from fifty parametric bootstraps

[4]And it is an estimator of the probability that the interval estimator contains the true value

4.3.1 Correct model specification

First, we estimate the selection coefficient conditional on the true allele frequency $p(0)$ and effective population sizes $N_e(t)$. We consider the following choices: $s \in [0.02, 0.03, 0.04]$; $p\,(0) \in [0.25, 0.50, 0.75]$; and population bottleneck (BN), three phases of exponential growth (G3), or constant population of twenty-five thousand individuals (C25) demographic scenarios (Section 2.5.1). We specify the perturbed model parameter for each row in Table 4.1.

Table 4.1 shows inference results across a range of scenarios, conditional on correctly specified $p(0)$ and $N_e(t)$. For the BN demographic model and $p\,(0) = 0.5$, average estimates over ten thousand simulations are roughly equal to the true $s \in [0.02,\ 0.03,\ 0.04]$. For $s = 0.03$, the estimator is likewise accurate if allele frequency $p\,(0) \in [0.25,\ 0.75]$ or under alternative demographic models. Average widths of ninety-five percent confidence intervals are less than 0.01 for all experiments. Moreover, the ninety-five percent confidence intervals contain the true selection coefficient roughly ninety-five percent of the time. Overall, in these simulations, our selection coefficient estimator has empirical behavior that is anticipated of a maximum likelihood estimator.

4.3.2 Weak model misspecification

Now, we estimate the selection coefficient conditional on model misspecifications: allele frequency or $N_e(t)$ are different from the simulations, or sweeps come from variants at one, two, and five percent frequency (abbreviated SV for standing variation). We specify the true simulated model and the model assumed for estimation for each row in Table S4. As an example, in the first row of Table S4, we simulate true IBD segments from a sweeping allele at forty percent in the present day, but we estimate the selection coefficient conditional on an allele frequency of fifty percent. For misspecified demography, 1.1x and 0.9x denote uniform scaling of effective population sizes $N_e(t)$. These scalar multipliers represent weak model misspecification but serve to indicate the non-identifiability of s and $N_e\,(t)$ around a single locus (Equation 4.4). (In data analysis, recent effective population sizes should be

Scenario	**True** s	**Estimate** $\bar{\hat{s}}$	**Width**	**Coverage**
$s = 0.020$	0.0200	0.0202	0.0085	97.7%
$s = 0.030$	0.0300	0.0300	0.0085	95.2%
$s = 0.040$	0.0400	0.0401	0.0099	95.1%
$p(0) = 0.25$	0.0300	0.0300	0.0087	95.3%
$p(0) = 0.50$	0.0300	0.0300	0.0066	94.9%
$p(0) = 0.75$	0.0300	0.0300	0.0058	95.2%
$N_e(t) = \text{BN}$	0.0300	0.0300	0.0066	95.3%
$N_e(t) = \text{G3}$	0.0300	0.0300	0.0062	95.0%
$N_e(t) = \text{C25}$	0.0300	0.0300	0.0066	95.8%

Table 4.1: Selection coefficient estimation based on true IBD segments and correct model specification. Summary statistics are the mean estimate, the average confidence interval width, and the percentage of confidence intervals containing the true selection coefficient (coverage). Ninety-five percent confidence intervals are based on fifty bootstraps. Estimation is conditional on IBD segments longer than 3.0 cM and known allele frequency. Each row represents the results of 10,000 replicates. Default settings are $s = 0.03$, $p(0) = 0.50$, a population bottleneck demography, the multiplicative genetic model, and an ongoing sweep from a new beneficial mutation. Demographic scenarios considered are population bottleneck (BN), three phases of exponential growth (G3), and constant N_e=25000 (C25). The sample size is five thousand diploids.

inferred using separate methods [19, 110, 53, 80].)

Table S4 shows inference results conditional on the aforementioned model misspecifications. If the allele frequency used for estimation differs from the true value by 0.10, the average estimates can be slightly biased. This bias affects the coverage of confidence intervals; nevertheless, the confidence intervals in these simulations retain at least ninety percent coverage. Effective sizes misspecified by a scalar factor affect bias and coverage similarly.

For example, to fit IBD counts to a misspecified model with elevated effective sizes, a greater selection coefficient is required, and vice versa for lower effective sizes. Interestingly, our selection coefficient estimator is robust to hard sweeps from low-frequency standing variation more than one hundred generations ago. We suspect this result is because IBD segments $\geq$ 3.0 cM mainly come from common ancestors within the last one hundred generations.

4.3.3 Time-varying selection coefficients

The estimator we propose is for a single selection coefficient s. Recent work by Mathieson and Terhorst [100] and Vaughn and Nielsen [152] provide evidence at human LCT and SLC45A2 loci in favor of a sweep model with time-varying selection coefficients. In our work, time-varying selection coefficients is a strong model misspecification.

Here we consider the impacts on our selection coefficient estimator in the face of a model with time-varying selection coefficients. Let $\mathbf{s}_{[0,\infty)} = (s_0, s_1, s_2, \ldots, s_\infty)$ be a vector of selection coefficients at each coalescent generation. Denote $\mathbf{s}_{[\tau_1,\tau_2)}$ as the selection coefficient between generation τ_1 to $\tau_2 - 1$ inclusive. Our simulations fall into two types: there is a single change point in the selection coefficient vector $\mathbf{s}_{[0,\infty)}$, or there are two selection coefficients oscillating between time periods. In both simulation types, the present-day sweeping allele frequency is 0.65, the constant population size is ten thousand diploids, the sample size is one thousand diploids, and the IBD segment length threshold is 3.0 cM. The sweeping allele frequency of 0.65 is close to the reported frequencies of the putatively selected allele at LCT in northern Europeans.

The change point simulation study involves change points twenty-five and fifty generations ago and selection coefficients of 0.02 and 0.05. Specifically, we have four scenarios: $\mathbf{s}_{[0:25)} = 0.05$ and $\mathbf{s}_{[25:\infty)} = 0.02$ (Figure S16A), $\mathbf{s}_{[0:50)} = 0.05$ and $\mathbf{s}_{[50:\infty)} = 0.02$ (FigureS16B), $\mathbf{s}_{[0:25)} = 0.02$ and $\mathbf{s}_{[25:\infty)} = 0.05$ (Figure S16C), and $\mathbf{s}_{[0:50)} = 0.02$ and $\mathbf{s}_{[50:\infty)} = 0.05$ (Figure S16D). We estimate a single selection coefficient in two hundred parametric bootstraps, and for each selection coefficient estimate we simulate a WF process back in time. We also simulate two hundred WF processes using the true time-varying selection coefficients. We plot and discuss

the median and percentiles of these bootstrap trajectories.

Figure S16 shows the historical allele frequencies of the sweeping allele for inference in the four scenarios. In the models where $\mathbf{s}_{[0:25)} = 0.02$ or $\mathbf{s}_{[0:50)} = 0.02$, we observe an overall trend that the inferred historical allele frequencies are less than the true historical allele frequencies before and slightly after the change point. In the models where $\mathbf{s}_{[0:25)} = 0.05$ or $\mathbf{s}_{[0:50)} = 0.05$, we observe an overall trend that the inferred historical allele frequencies are greater than the true historical allele frequencies before and slightly after the change point. In all models, the inferred historical allele frequencies approach the true historical allele frequencies tens of coalescent generations after the change point.

The selection coefficient estimates are between the two time-varying varying selection coefficients, but closer to the more recent selection coefficient. In the models where $\mathbf{s}_{[0:25)} = 0.02$ or $\mathbf{s}_{[0:50)} = 0.02$, we estimate a single selection coefficient $\hat{s}$ that is on average less than the midpoint 0.035. In the models where $\mathbf{s}_{[0:25)} = 0.05$ or $\mathbf{s}_{[0:50)} = 0.05$, we estimate a single selection coefficient $\hat{s}$ that is on average greater than the midpoint 0.035.

The oscillations simulation study involves change points every twenty generations that switch between selection coefficients of 0.02 and 0.05. Specifically, we use the model $\mathbf{s}_{[0,20)} = 0.02$, $\mathbf{s}_{[20,40)} = 0.05$, $\mathbf{s}_{[40,60)} = 0.02$, $\mathbf{s}_{[60,80)} = 0.05$, and $\mathbf{s}_{[80,\infty)} = 0.02$. Estimation and bootstrapping are the same as in the one change point simulation study. Figure S17A shows the inferred and true historical allele frequencies for this oscillation model. Overall, we observe inferred historical allele frequencies less than or greater than the true historical allele frequencies at the change points when selection coefficients change from 0.02 to 0.05 or from 0.05 to 0.02, respectively.

We also consider a similar oscillation model except the first selection coefficient $\mathbf{s}_{[0,20)}$ is 0.05: $\mathbf{s}_{[0,20)} = 0.05$, $\mathbf{s}_{[20,40)} = 0.02$, $\mathbf{s}_{[40,60)} = 0.05$, $\mathbf{s}_{[60,80)} = 0.02$, and $\mathbf{s}_{[80,\infty)} = 0.05$. The histogram in Figure S17B shows the selection coefficient estimates for the two oscillation models. Overall, we observe selection coefficient estimates less than or greater than the midpoint 0.035 when the first selection coefficient $\mathbf{s}_{[0,20)}$ is 0.02 or 0.05, respectively.

As expected, our IBD-based selection coefficient estimator is informed more by allele

frequency changes in recent generations as opposed to allele frequency changes in distant generations. Both the change point and oscillation simulation studies arrive at this finding. These experiments indicate that our estimates of a single selection coefficient "average", in a loose sense, over time-varying selection coefficients, with selection coefficients in the recent generations assigned more weight than those in the distant generations.

4.4 The effects of important model parameters on estimation

Beyond studying statistical properties like coverage in large-scale simulations, we can use smaller-scale simulations to explore the effects of sample size, the IBD segment length threshold, and the sweeping allele frequency on selection coefficient estimation.

4.4.1 Simulation setup

In these analyses, we consider individuals sampled from a population of constant effective size ten thousand diploids. This demographic scenario closely mirrors recent effective population sizes inferred for *Pongo abelii* (Bornean or Sumatran orangutan) [97], *Pan troglodytes* (Western chimpanzee) [91], and *Canis familiaris* (domestic dog) [96] before breed creation (stdpopsim catalog accessed April 30, 2024) [1, 95].

Unless otherwise specified, our default settings are a sweeping allele frequency $p(0) = 0.5$, a detection threshold of 3.0 cM, and a sample size of four hundred diploids. We perturb these parameters one at a time, holding the other parameters constant. For each unique combination of model parameters, we conduct two hundred simulations. Analyses of this scope and with these sample sizes can be performed on a personal laptop in a few minutes of runtime.

4.4.2 The effect of sample size on estimation

The rate of detectable IBD segments depends on the effective population sizes. With a constant effective size of ten thousand individuals, we determine an empirical lower bound

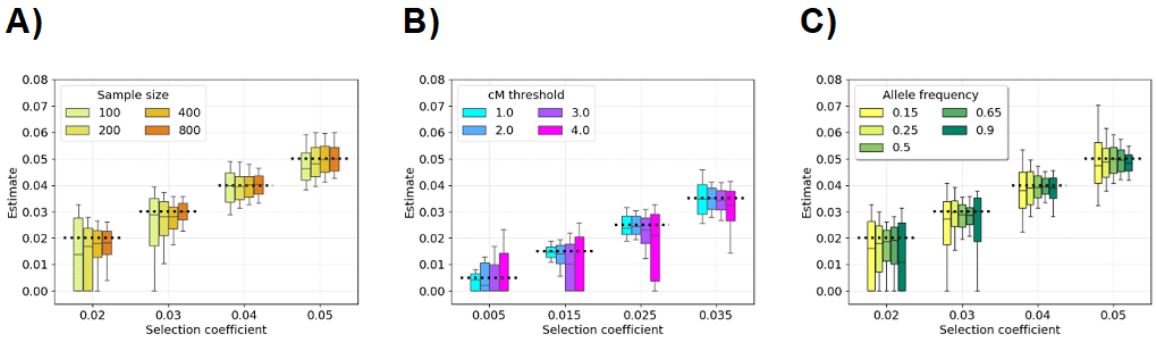

Figure 4.2: The effects of sample size, detection threshold, and allele frequency on selection coefficient estimation. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of selection coefficient estimates for each setting: (A) sample sizes of one, two, four, and eight hundred diploids, sweeping allele frequency of fifty percent, and 3.0 cM length threshold; (B) 1.0, 2.0, 3.0, and 4.0 cM length thresholds, sample size of four hundred diploids, and sweeping allele frequency of fifty percent; and (C) sweeping allele frequencies of fifteen, twenty-five, fifty, sixty-five, and ninety percent, sample size of four hundred diploids, and 3.0 cM length threshold. Horizontal dotted lines show the true selection coefficients. There are two hundred simulations for each combination of parameters. The constant population size is $N_e = 10,000$.

on sample size that is necessary for reliable estimation of $s \geq 0.02$. Figure 4.2A shows selection coefficient estimates $\hat{s}$ for $s \in [0.02, 0.03, 0.04, 0.05]$and sample sizes of one, two, four, and eight hundred diploids. When $s = 0.02$ and the sample size is one hundred, more than fifty percent of estimates are non-zero. Across all s values and sample sizes, the median estimates are within 0.01 of the true selection coefficients and the interquartile ranges contain them. The differences between median estimates and the actual s values, as well as the interquartile range of estimates, decrease as sample sizes increase. The trend that absolute deviations between estimates and the truth decrease as sample size increases is an empirical behavior expected of a consistent estimator. This *in silico* analysis can be

replicated with various demographic scenarios to establish a lower bound on the sample size that is required for our estimation method. For instance, in additional simulation studies (Section B.2, Figure S23), we show that estimates of very large selection coefficients $s \geq 0.1$ may be accurate in sample sizes as small as fifty.

4.4.3 The effect of IBD segment detection threshold on estimation

Longer IBD segments are more likely to trace back to more recent common ancestors than are shorter IBD segments, and hence segment lengths can serve to calibrate the time depth of IBD [149]. Figure 4.2B shows selection coefficient estimates $\hat{s}$ for $s \in [0.005, 0.015, 0.025, 0.035]$ and length thresholds 1.0, 2.0, 3.0, and 4.0 cM. Using a 1.0 cM threshold, we observe median estimates that are within 0.001 of the true selection coefficients, even for s as small as 0.005. We observe that the tenth percentile of estimates is zero for $s \leq 0.025$ when using segments longer than 4.0 cM. This length threshold is too stringent to provide much information about the recent genetic history. With a threshold of 1.0 or 2.0 cM, the interquartile ranges of estimates contain $s \geq 0.015$, and the tenth percentiles of estimates exclude zero. With a threshold of 3.0 cM, the median and seventy-fifth percentile of estimates are non-zero for $s = 0.015$ and $s = 0.005$, respectively, but the twenty-fifth percentiles are zero. The ability to model weaker hard sweeps improves when we use true IBD segments at a lower detection threshold. This *in silico* analysis can be replicated in other studies to calibrate the magnitudes of hard sweeps that our method can infer given reliable detection of IBD segments greater than some length threshold.

4.4.4 The effect of sweep frequency on estimation

Changes in allele frequency, especially between generations closest to the present day, impact the number of detectable IBD segments. Figure 4.2C shows estimates $\hat{s}$ for allele frequencies $p(0) = [0.15, 0.3, 0.5, 0.7, 0.85]$ and $s \in [0.02, 0.03, 0.04, 0.05]$. Across all s values and allele frequencies, the median estimates are within 0.005 of the true selection coefficients. When $s \geq 0.03$, the interquartile ranges exclude zero and contain the true selection coefficients.

The difference between the ninetieth and tenth percentiles is uniformly greatest for the low sweep frequency $p(0) = 0.15$. When $s = 0.02$ and $p(0) = 0.15$ or $p(0) = 0.85$, the lower quartile is zero. We can extrapolate that sweeps where $p(0) \geq 0.85$ or $p(0) \leq 0.15$ will be more difficult to model correctly with our IBD-based estimator. These results relate to the WF processes in Figure 4.1 and S14. Rapid allele frequency changes in the past few hundred generations are consistent with more IBD segments derived from a recent adaptive haplotype, which is the signal our estimator is designed for. This *in silico* analysis can be replicated in other studies to determine if our method can estimate very new sweeps or sweeps near fixation. For instance, in additional simulation studies (Section B.2, Figure S24), we show that estimates of very large selection coefficients $s \geq 0.10$ may be accurate in sweeps near fixation, whereas the simulation study in Figure 4.2 indicates that estimation may limited for sweeps near fixation when $s \leq 0.03$.

4.5 The coverage of confidence intervals

The confidence intervals we provide from the parametric bootstrap are motivated by the central limit theorems in Chapter 3: we use standard normal quantiles and the sample standard deviation over bootstrap estimates. However, we know from the simulation study in Section 3.4.1 that in finite samples the detectable IBD rate may have heavier upper tails than those of Gaussian distributions. Since our IBD-based selection coefficient estimator $\hat{s}$ is a monotone and "smooth" function of the detectable IBD rate, we anticipate that an analogous finite-sample result holds for $\hat{s}$. For example, in a specific simulation setting, Figure S18 illustrates that our selection coefficient estimator $\hat{s}$ also resembles a Gaussian distribution but for a heavier upper tail as the true selection coefficient s increases.

We simulate selection coefficient estimates using Algorithm 1 and Equation 4.6 for true selection coefficients $0.01 \leq s \leq 0.03$, $p(0) = 0.33$, the population bottleneck demographic scenario, and a sample size of five thousand. We split 37,500 simulations of each parameter combination into 150 replicates with 250 bootstraps. For each replicate, we use the 250 bootstraps to estimate a sample standard deviation $\hat{\sigma}_{1:B}(\hat{s})$, and calculate lower and upper

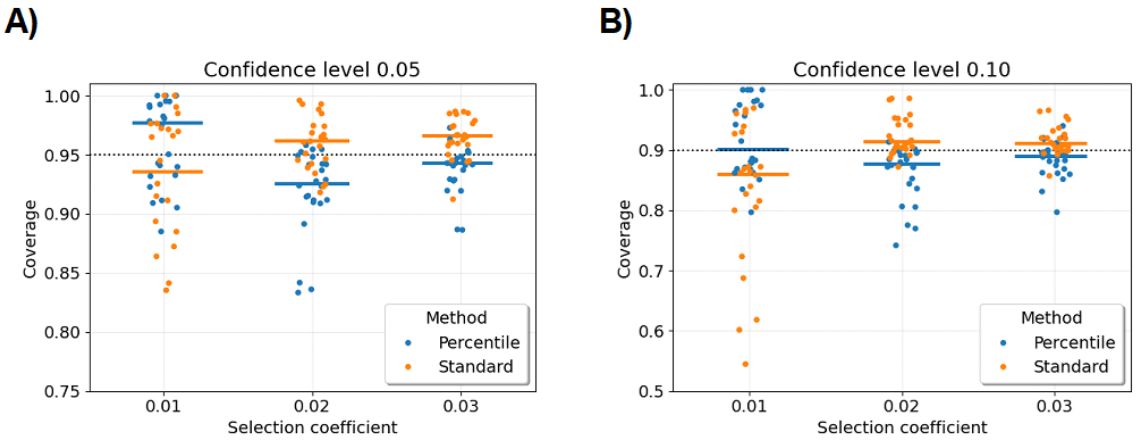

Figure 4.3: Coverage of selection coefficient confidence intervals in exhaustive simulations. Each point in the scatter plot represents the results of one parameter combination (see main text). Standard normal (orange) and percentile (blue) confidence intervals are based on 2500 parametric bootstraps. Coverage (y-axis) of true selection coefficients (x-axis) is calculated over 500 replicate simulations. Dotted horizontal black lines denote $1-\alpha$, where α is the significance level A) 0.05 or B) 0.10. For visibility, the y-axes have different scales.

confidence interval bounds as $s \pm \hat{\sigma}_{1:B}(\hat{s}) \cdot z_{0.05}$. Figure S19 shows on average which percentile the lower and upper bounds are among all 37,500 simulations. The lower bounds are on average less than the fifth percentile of the sampling distribution. The upper bounds are on average less than the ninety-fifth percentile of the sampling distribution. At the same time, the differences between lower bound and lower percentile and upper bound and upper percentile are similar, which could protect against under- or over-coverage.

Given this observed skew in the sampling distribution, we test confidence interval coverage for selection coefficients $s = 0.01, 0.02, 0.03$ in our simulation studies. We consider two demographic scenarios (population bottleneck and constant twenty-five thousand samples), sample sizes of 2500 and 5000, cM length thresholds of 2.0 and 3.0, and sweeping allele frequencies $p(0) = 0.25, 0.50, 0.75$. For each replicate of five hundred simulations, we create

confidence intervals from twenty-five hundred parametric bootstraps. We also compare to percentile-based confidence intervals (see Manly [98] for a review). Each measurement of coverage is taken over five hundred replicates.

Scatter plots in Figure 4.3 display the empirical coverage for simulation of the different parameter settings. The confidence intervals based on standard normal quantiles tend to over-cover for $s \geq 0.02$ and under-cover for $s < 0.01$. Conversely, the confidence intervals based on bootstrap percentiles tend to under-cover for $s \geq 0.02$ and over-cover for $s < 0.01$. One disadvantage of the percentile-based approach is that many bootstraps could be required, especially for small significance levels. For sample sizes on the order 10^5, percentile-based confidence intervals may be less practical to compute. With standard normal confidence intervals based on fifty bootstraps, we observe coverage estimates close to one minus the nominal significance level (Section 4.3.1).

4.6 Discussion

The selection coefficient estimator we propose in this chapter is easy to interpret: the more detectable IBD segments observed, the larger the selection coefficient estimate. Over simulations of many different model combinations here and in Appendix B, we find that mean selection coefficient estimates are close to the true values $0.01 \leq s \leq 0.40$. In humans, there is little evidence of any recent sweep where $s > 0.05$. However, it could be important to have a reliable selection coefficient estimator for $s \geq 0.05$ when studying other organisms. We believe our selection coefficient estimator's ability to estimate $s \geq 0.01$ fills a gap in the current literature (Chapter 5). The main limitation of our approach is that IBD segment detection must be accurate and that the sample size must be large enough to detect at least a few hundred long IBD segments.

The main contribution of our approach is to give confidence intervals. Many existing selection coefficient estimators do not give uncertainty quantification (Chapter 5). We do not offer specific statistical guarantees for our confidence intervals, except to say that they tend to have empirical coverage within 0.10 of one minus the nominal significance level.

Beyond our simulation studies, the fast parametric bootstrap in Section 2.4 can be used to estimate bias and coverage in virtually any hard and recent selective sweep model.

Chapter 5

METHODS TO STUDY SELECTION IN GENETIC DATA

5.1 Introduction

When analyzing genetic data, we often do not know the sweeping allele's location or its frequency. Initially, a scan is performed to catalog regions where some normalized statistic deviates from a genome-wide central tendency. Afterward, a litany of different methods have been developed to identify candidate sweeping alleles and estimate selection coefficients. In this chapter, we focus on estimating the selection coefficient when the sweeping allele is unknown among thousands of variants in an 8.0 cM region. We assume that the 8.0 cM region under selection is identified from a prior scan. (In the subsequent Chapter 6, we propose a genome-wide scan with a multiple hypothesis testing correction.)

To identify possible sweeping alleles, some early methods are Hudson's haplotype test [81] and extended haplotype homozygosity (EHH) [124]. Both these approaches consider correlations within short genetic sequences: Hudson's haplotype test evaluates if the probability that a subset of a sample having fewer than a specified number of polymorphisms in a region is compatible with neutral simulations [81], and EHH measures the proportion of shared haplotypes at increasing physical distances [124]. Additional methods derive from these initial ideas [155, 2, 50, 144, 157, 58]. EHH and many methods derived from it devise statistics based on population genetics, normalize these statistics over the genome, and catalog regions where these statistics differ from a genome-wide central tendency. Alleles or haplotypes are deemed to be good candidates for selection based on the magnitude of deviations from the genome-wide central tendency.

To estimate the selection coefficient, four different paradigms have been considered. Mathieson and McVean [99] and Mathieson and Terhorst [100] use changes in allele frequency

over time to estimate a constant selection coefficient or time-varying selection coefficients, respectively. These methods are limited to studies where allele frequency data is collected at multiple time points. An approximate Bayesian computation (ABC) framework is to calculate summary statistics in simulations of varying s and then to choose s based on the simulations that are close to those in the actual data [116]. The coalescent-based approaches of Stern et al. [142] and Vaughn and Nielsen [152] optimize the likelihood of an ancestral tree in terms of s. Their methods require an estimate of the ancestral tree and can thus be sensitive to the initial task of tree estimation. Some tree inference algorithms apply ad-hoc rules to handle increasing sample sizes, but these methods do not quantify uncertainty over the space of possible tree sequences [83, 140, 160]. Alternative approaches express tree uncertainty by sampling from a Bayesian posterior distribution, but these approaches scale poorly with increasing biobank sizes [93, 92, 121]. Machine learning approaches train neural networks on simulated genetic data for known selection coefficients [151, 75, 103, 122]. Predictive accuracy depends on the model architecture and training.

Many of these methods demonstrate reasonable localization or estimation performance in simulations of hard selective sweeps. One concern in evolutionary biology is that alternative mechanisms could explain the patterns of haplotype structure that these methods leverage. For instance, authors repeatedly suggest that the selection signal at the major histocompatibility complex (MHC) found in many European population studies is an example of balancing rather than positive selection [56, 112]. Recent work has attempted to classify different types of selective sweeps using an ABC framework [116], an EHH-based method [58], or deep learning [84], but it can be difficult to diagnose the important distinguishing features in this case. Further confounding factors include population structure, assortative mating, and unaccounted for familial relatedness. These issues can be partially addressed by preliminary data analysis [5, 119, 118].

In applied work, different combinations of scanning, fine mapping, and estimation methods, developed by different labs and making different assumptions about the data, may be combined to analyze positive selection in a sample. A complete and unifying methodology

that is motivated by statistical and population genetics theory, makes as few assumptions as possible, and is easily interpretable would broadly contribute to the research community. We propose a suite of methods to study selective sweeps that are all based on the effects of strong positive selection $s \geq 0.015$ on identity-by-descent (IBD) within a population. Our methodology addresses four aspects of a selective sweep analysis: 1) scanning for putative genomic regions under selection, 2) estimating frequencies and locations of sweeping alleles, 3) estimating selection coefficients, and 4) providing some diagnostic checks.

The outline of this chapter is as follows. In Section 5.2, we formally describe our methods to investigate evidence in genetic data of recent hard selective sweeps. Our method to scan for putative genomic regions under selection is reserved for Chapter 6. Our selection coefficient estimator is described in Chapter 4. In Sections 5.3.1 and 5.3.2, we describe our simulation studies and the methods we compare to, respectively. In Section 5.3.3, we present our results for estimating sweep frequency and location and the selection coefficient using IBD segments inferred from genetic data. We conclude with a discussion of our methods relative to existing methods that estimate selection coefficients.

5.2 A complete analysis workflow for hard and recent sweeps

5.2.1 Overview and motivation

Here we offer a unifying framework for recent and strong selective sweeps that connects shared haplotypes to the statistical model for recent ancestry in Chapters 2 and 4. Around a locus, haplotypes carrying and not carrying the sweeping allele bifurcates forward in time into two subtrees [158, 74]. Backward in time, the subtree with the adaptive allele coalesces more quickly than expected given genome-wide relatedness. This imbalance is more profound the greater the selection coefficient. IBD segments overlapping a locus provide information about the underlying coalescent tree.

Selective sweeps can create extremely high local IBD rates [20, 148] and large clusters of IBD haplotypes compared to average genome-wide relatedness (Tables S2 and S3). Around

loci with elevated IBD rates, we infer a haplotype group of excess IBD rate, and then we calculate scores for alleles based on their frequency difference between the inferred group and the rest of the sample. These scores serve as the basis for ranking alleles as possible targets of selection. The location and frequency where alleles have the highest scores we estimate to be the location and frequency of the sweeping allele. Next, conditional on the estimated allele frequency and detected IBD segments around the estimated location, we use Equation 4.6 in Chapter 4 to estimate a selection coefficient. Finally, we perform the fast parametric bootstrap procedure in Chapter 4 to calculate confidence intervals (CIs) for the selection coefficient. We implement this suite of methods `iSWEEP` (incomplete Selective sweep With Extended haplotypes Estimation Procedure) in freely available software[1].

5.2.2 Detecting identity-by-descent segments

Our methods use IBD segments inferred from genetic data. We use two existing software programs to infer IBD segments: `hap-ibd` [163] and `ibd-ends` [20]. The accuracy of IBD segment detection depends on the cM length threshold: longer segments are more reliably inferred. We use different length thresholds in different contexts. We use segments longer than 2.0 cM to scan for selected loci and estimate the selection coefficient, and we use segments longer than 1.0 cM to infer an outgroup of excess IBD sharing around a locus. Estimating the selection coefficient is sensitive to inaccuracies in IBD segment detection, whereas determining a haplotype cluster is an intermediate step in our procedure that tolerates more false positive segments. We specify algorithm settings for detecting IBD segments $\geq$ 2.0 and 1.0 cM in Table S5.

5.2.3 Inferring an outgroup with excess identity-by-descent rates

Haplotypes descendant from an adaptive haplotype may belong to large clusters if selection is recent and strong (Tables S2 and S3). We develop a procedure to infer an outgroup,

[1] https://github.com/sdtemple/isweep

denoted A, with excess IBD sharing compared to the rest of the sample, denoted B. First, we detect IBD segments with length 1.0 cM or greater that overlap a focal location. Let haplotypes be nodes and IBD segments be edges in the context of a network graph. Up to now, this detectable IBD graph formation is the same as the detectable IBD graph formation in Chapter 3, Section 3.4.2. Because IBD segments longer than 1.0 cM are infrequent in a population sample, this graph should contain many disconnected haplotype clusters [135]. We also impose the rule that a node in a cluster may only be three edges away from its highest degree node (the haplotype sharing the greatest number of IBD segments), which is different from the graphs in Chapter 3, Section 3.4.2. This condition ensures that there are no bridging edges on the periphery of clusters that would otherwise be disconnected.

Sizes of these haplotype clusters amount to an empirical distribution for IBD cluster size overlapping a focal location. Recall from Chapter 3, Section 3.4.2 that most haplotype clusters are trees of small order, even under strong positive selection where the main effect is the growth of a single mega-cluster. We compute a sample mean and standard deviation for cluster size. Then, we identify clusters whose size exceeds the mean plus three times the standard deviation. These inferred clusters $\hat{a}$ together form the outgroup $\hat{A}$; the rest of the sample we denote as $\hat{B}$. We use multiple clusters to construct the outgroup when the sweeping allele is at frequency $p\,(0) \leq 0.3$. We show in the simulation results (Section 5.3.3) that there is a single inferred excess IBD cluster in most of our simulations of $p\,(0) > 0.3$.

The primary reason for the IBD outgroup algorithm is to determine alleles on a putatively adaptive haplotype. We apply the IBD outgroup algorithm twice in our analysis, once to pinpoint a focal position of the sweeping allele, and once to estimate the frequency and location of the sweeping allele. We denote $\hat{A}_1, \hat{A}_2$ and $\hat{B}_1, \hat{B}_2$ for the first and second applications.

5.2.4 Initializing a focal point near the sweeping allele

Genomic regions passing our selection scan encompass many Mb of genetic sequence. The IBD outgroup algorithm is sensitive to the initial location of IBD detection. We develop

an intermediate step to focus on smaller regions where the sweeping allele may lie. Variants hitchhiking on the sweeping allele should segregate at similar frequencies, suppressing common variation locally about the position of the causal allele. We take a sliding window approach to localize the signal. Consider $Q_w = \sum_{k=1}^{10} p_{(k)}$ where is $p_{(k)}$ is the k^{th} smallest percentile of allele frequencies in the window w [2]. Allele frequencies are not derived allele frequencies, but rather those of the allele more frequent in the inferred group $\hat{A}_1$. High Q_w values correspond to windows where common variation is greater than that of a site frequency spectrum at neutrality. The window with the highest Q_w is a refined focal location for further steps in the estimation procedure. We use window sizes and step sizes of 250 kb and 50 kb, respectively. To address low marker density regions, we do not consider windows where there is a stretch of 25 kb without SNPs.

We contrast our approach to that of the singleton density score from Field et al. [51]. Their approach computes the physical distance between singleton variants. They note that the coalescent model with strong positive selection has long terminal branches, which should result in more *de novo* mutations close to each other. Singleton density score is based on rare variants; high Q_w is based on common variants being at intermediate to high frequency in a window. We only use this statistic to refine a location for IBD calling, whereas singleton density score is applied to genome-wide selection scans.

5.2.5 Estimation of the sweep frequency and location

At the refined location, we apply the IBD outgroup algorithm to infer the $\hat{A}_2$ group with excess IBD sharing. For each variant j within 150 kb of the refined location, we compute a scaled difference in frequencies. Let p_j and $p_{j,\hat{A}_2}$ be the present-day allele frequency in the entire sample and in the inferred outgroup $\hat{A}_2$, respectively. The scaled difference in frequencies is:

$$z_j = \frac{p_{j,\hat{A}_2} - p_{j,\hat{B}_2}}{\sqrt{p_j\,(1-p_j)}} \tag{5.1}$$

[2] Summation over ten is an arbitrary choice that works well in our simulation studies.

Except for a constant scalar, these scores represent the two-sample difference in proportions test statistic[3]. They are similar to the change in derived allele frequency (Δ_{DAF}) statistic in Grossman et al. [64]. However, the Δ_{DAF} statistic is based on derived alleles between populations, whereas we consider a detected outgroup within a single population. We rank SNPs as candidates for selection (or in strong LD with a sweeping allele) based on the z_j scores.

When the selected allele is not known, we aggregate evidence across variants to estimate a frequency for the selected allele. We take a sliding window approach in terms of subsets of two dimensions: physical distance and allele frequency. We consider rectangles of 10% allele frequency length and 250 kilobase pair (kb) width, and we slide them by increments of 2.5% allele frequency and 50 kb. For each allele frequency by base pair rectangle, we average the z_j variant scores contained in it. To address sparse marker coverage, we require that each allele frequency by base pair rectangle includes at least five markers[4]. The middle frequency and middle base pair of the rectangle with the highest average z_j are our estimates of the selected allele's frequency and location. For example, if the subset between 50 – 60% frequency and 250 to 500 kb has the highest average score, the middle frequency is 55% and the middle base pair is 375 kb. It is possible to instead aggregate over variant scores from other methods like `iSAFE` [2] and iHS (integrated Haplotype Score) [155].

5.2.6 Lack of cluster heterozygosity supports the hard sweep model

Loci that pass our selection scan may not be the result of a selective sweep. Other biological processes may give rise to excess IBD rates. For instance, under balancing selection, there could be multiple beneficial haplotypes that each maintain appreciable frequency. In the recent and strong sweep model, we expect elevated extended haplotype homozygosity. Let $q_{\hat{a}}$ be the proportion of haplotypes in cluster $\hat{a}$ out of inferred outgroup $\hat{A}$. We propose to

[3]This observation does not allude to a hypothesis testing framework, as many of the variants in these selected regions have large z_j values.

[4]This threshold choice should depend on the marker density of the dataset

use the Gini impurity index:

$$\text{Gini Impurity Index} := 1 - \sum_{\hat{a} \in \hat{A}} q_{\hat{a}}^2. \tag{5.2}$$

The expected heterozygosity under Hardy-Weinberg equilibrium, a measurement over allele frequencies of the genetic diversity in a population, has the same formula as the Gini impurity index. Here the Gini impurity index of IBD clusters measures the genetic diversity across clusters of excess IBD rate. It increases with more clusters and/or decreasing size of the largest cluster. For example, if $\hat{A}$ consists of one cluster, the Gini impurity index is 0.00; if $\hat{A}$ consists of two equally sized clusters, the Gini impurity index is 0.50; and, if $\hat{A}$ consists of three equally sized clusters, the Gini impurity index is 0.67.

5.2.7 An automated analysis pipeline

The analysis is performed via an automated bioinformatics pipeline [89]. The inputs are phased variant call format files (VCFs) and a pedigree-based recombination map. Samples are assumed to come from a panmictic population and to be unrelated. The procedure is as follows:

1. Detect IBD segments ≥ 2.0 cM along the genome.
2. Scan for loci with excess IBD rates (Chapter 6).
3. Around loci of excess IBD rate,
   (a) Infer outgroup $\hat{A}_1$ and the rest of the sample $\hat{B}_1$ at the location of the maximum IBD rate.
   (b) Calculate Q_w for sliding windows of size 250 kb and step 50 kb using $\hat{A}_1$,$\hat{B}_1$. These windows span all positions at which the IBD rate exceeds some multiple testing adjusted threshold.
   (c) Infer outgroup $\hat{A}_2$ and the rest of the sample $\hat{B}_2$ at the location $\arg\max_w Q_w$.

(d) Compute Gini impurity index given $\hat{A}_2$ and $\hat{B}_2$.

(e) Compute z_j for markers j based on $\hat{A}_2$ and $\hat{B}_2$. These scores are calculated for each SNP within 150 kb of the refined location $\arg\max_w Q_w$.

4. Estimate the current frequency and location of the sweeping allele.

5. Estimate the selection coefficient conditional on the estimated allele frequency and location (Chapter 4).

6. Make confidence intervals via the parametric bootstrap (Chapter 2, Algorithm 1).

Our `iSWEEP` software automates this analysis pipeline and includes a Python package for statistical inference of the selection coefficient.

5.3 Simulation studies

5.3.1 Simulation setup

We investigate the performance of our methods to identify selected alleles, estimate the frequency and location of an unknown sweeping allele, and infer its selection coefficient. We use `SLiM` [71, 72] to model the evolution of an adaptive allele and output a tree sequence. Then, we use `msprime` [9] to place neutral mutations on the tree sequence.

Our default settings for the simulation of sequence data are as follows. Analyses are based on five thousand diploid individuals sampled from the present-day population. We consider constant mutation and recombination rates of 1e-8, gene conversion rate 2e-8, and mean gene conversion tract length of three hundred base pairs, which are identical to the simulation settings in Cai et al. [26]. We introduce genotyping errors at a rate of 0.02%, which is identical to the simulation setting in Browning and Browning [20]. We use the true haplotype phase for IBD segment detection. We use the true effective sizes in each demographic model to estimate the selection coefficient. The demographic scenarios we study

are population bottleneck (BN), three phases of exponential growth (G3), and constant size twenty-five thousand (C25), described in Section 2.5.1 and Figure 2.2

To simulate a selective sweep, we place a sweep mutation at the center of an 8 cM genomic region. We introduce this *de novo* mutation at times ranging from two hundred to five hundred generations ago. We focus on partial sweeps, so we re-run simulations if the *de novo* mutation is not segregating at a present-day allele frequency between ten and ninety percent.

For simulation studies with the BN demographic model, we study selection coefficients $s \in [0.015,\ 0.02,\ 0.025,\ 0.03,\ 0.035,\ 0.04]$. We perform enough simulations to generate more than fifty observations for each combination of selection coefficient and allele frequency bin in $[0.1-0.3, 0.3-0.5, 0.5-0.7, 0.7-0.9]$. Next, we randomly downsample to fifty simulated datasets for each combination of selection coefficient and allele frequency bin. To explore the effect of sample size, we randomly select one thousand and twenty-five hundred samples from these datasets of five thousand individuals.

For simulation studies on the different BN, G3, and C25 demographic models, we set $s = 0.03$ and introduce *de novo* mutations 250 and 300 generations ago. We simulate data one hundred times for each pair of selection coefficient and mutation time. Our simulated data has at least twenty observations for each combination of demographic model and allele frequency bin.

Cryptic population structure contributes to linkage disequilibrium (LD) and can thus confound selective sweep signals. We consider a setting of two equally sized subpopulations with *continuous* random mixing at migration rates forty, ten, and one percent that began five hundred generations ago. The effective population sizes follow the BN demographic model.

We also examine the robustness of our method to different mutation, recombination, and gene conversion rates. For these studies, we simulate sixty replicates for five thousand samples in the BN demographic model. In evaluating how our methods handle different mutation and recombination rates, we vary them between 1e-8 and 2e-8 with fixed gene conversion rate 2e-8 and $s = 0.03$. The ARG-based method we compare to in Section 5.3.3

can be sensitive to mutation and recombination rates [152]. In evaluating how our methods handle gene conversion, which can disrupt IBD segments, we vary gene conversion rates between zero and 2e-8 with $0.015 \leq s \leq 0.035$ and fixed mutation and recombination rates equal to 1e-8.

5.3.2 Comparing against other methods

Approaches to estimate the frequency and location of a sweep

Our method to estimate sweep frequency and location uses scores that are designed to identify plausible sweep variants. Numerous methods exist to score and rank plausible sweep variants. We compare our estimator using scores from either `iSWEEP` or the EHH-based method `iSAFE` [2]. The `iSAFE` score is a scaled difference between two haplotype-based estimators of derived allele frequency, where the estimators skew in different directions in cases of strong positive selection. Akbari et al. [2] show that `iSAFE` scores are better able to identify true sweeping alleles than three alternative methods, including iHS [155].

The input to `iSAFE` is phased genetic data 1.2 Mb to the left and right of the sweeping allele. The input to `iSWEEP` is genetic data for the entire 8.0 cM region. Default settings are used for both methods. The data input and algorithm settings for each method are summarized in Table S6.

We evaluate the `iSWEEP` and `iSAFE` -based estimates in terms of the absolute deviations between their frequency and location estimates and the true frequency and location of the sweeping allele, respectively. The accuracy of these estimates depends on highly scored variants being close in frequency and location to the selected allele. We also compare the `iSWEEP` and `iSAFE` ranks for the true variant, where one is the best possible rank estimate. Identifying the sweeping allele is the primary goal of the `iSAFE` method whereas we aim to estimate allele frequency and location for the purpose of selection coefficient estimation.

Approaches to estimate selection coefficients

Few methods exist to estimate selection coefficients with modern sequences only. We compare the methods `CLUES2` [152] and `ImaGene` [151] against our selection coefficient estimator. `CLUES2` finds a selection coefficient that optimizes a likelihood for the ancestral recombination graph (ARG). (The likelihood calculation of the original `CLUES` [142] method and `CLUES2` are nearly identical, but `CLUES2` implements an approximation that decreases runtime [152].) `ImaGene` [151] trains a convolutional neural network on simulated data, where inputs are genotype matrices and the response variable is the selection coefficient. To apply these methods, we downsample to five hundred of the five thousand simulated diploids due to limitations in computer memory and runtime. The data input and algorithm settings for each method to estimate selection coefficients are summarized in Table S6.

`CLUES2` uses importance sampling from a distribution on ARGs. To sample the posterior distribution of ARGs, we run `Relate` [140] with a buffer of 100 kb to the left and right of the selected allele and the true population effective sizes discretized every ten generations. We generate ten thousand MCMC samples for the branch lengths. We choose these parameters to provide as much data to `CLUES2` as possible while keeping compute times within two days for each replicate. We run `CLUES2` using the exact allele frequency and position of the sweep mutation, six hundred allele frequency bins, and otherwise default parameters.

To train `ImaGene` models, we generate 200 kb of sequence data for each observation, maintaining consistency with the biological parameters and BN demographic scenario used in the simulation study. We use the `msms` software to simulate large amounts of training data because it is faster and consumes less memory than `SLiM` for our examples [45]. The response variables to learn are selection coefficients discretized every 0.005 between 0.00 and 0.05. More precise selection coefficient estimates may be possible with finer scale discretization. At each training epoch, we split one hundred thousand newly generated datasets into 80% training and 20% validation datasets. In total, we simulate two million replicates uniformly proportioned among the different selection coefficients.

By treating the selection coefficient as either a continuous or categorical response variable, we train two `ImaGene` neural network models. The continuous response models use the same settings as in the `ImaGene` tutorial repository. The categorical response model uses the same settings as in the `ImaGene` paper [151]. We use sampling from the predicted probabilities for each category to compute posterior means and high posterior density intervals as in the `ImaGene` paper [151].

Fitting each neural network model takes more than six days of wall clock time using a 24-core computing node. The time to train the `ImaGene` model may be quicker if run on graphics processing units. After the models are fit, the prediction of selection coefficients takes seconds. We do not explore other demographic scenarios because we would have to fit a new `ImaGene` model each time.

5.3.3 Results

Estimating the sweep frequency and location

In simulated sequence data, we compare the use of `iSWEEP` versus `iSAFE` scores as inputs to our sweep frequency and location inference. For varying present-day sweep frequencies, Figure 5.1 conveys the absolute deviations between frequency and location estimates relative to the true values. Except for estimating sweep frequencies $p(0) \leq 0.3$, our inference of sweep frequency and location is uniformly better with `iSAFE` scores instead of `iSWEEP` scores. However, `iSWEEP` has practical advantages that we list later in this section. The ninetieth percentiles of `iSWEEP`-based frequency estimates are always within twenty percent, whereas the ninetieth percentile of `iSAFE`-based frequency estimates has an absolute deviation of more than forty percent when $p(0) \leq 0.3$. The median frequency and location estimates are within 10% and 100 kb of the true values, respectively, using either `iSWEEP` or `iSAFE` scores. Figure S32 shows that the median frequency and location estimates are also within 10% and 100 kb of the true values, respectively, for selection coefficients $0.015 \leq s \leq 0.04$. For these selection coefficients, the use of `iSAFE` instead of `iSWEEP` has smaller absolute deviations.

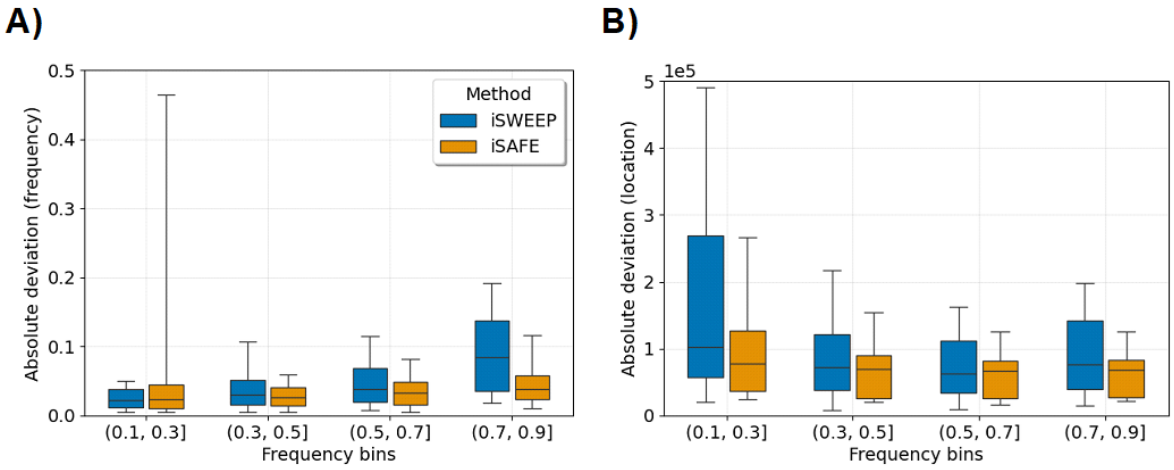

Figure 5.1: Estimating the frequency and location of the sweeping allele for varying allele frequencies. Comparing our estimator of the sweeping allele frequency and base pair location using `iSWEEP` versus `iSAFE` variant scores in terms of: (A) the absolute deviation between estimates and the true sweeping allele frequency, and (B) the absolute deviation between estimates and the true sweeping allele location. Box plots show 10th, 25th, 50th, 75th, and 90th percentile of estimates for each allele frequency bin. There are three hundred simulations of sequence data for each allele frequency bin. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

Figure S33 breaks down the results in terms of the selection coefficient and the allele frequency bin. The pathological performance of `iSAFE`-based scores when $p(0) \leq 0.3$ occurs only when $s \leq 0.02$. The authors of `iSAFE` report that their method is worse at ranking putative sweep alleles when $p(0) < 0.3$ as opposed to $p(0) > 0.3$ [2]. Regardless of the selection coefficient, the absolute deviations in frequency when using `iSWEEP` scores increase as $p(0)$ increases. Our observations suggest that `iSWEEP` and `iSAFE` may complement each other as methods more appropriate for identifying a sweeping allele, and those correlated with it, at low and high frequencies, respectively.

Figures S34 and S35 contrast iSWEEP and iSAFE in terms of their rank estimates of the true sweeping allele. The trends in rank estimation performance exactly coincide with our results for sweep frequency and location estimation, indicating that iSAFE is generally better at identifying the sweeping allele and those correlated with it in our simulation study of *high marker density* data. Figure S36 shows that iSWEEP almost always ranks the true sweeping allele lowest when IBD is called at the true location of the sweeping allele. The main challenge that iSWEEP faces is thus determining a focal location for calling IBD segments.

There are some practical advantages to using iSWEEP scores over iSAFE scores. Table S6 reports that iSWEEP runs six times faster than iSAFE on one CPU and that iSWEEP can be run in parallel. Additionally, the iSWEEP scores are easy to interpret as scaled differences in proportions. The iSWEEP scores are applied to the allele that is more frequent in the inferred excess IBD outgroup (relative to the rest of the sample), whereas iSAFE assumes that an a priori derived allele is the favored allele (Table S6). The inference of an excess IBD outgroup may be useful in other pursuits; for example, we propose a Gini impurity diagnostic that measures a lack of heterozygosity in the excess IBD outgroup. Finally, we observe that iSAFE often does not give scores to more than ninety percent of SNPs, which may limit our inference of sweep frequency and location in lower marker density data.

We also explore the performance of iSWEEP scores in cases of smaller sample sizes, different demographic models, and cryptic population structure. Figure S37 demonstrates that iSWEEP's estimation of the sweep frequency, location, and rank improves rapidly with increasing sample size. In simulations of one thousand diploids from the BN demographic model, there are often less than a few hundred IBD segments longer than 1.0 cM. We caution that the use of iSWEEP scores can lead to inaccurate frequency and location estimates in this case; meanwhile, iSAFE has been applied to much smaller sample sizes [2]. For the G3 and C25 demographic models, Figure S38 exhibits the same trend of decreasing estimation performance as the sweep frequency increases, but performance is marginally worse relative to the BN simulations. Figure S39 conveys that low continuous mixing rates between two cryptic subpopulations can seriously negatively impact the iSWEEP-based sweep frequency

and rank estimation.

Figure S40 shows similar summary statistics for zero versus 2e-8 gene conversion rates. This result may indicate that the `ibd-ends` tool properly addresses small gene conversion tracts in sequence data. Likewise, we find indistinguishable results in Figure S41 in different combinations of mutation and recombination rates. Accurate detection of long IBD segments, regardless of changes to these biological rates, may explain these robust results.

Lack of heterozygosity across inferred excess identity-by-descent clusters

Figure S42 reports Gini impurity indices calculated over inferred clusters of excess IBD rate. We observe that the Gini impurity index is less than 0.60 and equal to zero in ninety-seven and eighty-nine percent of the BN simulations, respectively. Figure S43 shows that the majority of our simulations where the Gini impurity index exceeds 0.6 are when $s \leq 0.02$ and $p(0) \leq 0.3$. Moreover, the ninetieth percentiles of Gini impurity indices are zero for all simulations where $p(0) \geq 0.3$. This finding indicates that Gini impurity indices less than 0.6, and especially equal to zero, may be sufficient but not necessary evidence to support the recent and strong hard sweep model at intermediate to high frequency.

Estimating selection coefficients

Given an allele frequency estimate and accurate IBD detection around an inferred location, we expect inference of $\hat{s}$ to be in line with those of our idealized coalescent-based simulation study. For the population bottleneck simulation study, Table 5.1 presents average selection coefficient estimates, mean absolute deviations between estimates and the true selection coefficients, the average widths of ninety-five percent confidence intervals, and the coverages of ninety-five percent confidence intervals. We observe empirical coverages of at least seventy-five percent for all selection coefficients and nearly ninety-five percent when $s > 0.02$. We measure average confidence interval widths of less than or equal to 0.01 when $s \geq 0.02$. The mean selection coefficient estimates are within 0.003 in all cases. For $s \leq 0.025$, the

True s	Estimate $\bar{\hat{s}}$	MAD	Width	Coverage
0.040	0.0392	0.0023	0.0106	90.9%
0.035	0.0347	0.0020	0.0090	92.5%
0.030	0.0302	0.0015	0.0078	95.0%
0.025	0.0263	0.0018	0.0077	95.0%
0.020	0.0224	0.0028	0.0104	84.8%
0.015	0.0184	0.0044	0.0142	74.1%

Table 5.1: Selection coefficient estimates based on IBD segments inferred from sequence data. For each row, the average of estimates $\hat{s}$, mean absolute deviation (MAD) between estimates and the truth, average confidence interval width, and confidence interval coverage are aggregated over two hundred simulations of sequence data. Estimation is conditional on IBD segments longer than 3.0 cM and inferred frequency and location of the sweeping allele. Ninety-five percent confidence intervals are based on one hundred bootstraps. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

estimation bias is greater the lower the selection coefficient. IBD rates can be similar between neutral evolution and sweeps when $s \leq 0.02$.

Table S7 and Figure S44 show selection coefficient estimates conditional on the true frequency and location of the sweeping allele. We do not observe a discernible difference in estimation results when the frequency and location of the sweeping allele are inferred. Figure S45 reports selection coefficient estimates using detected versus true IBD segments longer than 3.0 cM. (The true IBD segments are derived from the tree sequence using `tskibd` [65].) We do not observe a discernible difference in estimation results when the IBD segments longer than 3.0 cM are inferred. Figure S46 shows the difference between estimates and the true selection coefficient split into allele frequency bins. We do not observe a discernible effect from the sweeping allele frequency in this simulation study. This observation differs from

Figure 4.2C in which low and high-frequency sweeps are shown to have poorer estimation than intermediate-frequency sweeps in a small sample.

We contrast our estimation procedure with the deep learning method `ImaGene` [151] and the tree inference method `CLUES2` [152]. Box plots in Figure 5.2 report percentiles of estimates for varying selection coefficients and strip plots in Figure S47 show the individual estimates. The interquartile range of `iSWEEP` estimates contain the true selection coefficient for $s \geq 0.025$. The tenth and ninetieth percentiles of `iSWEEP` estimates contain the true selection coefficient for $s \geq 0.015$; the interquartile range of our estimates contain the true selection coefficient for $s \geq 0.025$.

Selection coefficient estimates from `ImaGene` are consistently inflated. Figure S48 replicates this finding regardless of the two model fitting procedures explored here. Figure S47 shows that the distributions of predicted values are similar for $0.015 \leq s \leq 0.02$ and $0.025 \leq s \leq 0.035$ simulations, respectively. From the fitted `ImaGene` categorical response model, we compute high posterior density intervals for $s \geq 0.02$ that are more than 200% wider than the corresponding `iSWEEP` confidence intervals (Table S8).

Selection coefficient estimates from `CLUES2` do not increase when we increase the true selection coefficient. In fact, estimates from `CLUES2` are around or below 0.01 for most simulations. `CLUES2` estimates negative selection coefficients or $s = 0.0192$ with an infinite negative log p-value in thirty-four and nineteen percent of our simulations, respectively. The authors of `CLUES` and `CLUES2` acknowledge that their method is limited to $s \leq 0.01$, and they only test performance on simulations where $s \leq 0.01$ [142, 152]. An analysis of `CLUES` from separate authors observes results like ours for $0.01 \leq s \leq 0.02$. On the other hand, they report that `CLUES` can be accurate in this range when given the true simulated tree [75]. Knowing the true coalescent tree is unrealistic for any analysis. One explanation for the results we observe for `CLUES2` is that inferring the high-dimensional ARG is extremely difficult [73], which we do using the `Relate` [140] method. Vaughn and Nielsen [152] show that using ancient DNA in the true ARG can improve `CLUES2` inference. Overall, we suggest our method as suited for recent strong positive selection $s \geq 0.015$ sweeps whereas current

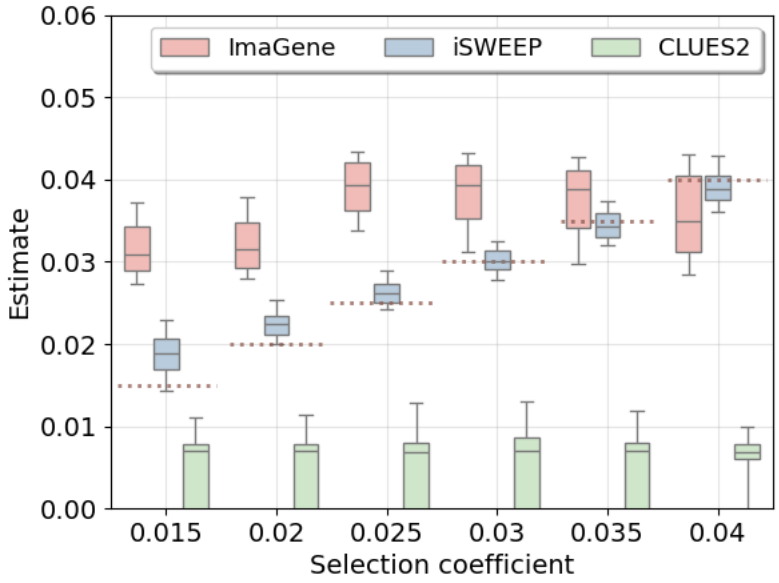

Figure 5.2: Estimating selection coefficients from simulated sequence data. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of estimates from `iSWEEP`, `CLUES2`, and `ImaGene` over two hundred simulations of each selection coefficient. Horizontal dotted brown lines correspond to the true selection coefficient. The sample size is five thousand diploids for `iSWEEP`. Due to memory and runtime limitations, the sample size is five hundred diploids for `CLUES2` and `ImaGene`. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

alternatives may address weak selection $s < 0.015$ sweeps [142, 151, 75, 152].

Figure S37 suggests that the accuracy of our selection coefficient estimator can carry over to examples of smaller sample size in the BN demographic scenario. We observe median estimates within 0.005 of the true selection coefficients for $s \geq 0.015$ and sample sizes greater than one thousand diploids. The interquartile ranges of estimates contain the true selection coefficients for $s \geq 0.015$ and for sample size greater than or equal to twenty-five hundred diploids. We cannot reliably distinguish between $s = 0$ and $s \leq 0.02$ for a sample size

of one thousand diploids in the BN demographic scenario. There are generally fewer than fifty inferred IBD segments longer than 3.0 cM in this case. While Figure 4.2B indicates that smaller selection coefficients can be estimated with smaller detection thresholds, Figure S49 demonstrates that inaccuracies in IBD segment detection for lengths < 3.0 cM can bias selection coefficient estimation. Using `hap-ibd` and `ibd-ends` and our choice of algorithm settings, fewer IBD segments longer than 3.0 cM are inferred than the truth, deflating our estimates.

For the different demographic scenarios and $s = 0.03$, Table S9 presents average selection coefficient estimates, mean absolute deviations between estimates and the true selection coefficients, the average widths of ninety-five percent confidence intervals, and the coverages of ninety-five percent confidence intervals. The mean selection coefficient estimates are within 0.002 in all demographic models. We observe empirical coverages of at least seventy-five percent for all demographic models and nearly ninety-five percent in the BN and C25 simulations. We measure an average confidence interval width greater than 0.01 in the C25 simulations, which is more than 0.005 wider than in the other demographic models. This finding may be due to increased variance in the coalescent process when $N_e(t)$ is smaller in the most recent generations. Finally, the accuracy of our selection coefficient estimator carries over to cryptic population substructure (Table S9) and varying mutation, recombination, and gene conversion rates as well (Figures S41 and S40).

5.4 Discussion

Our IBD-based approach to model recent and hard selective sweeps $s \geq 0.015$ pinpoints selection signals to a resolution of 250 kb, estimates frequencies and locations of unknown sweeping alleles within 10%, and infers selection coefficients within 0.005 of the true value in most of our simulations of sequence data. We also suggest the Gini impurity index as a diagnostic check when a lack of heterozygosity in IBD clusters is suggestive of a hard sweep. We say that our methods are suitable for $s \geq 0.015$ selective sweeps because our estimates become increasingly inflated as the selection coefficient decreases past this lower bound.

These hard sweeps may not pass our genome-wide selection scan (Chapter 6). Moreover, Riley et al. [122] show that a neural network they trained can seldom discriminate between $s \geq 0.01$ and $s = 0.00$. When $s \geq 0.015$, we find our selection coefficient estimates to be favorable to those of an ARG-based method and a deep learning-based method.

Methods to estimate selection coefficients should quantify uncertainty, especially when the true selection coefficients are less than 0.01. Ideally, ninety-five percent confidence, credible, or high posterior density intervals should be demonstrated to contain true selection coefficients in ninety-five percent of simulations. The confidence interval coverages we observe in simulations of sequence data are reasonable for $s \geq 0.015$, even when the sweeping allele is unknown and IBD segments are inferred. We also contrast the widths of high posterior density intervals from a deep learning method to be at least 150% larger than the widths of our confidence intervals. Point estimates for an ARG-based method and two deep learning-based methods have been observed to include zero when $s \leq 0.01$ [142, 75]. Given the potential harms of research on natural selection being misappropriated or misinterpreted [78], we believe that it is crucial to quantify uncertainty when modeling a hard sweep.

All of our results are obtained from automated bioinformatics pipelines that run the statistical methods presented here[5]. Algorithm settings can be specified in a configuration file to reproduce our results. The most computationally intensive aspects of our methods involve the use of efficient `hap-ibd` and `ibd-ends` algorithms to infer IBD segments [20, 163]. We benchmark that our methods to estimate selection coefficients run in minutes versus an ARG-based method that takes hours to days and a deep learning method that could take weeks of training time.

The input data to our methods is high-coverage whole genome sequences from the current generation as opposed to other methods based on ancient or archaic DNA. Increasing sequencing to a few hundred samples may render our approach feasible for the statistical analysis of positive selection in non-human populations with recent effective sizes on the or-

[5] https://github.com/sdtemple/isweep

der of ten thousand or less (Chapter 4). One current example of such a non-human dataset concerns *Plasmodium falciparum* parasites which have developed resistance to antimalarial drugs [65]. Technical limitations remain in applying our methods to non-humans, especially obtaining accurate assemblies, genotype calls, recombination maps, and haplotype phasing. Depending on the dataset, another IBD segment detection method may be more suitable for that aspect of our analysis workflow [54, 65, 105, 104, 134]. Once genomic regions putatively under selection are identified, our methods facilitate modeling selective sweeps without knowing the phenotype being selected for nor requiring that the sweeping allele is genotyped.

Chapter 6

SCANNING FOR EXCESS IDENTITY-BY-DESCENT RATES

6.1 Introduction

Before modeling a hard sweep with estimates of the allele frequency and selection coefficient, we scan the genome for regions in which there is significant evidence to reject the no selection null model. The scanning statistic may be designed to capture different alternative hypotheses that deviate from the neutral theory of Kimura [86]. To scan for targets of recent positive selection, standard methods study either decay patterns of extended haplotypes [124, 155], excess recent shared ancestry [20, 3, 104, 111], differences between linkage disequilibrium (LD) and pedigree-based maps [108], changes to the site frequency spectrum [145, 46], haplotype differences between ancestrally diverged populations [117, 125, 144], allele frequencies in ancient DNA samples [101], or excess ancestry proportions in admixed samples [127]. Many of the cited methods offer p-values without determining the multiple testing burden [111, 51, 140] or explicitly state that their scanning statistic is not a formal hypothesis test [20, 155].

We give the opinion that an alternative hypothesis of strong adaptive evolution warrants a formal hypothesis testing framework with multiple testing adjustments. Two main approaches to develop multiple testing adjustments are to control the family-wise error rate (FWER) or the false discovery rate (FDR). FWER is the probability of rejecting the null hypothesis one or more times when the null hypothesis is true (see Sidak [136] for an example). FDR is the expected proportion of rejected null hypotheses when the null hypothesis is true (see Benjamini and Hochberg [11] and Holm [79] for examples). In this work, we derive a multiple testing correction designed to control FWER.

The Bonferroni correction controls FWER under the null model, but it can be very

conservative in genetics when the scanning statistics are correlated. Palamara et al. [111] use the Bonferroni correction when evaluating significance for their scanning statistic at user-defined cM spacings. A related example in genetic studies is using a p-value threshold of 5e-8. This significance level is based on the 0.05 family-wise significance level with the Bonferroni correction and an assessment of the number of effective hypothesis tests in human genotype array data from the early 2000s [35]. Field et al. [51] and Speidel et al. [140] use this p-value threshold even though their data is far denser than human genotype array data in the early 2000s.

Permutation or simulation-based approaches can control FWER under valid permutation or simulation frameworks, but these procedures can be computationally intensive and difficult to design [21, 62, 63, 37, 131, 137]. Some of these simulation-based approaches were applied to sample sizes less than a few thousand [21, 131], or they leveraged the fact that Wald and score statistics from linear models are asymptotically normally distributed [37, 63].

Another approach is to determine the threshold such that the probability of the first hitting time in a stochastic process is equal to the nominal significance level. In an IBD mapping study, Browning and Thompson [21] approximate transitions between IBD and non-IBD states as a Markov process and then derive a genome-wide significance threshold with this approach. In an admixture mapping study, Grinde et al. [63] approximate their Wald test statistics as an Ornstein-Uhlenbeck process and then calculate the genome-wide significance level with this approach.

Here we determine an approximate significance threshold with a rigorously supported multiple testing correction for the genome-wide scan for excess (detectable) identity-by-descent rates in Browning and Browning [20] and Temple et al. [148]. Similar to Grinde et al. [63], we treat our genome-wide scan for excess IBD rates as an Ornstein-Uhlenbeck process, and we use the analytical approximation proposed in Siegmund and Yakir [137]. We calculate the threshold such that the approximate probability that the maximum IBD rate exceeds it is equal to a family-wise significance level. The Ornstein-Uhlenbeck process is normally distributed at every point, spatially homogeneous, and Markov. The assumption

of normality at every point is supported by our central limit theorems in Chapter 3. Spatial homogeneity is an assumption consistent with neutral evolution. The IBD rate along the chromosome is not a Markov process[1], but assuming the Markov property has been useful in prior works [21, 47].

The outline of this chapter is as follows. In Section 6.2, we review the detectable IBD-based selection scan, and we derive an approximate multiple testing correction. After IBD segment detection, this method is the first method in our analysis workflow (Chapter 5, Section 5.2). In Section 6.3, we conduct empirical studies to evaluate FWER control in neutral simulations and statistical power in simulations of strong positive selection. In Appendix A.3, we provide an interpretation of our multiple testing correction in terms of the effective number of hypothesis tests. In Chapter 7, we use our multiple testing correction to reanalyze the European ancestry selection scan in Temple et al. [148] as well as to study selection in non-European ancestry populations. To conclude, we discuss our scanning statistic in the context of other methods, and we suggest directions for future work.

6.2 Multiple testing in IBD-based selection scans

Let $Y_{a,b}(m)$ be the indicator that the IBD segment between haplotypes a and b is detectable and overlapping the m^{th} focal position. The IBD rate at the m^{th} locus is $\bar{\mathbf{Y}}_m = f(n)^{-1} \sum_{(a,b)} Y_{a,b}(m)$, where $f(n) = 2n(2n-1)/2 - 2n$ in diploids and $\binom{n}{2}$ in haploids. The hypothesis test we consider is

$$H_0 : \mathbb{E}[\bar{\mathbf{Y}}_m] = \mu_0 \tag{6.1}$$

$$H_1 : \mathbb{E}[\bar{\mathbf{Y}}_m] > \mu_0 \tag{6.2}$$

where μ_0 is a genome-wide mean IBD rate around a locus. This null model is consistent with no positive selection. The alternative model is consistent with positive selection *or* other

[1] As a counterexample, consider two haplotypes a and b with three IBD segment indicators $Y_{a,b}(m_1), Y_{a,b}(m_2), Y_{a,b}(m_3)$ at the m_1^{th}, m_2^{th}, and m_3^{th} positions. For constant population size N, we can calculate $P(Y_{a,b}(m_3)|Y_{a,b}(m_2), Y_{a,b}(m_1))$ and $P(Y_{a,b}(m_3)|Y_{a,b}(m_2))$ for $\{0,1\}^3$. These probabilities are not equal. Their calculations for any choice of m_1, m_2, and m_3 are left as an exercise.

evolutionary mechanisms.

Under certain asymptotic conditions, the IBD rate around the m^{th} locus is normally distributed (Theorems 3.3.1, 3.3.2, and 3.3.3). Let $\hat{\mu}_{1:M}$ and $\hat{\sigma}_{1:M}$ be the sample mean and sample standard deviation of M IBD rates along the genome:

$$\hat{\mu}_{1:M} := M^{-1} \sum_{m=1}^{M} \bar{\mathbf{Y}}_m; \tag{6.3}$$

$$\hat{\sigma}_{1:M} := \sqrt{(M-1)^{-1} \sum_{m=1}^{M} (\bar{\mathbf{Y}}_m - \hat{\mu}_{1:M})^2}. \tag{6.4}$$

Browning and Browning [20] and Temple et al. [148] (our initial work) suggest a heuristic significance threshold of $\hat{\mu}_{1:M} + 4 \times \hat{\sigma}_{1:M}$ [2]. Under the standard normal model, this threshold corresponds to a significance level of $1 - \Phi(4) = 3.17 \times 10^{-5}$.

6.2.1 Two possible multiple testing corrections

We assume that the IBD rates along the genome follow an asymptotically standardized Ornstein-Uhlenbeck process

$$\{\tilde{\mathbf{Z}}\}_{1:M} := (\{\bar{\mathbf{Y}}\}_{1:M} - \hat{\mu}_{1:M})/\hat{\sigma}_{1:M}. \tag{6.5}$$

Each $\tilde{\mathbf{Z}}_m$ is a standard normal random variable with mean zero under the null model. The distance between consecutive focal positions m_{i+1} and m_i is set to be constant Δ. The covariance between standardized IBD rates $\tilde{\mathbf{Z}}_{m_1}$ and $\tilde{\mathbf{Z}}_{m_2}$ at different loci depends on the genetic distance between the loci and an exponential decay parameter θ.

$$\text{Cov}(\tilde{\mathbf{Z}}_{m_1}, \tilde{\mathbf{Z}}_{m_2}) = \exp(-\theta \cdot \Delta). \tag{6.6}$$

The exponential decay parameter θ is not known for the IBD rate process but must be estimated, whereas θ is the time of admixture in Grinde et al. [63], which can be estimated or assumed from prior knowledge.

---

[2] Technically speaking, they use the genome-wide median, not the mean, which can be more robust to outlier effects. If the IBD rate around a locus is normally distributed, the true mean and median are the same.

Analytical approximation

To control FWER, we must determine a threshold z such that $P(\max_m \tilde{\mathbf{Z}}_m \geq z) = \alpha$, where α is the nominal significance level. Let L be the total length of the genome, C be the number of chromosomes, and Φ and ϕ be the cumulative distribution and density functions of the standard normal random variable. The Siegmund and Yakir [137] analytical approximation is

$$P(\max_{1 \leq m \leq M} \tilde{\mathbf{Z}}_m \geq z) \approx 1 - \exp(-C[1 - \Phi(z)] - \theta \cdot L \cdot z \cdot \phi(z) \cdot \nu(z\{2\theta\Delta\}^{1/2})), \tag{6.7}$$

where $\nu(\cdot)$ is defined in Siegmund and Yakir [137] to address a discretized stochastic process. When $\Delta \to 0$ (the continuous process), $\nu(0) = 1$. The approximate multiple testing threshold is z such that the right-hand term in Equation 6.7 is equal to the nominal significance level. We provide an interpretation in Appendix A.3 where $\exp(-C[1 - \Phi(z)])$ is related to an infinite number of independent tests and $\exp(-\theta \cdot L \cdot z \cdot \phi(z) \cdot \nu(z\{2\theta\Delta\}^{1/2}))$ corrects for correlation in the stochastic process.

Simulation-based approach

Another way to control FWER is to simulate the Ornstein-Uhlenbeck process for known or estimated θ. Let J be the number of simulations and $M := \lfloor L \div \Delta \rfloor$. The simulation approach goes as follows.

1. Let $\mathbf{z}_{1:J}$ be an empty vector.
2. For j in 1 to J:
   (a) Draw $z_1 = Z_1 \sim N(0, 1)$.
   (b) For m in 2 to M:
      i. Draw $z_m = Z \,|\, z_{m-1} \sim N(z_{m-1} \cdot \exp(-\theta \cdot \Delta), 2 - 2 \cdot \exp(-\theta \cdot \Delta))$.
   (c) Append $\max_m z_m$ to the vector $\mathbf{z}_{(1:J)}$

3. Return the $(1-\alpha)\%$ quantile of $\mathbf{z}_{(1:J)}$.

For nominal significance levels like 0.01 or 0.05, this simulation approach requires a few thousand simulations J and runs within a few minutes to hours (depending on the genome length L). A precise simulation algorithm would simulate each chromosome individually based on its length, but we use the total genome length instead. This multiple testing correction is valid when the true model is the Ornstein-Uhlenbeck process.

We conduct a simple validation study to determine if the analytical approximation and simulation-based thresholds control FWER. We also compare to the continuous-time analytical approximation of Equation 6.7 and the standard normal quantile with the Bonferroni correction. The simulation settings are $\Delta = 0.0005$ Morgans, ten chromosomes, each chromosome is 1 Morgan long, θ ranges from 1 to 100, ten thousand simulations are run to determine the simulation-based threshold, and ten thousand simulations are run to compute FWERs. Figure S50 shows the quantile thresholds based on the different techniques and the FWERs at a nominal significance level of 0.05. For all θ, the Bonferroni and continuous-time analytical corrections have FWERs less than 0.025 (very conservative). The discrete-time analytical and simulation-based corrections have FWERs close to 0.05 when $\theta \geq 50$. The FWERs for the discrete-time analytical approximation are slightly less than 0.05 when $\theta \geq 50$ and considerably less than 0.05 when $\theta < 50$. The FWERs in the simulation-based approach are exact (up to some noise) for all θ. Grinde et al. [63] also finds that the discrete-time analytical approximation is slightly conservative in their admixture mapping study where $\theta \approx 10$.

6.2.2 Estimator of the exponential decay parameter

Before standardizing the IBD rates, we adjust for extreme outliers that could be present in real genetic data. First, we compute an initial genome-wide mean IBD rate plus four standard deviations. Second, we compute a revised genome-wide mean IBD rate and standard deviation, excluding IBD rates that exceed the initial threshold. We standardize the IBD

rates with the revised mean and standard deviation estimates.

To estimate the exponential decay parameter θ, we fit the slope parameter in a simple log-linear model. There is no intercept parameter in the log-linear model. We apply linear interpolation to the recombination map to hold the spacings between IBD rates constant. We estimate the covariance between standardized IBD rates at genetic positions Δ times some integer constant apart, excluding IBD rates that exceed the initial threshold. The integer scalars increment by 1 up until the covariance is between positions 4.0 cM apart. We treat the integer-scaled Δ's as covariates and the estimated covariances as response variables.

Using the same Ornstein-Uhlenbeck process setup as our prior validation study, we find that this estimator is accurate for $30 \leq \theta \leq 90$, with a slight bias to overestimate θ (Figure S51). Additionally, if we calculate FWERs using the estimates $\hat{\theta}$, Figure S52 shows similar FWERs as those in Figure S50B where the true θ is given. At the family-wise significance level 0.05, the FWERs are slightly less than 0.05, which may be due to larger than necessary multiple testing adjustments (Figure S50) when overestimating θ.

6.3 Simulation studies

6.3.1 Simulation setup

Now, to evaluate the Ornstein-Uhlenbeck process approximation, we study our multiple testing correction procedure for the IBD rate process in simulated genetic data. We estimate the exponential decay parameter θ from simulated IBD segments, and then we use the estimate $\hat{\theta}$ to calculate our multiple testing adjusted thresholds. For these calculations of the genome-wide significance level, we consider different step sizes 0.02, 0.05, and 0.10 cM.

To calculate FWERs, we consider five hundred simulations of entire genomes from twenty-five hundred diploids. FWER is calculated as the percentage of the five hundred null model simulations in which there is at least one significant result. The constant recombination rate is 1e-8, and the demographic scenario is the population bottleneck. We use `msprime` [9] to simulate ten chromosomes each of length 1 Morgan. We use `tskibd` [65] to get IBD segment

lengths longer than 2.0 and 3.0 cM from the tree sequence output by `msprime`. Figure S53A illustrates one simulation of the IBD rate process across the genome. The multiple testing threshold is greater than the four standard deviations rule suggested in Browning and Browning [20] and Temple et al. [148].

The data for these simulations amounts to 1 Tb disk storage, predominantly due to the `msprime` tree sequences. We are prohibited from making VCF marker data and inferring IBD segments, which would create more than 1 TB of additional disk memory. In a pilot study of ten simulations, we place mutations on the ARG with genome-wide rate 1e-8 and then infer IBD segments with our `hap-ibd` and `ibd-ends` analysis workflow (Table S5). Figure S53B illustrates one simulation of the *inferred* IBD rate process across the genome. We observe similar genome-wide median IBD rates and significance thresholds between the true and inferred IBD rate processes. The inferred IBD rates are within 95 to 105% of the corresponding true IBD rates (Figure S53C). Across the ten simulations, the average estimates $\hat{\theta}$ are 69 and 75 and the average standard deviations $\hat{\sigma}_{1:M}$ are 19 and 20 for the true and inferred IBD rate processes, respectively.

To calculate statistical power, we simulate IBD segments overlapping a focal point (Algorithm 1) for $s \geq 0.006$ and $p(0) = 0.25, 0.50, 0.75$. For these simulations, we do not simulate inferred IBD segments in small selected regions because it is experimentally challenging to make uniform numbers of simulations in different allele frequency bins for varying selection coefficients. Based on the accurate selection coefficient estimates in Chapter 5, we believe that Algorithm 1 simulates IBD rates around a locus similar to those inferred from simulated marker data.

Power is calculated as the proportion of our selective sweep simulations (alternative hypotheses) in which we reject the null model. The threshold in our power calculations is the average over the multiple testing adjusted thresholds in our five hundred neutral simulations. We estimate power using two hundred simulations for each pair of selection coefficient and sweeping allele frequency.

6.3.2 Estimating the exponential decay parameter

Box plots in Figure S54 show the percentiles of estimates $\hat{\theta}$ using IBD segments ≥ 2.0 and ≥ 3.0 cM. Regardless of the step size Δ, the distribution of estimates $\hat{\theta}$ is the same. The medians of estimates $\hat{\theta}$ for the ≥ 2.0 and ≥ 3.0 cM processes are roughly 40 and 62.5, respectively. As θ increases, and holding the genetic distance between two IBD rates constant, the covariance between the two IBD rates decreases, which could be interpreted in terms of fewer detectable IBD segments overlapping nearby loci on average. This interpretation is consistent with the smaller estimates $\hat{\theta}$ for the 2.0 cM versus the 3.0 cM threshold.

6.3.3 Controlling family-wise error rate

Table 6.1 reports the multiple testing adjusted significance levels and empirical family-wise error rates for the analytical approximation, simulation-based approach, and the Bonferroni correction in the ≥ 2.0 cM IBD rate processes. The adjusted significance levels from analytical and simulation-based methods are nearly an order of magnitude larger than those using the Bonferroni correction. At the family-wise significance level of 0.05, the FWERs of our analytical and simulation-based methods are inflated by more than 150%, whereas the family-wise error rates of the Bonferroni method are deflated by less than 50%. At the family-wise significance level 0.10, the average multiple testing quantiles of the analytical and simulation-based methods are 4.196 and 4.176, which both exceed the four standard deviations rule used in Browning and Browning [20] and Temple et al. [148]. Strictly speaking, our multiple testing adjustments still do not control FWER, which we predicted in Temple and Thompson [147] (Chapter 3). One reason for the anti-conservativeness of the hypothesis test with ≥ 2.0 cM segments may be that the upper tail of the IBD rate's distribution may be heavier than the upper tail of a normal distribution (Figures 3.4 and S9).

Table S10 reports the adjusted significance levels and family-wise error rates, except using the 3.0 cM threshold. In this case, the IBD rate overlapping a locus may be better approximated by a normal distribution than in the ≥ 2.0 cM selection scan (conditions on the

Nominal level	**Adjusted** Analytical	Simulation	Bonferroni	**FWER** Analytical	Simulation	Bonferroni
0.01	1.08e-6	1.30e-6	2.08e-7	0.024	0.028	0.006
0.05	6.24e-6	7.03e-6	1.04e-6	0.088	0.098	0.024
0.10	1.36e-5	1.49e-5	2.08e-6	0.140	0.146	0.040

Table 6.1: Significance levels and family-wise error rates after multiple testing corrections. Significance levels are adjusted for multiple testing based on scans over 10 chromosomes of size 100 cM and tests every 0.02 cM (50,000 total tests). The analytical approximation of Siegmund and Yakir [137] and the simulation method are based on a fitted Ornstein-Uhlenbeck process. Each simulation has a different threshold as a result of estimating θ. Family-wise error rate (FWER) is the percentage of five hundred genome-wide scans that have at least one statistically significant result. The sample size is twenty-five hundred diploids. The demographic scenario is the population bottleneck. The IBD segment detection threshold is 2.0 cM.

detection threshold in Theorems 3.3.1, 3.3.2, and 3.3.3). The FWERs of the analytical and simulation-based methods are indeed conservative in the $\geq$ 3.0 cM selection scan. Recall that the multiple testing adjustments can be conservative when $\theta < 50$ (Figure S50) or when θ is estimated (Figure S52). We thus make the important observation that there are two counteracting factors affecting FWER control: the multiple testing adjustments are conservative in true Ornstein-Uhlenbeck processes, but the test could be anti-conservative if the Ornstein-Uhlenbeck process is a poor approximation for the IBD rate process.

For the anti-conservative $\geq$ 2.0 cM selection scan, we consider a modification to the test that explores if the significant results barely exceed the threshold. We calculate at each locus the minimum of its value and the flanking values both to its left and right. Next, we

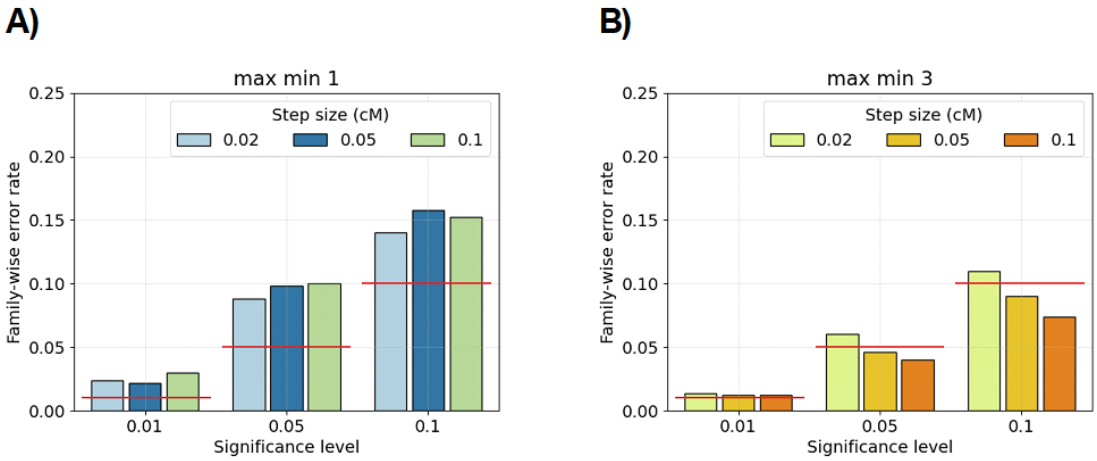

Figure 6.1: Family-wise error rates for genome-wide hypothesis testing in null model simulations. Bar plots show family-wise error rates (y-axis) using true IBD segments ≥ 2.0 cM from simulated IBD rate processes. The hypothesis testing method is the Siegmund and Yakir discrete approximation. In each non-overlapping window of size A) 1 or B) 3 marginal test statistics, we compute the minimum of IBD rates at each step, and the test is if the maximum over all windows is less than or greater than the multiple testing quantile. Estimates $\hat{\theta}$ are based on covariance decays up to 4.0 cM. There are five hundred simulations for each pair of significance level (x-axis) and step size (colors in legend). Family-wise significance levels are denoted with horizontal red lines. The demographic model is the population bottleneck. There are twenty-five hundred diploid samples. The amount of data for each simulation is equal to ten chromosomes of uniform length 100 cM.

calculate the maximum over the entire genome of these aggregated minimum values:

$$\max_{1 \leq m \leq M} \min\{\hat{\mathbf{Z}}(m-1), \hat{\mathbf{Z}}(m), \hat{\mathbf{Z}}(m+1)\}. \tag{6.8}$$

Using the same multiple testing threshold as the analytical approximation, Figure 6.1 shows that family-wise error rates decrease using the max-min statistic. This result indicates that a considerable proportion of the Type 1 errors correspond to marginally significant results. We also note that the choice of step size does not noticeably affect FWERs. We attribute this to

proper handling of Δ in the analytical approximation, whereas the conservative Bonferroni correction indiscriminately penalizes the number of user-specified hypothesis tests (Tables 6.1 and S10).

Next, when there is a significant result, we investigate how many significant results there are. Since the IBD rate process has non-negligible correlations, we anticipate there to be multiple significant results adjacent to each other. Across non-overlapping windows of varying sizes, we count the number of windows that have a significant result. Figure S55 shows that the number of windows with a significant result decreases to a median of 1 when the window size is 0.40 cM and the family-wise significance level is less than 0.05. Coupled with our previous observations, when a Type 1 error is made, we tend to find only one or a few marginally significant results in aggregated regions less than 0.5 cM.

6.3.4 Statistical power in selective sweeps

Figure 6.2A shows the power estimates for $0.006 \leq s \leq 0.014$, $0.25 \leq p(0) \leq 0.75$, and the ≥ 2.0 cM selection scan. Power estimates are uniformly greater when allele frequency $p(0) = 0.50$ as opposed to $p(0) = 0.25$ or $p(0) = 0.75$. The increased ability to detect positive selection when the sweep is at an intermediate frequency is consistent with our analyses in Chapter 5. Power estimates are less than 5% when $s \leq 0.008$ but greater than 90% when $s \geq 0.014$. In between these extremes, power estimates range from 15% to 40% when $s = 0.010$ and from 55% to 85% when $s = 0.012$.

Figure 6.2B shows the power estimates for $0.006 \leq s \leq 0.014$, $0.25 \leq p(0) \leq 0.75$, and the ≥ 3.0 cM selection scan. We estimate zero power for all combinations of selection coefficients and allele frequencies. The biased estimation of selection coefficients $s \leq 0.02$ in Chapters 4 and 5 is consistent with the lack of power to detect $s < 0.015$ sweeps in the ≥ 3.0 cM IBD-based selection scan.

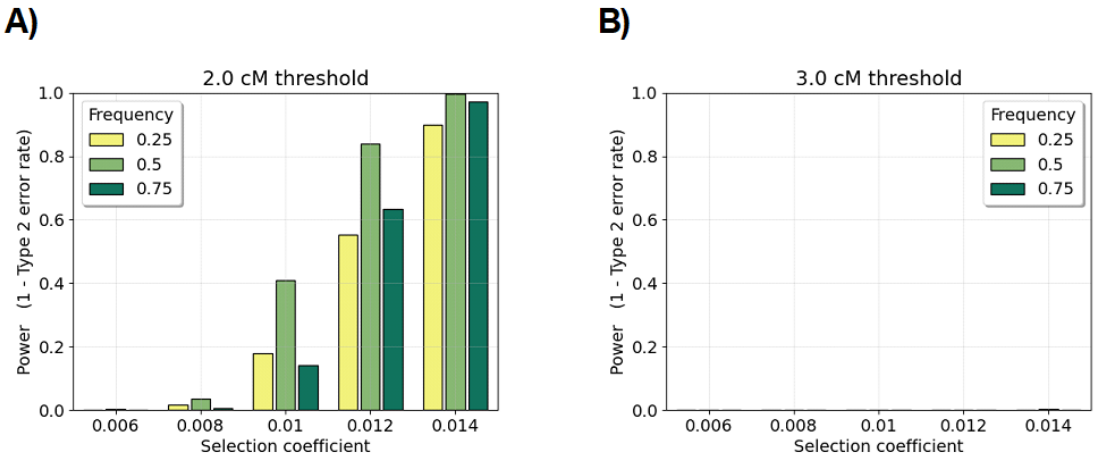

Figure 6.2: Power simulations for different selection coefficients s and sweeping allele frequencies $p(0)$ based on coalescent IBD segment lengths. Bar plots show power (1 – Type 2 error rate) (y-axis) using true IBD segments A) $\geq$ 2.0 cM or B) 3.0 cM overlapping the selected allele. Hypothesis testing is based on the Siegmund and Yakir discrete approximation for step size 0.02 cM and family-wise significance level 0.05. This data is simulated under an alternative hypothesis using our coalescent simulator (Algorithm 1). Power is the proportion of tests where the null model is rejected at the p-value threshold corresponding to the family-wise significance level of 0.05. There are two hundred simulations for each pair of selection coefficient (x-axis) and allele frequency (colors in legend). The demographic model is the population bottleneck. There are twenty-five hundred diploid samples. Additive selection is simulated.

6.4 Discussion

In this chapter, we present two multiple testing corrections for a genome-wide scan of excess IBD rates. The multiple testing corrections account for correlation in the IBD rate process, whereas independence assumptions could lead to very conservative tests with low power to reject null models when an alternative model is true. Our multiple testing corrections should be reasonable under certain asymptotic conditions on sample size and scaled population size

(Theorems 3.3.1, 3.3.2, and 3.3.3). On the other hand, our asymptotic conditions may be more strict than those of Wald and score statistics (see Grinde et al. [63]). In practice, we indicate that the adjusted significance thresholds are anti-conservative, consistent with our simulation studies in Chapter 3. At the same time, our adjusted significance levels are more conservative than the heuristic rule used by Browning and Browning [20]. Compared to computationally intensive permutation- or simulation-based approaches, we calculate our analytical approximation in seconds.

Our simulation studies suggest that our adjusted significance thresholds are only slightly anti-conservative. Rejected hypothesis tests under the true null model might barely exceed our significance threshold in small cM windows. Similar to the discussion in Salyakina et al. [128], we recommend that any statistically significant results from our hypothesis testing framework be scrutinized in follow-up studies. Some of the methods in Chapter 5 may be useful for these purposes. For instance, at a statistically significant locus, one could calculate the Gini impurity index of excess IBD clusters or investigate if there is a singular predominant excess IBD cluster around a locus.

Although the $\geq$ 2.0 cM selection scan can be anti-conservative, we still recommend it over the $\geq$ 3.0 cM selection scan which may be conservative but has limited power. Moreover, greater sample sizes are required for larger detection thresholds, which can be an important practical consideration. A final consideration is the IBD segment detection method, which can be very accurate for $\geq$ 2.0 cM segments in high-quality sequence or array data [163, 20, 54, 134, 105, 104].

One important feature of our multiple testing procedure is that estimation of the exponential decay parameter θ and FWER control do not depend on the user-specified cM spacing. Regions of excess IBD rate can be located with finer precision when using a smaller cM spacing. There is a trade-off though in compute time and the number of hypothesis tests to account for. We find that spacing hypothesis tests every 0.02 cM has high power, FWERs are close to the family-wise significance level, and computation is within tens of seconds. In an IBD mapping study, Chen et al. [34] perform hypothesis tests every 1.0 cM, which gives

limited precision in locating phenotype-genotype associations. In a selection scan, Palamara et al. [111] use a Bonferroni correction when performing hypothesis tests every 0.05 cM, which may be conservative if the tests are highly correlated.

We use five hundred whole-genome simulations to estimate FWERs for our hypothesis test. This simulation study is possible because detectable IBD segments can be derived directly from simulated ancestral recombination graphs. Many of the existing selection scans require marker data as input [51, 151, 155, 124], which is not feasible to simulate at a large scale due to our disk memory constraints. As a result, we do not know if their methods control FWER at nominal significance levels when multiple tests are performed. The area under the receiver operating curve from small simulated regions under positive selection versus no selection has been calculated for many existing selection scans [142, 155, 151, 84, 75, 103, 145, 140, 50]. This classification diagnostic is not the same as a formal hypothesis test and it does not account for multiple testing. We encourage formal hypothesis tests before conclusions about adaptive evolution are made.

Future work may include simulation studies that compare FWER among methods whose scanning statistics can be derived from ancestral recombination graphs (ARGs). The selection test in Speidel et al. [140] is one example, but they do not provide a multiple testing correction, and, in practice, one must infer the ARG. Inferring the ARG is extremely difficult and may lead to poor results compared to IBD-based methods (Chapter 5). The method of Palamara et al. [111] is the most similar method to our approach. They calculate the density of pairwise coalescent events within the last 20 or 150 generations. According to our work, the coalescent times of common ancestors whose descendant IBD segments ≥ 2.0 cM are mostly within the past 200 generations (Figures S4, S5, and 2.4). To compute p-values, they approximate their scanning statistic as a Gamma random variable. This approximation is not supported by theory whereas our scanning statistic is asymptotically normally distributed. They use a Bonferroni correction which could be very conservative if their scanning statistics are highly correlated, whereas our methods adjust for the spacing between hypothesis tests. Beyond controlling FWER, we could estimate the power of these tests

to reject the null hypothesis under alternative models. We estimate that our test has high power in selective sweeps where $s \geq 0.012$.

Scanning for IBD rates that exceed the mean or median IBD rate around a locus may be transferable to case-control studies. Specifically, we assume that the mean IBD rates in the case and control samples are the same under the null model and normally distributed in large samples. Next, we assume that these case-control differences in IBD rates along the genome follow an Ornstein-Uhlenbeck process. Finally, using the Siegmund and Yakir [137] analytical approximation or the simulation-based approach, we could derive a multiple testing adjusted significance level. As opposed to the IBD mapping test in Browning and Thompson [21] where a default value is set for an important parameter, we would learn an exponential decay parameter θ from the data. The interpretation of an alternative model would be similar to hard and recent selective sweeps: the case population has an inherited allele from some founder or small set of founders in the recent past.

Chapter 7

MODELING RECENT POSITIVE SELECTION IN HUMANS

7.1 Introduction

Selective sweep events can leave profound footprints on genomic diversity around a selected locus. Extended haplotype homozygosity about the LCT gene in northern Europeans spans multiple megabases (Mb) of genetic sequence [124]. In the HLA region, the probability that two people in a European ancestry sample share alleles identical by descent has been estimated to be more than six percent [3]. Studying positive selection can be complementary to large association studies; for instance, loci associated with inflammatory disorders show evidence of strong and weak selection [112]. The purpose, magnitude, and prevalence of such adaptive evolution remain open questions in population genetics, medical genetics, and conservation genetics.

In this chapter, using our methods in Chapters 4, 5, and 6, we perform an entire analysis of positive selection in some human populations. We describe the datasets and our quality control procedures in Section 7.2. Next, in Section 7.3, we derive a multiple testing adjusted significance threshold for different ancestry-specific sample sets. In Section 7.4, our new multiple testing adjusted significance thresholds are juxtaposed to our initial heuristic thresholds in Temple et al. [148]. In Section 7.5, we model selection in Europeans for a few loci where there is an abundance of evidence supporting the recent and hard selective sweep model. We close with a discussion on how the speed and automation of our analysis and the rigor in our hypothesis testing can promote reproducible science.

7.2 Pre-processing and quality control of real data

In our study, we focus on selection scans and modeling selective sweeps in African, European, and South Asian ancestry groups from the Trans-Omic for Precision Medicine (TOPMed)

project [146] and the United Kingdom Biobank (UKBB) [24]. We use the 318,858,817 filtered autosomal markers from the TOPMed phased data in Browning et al. [18]. The TOPMed database includes more than thirty thousand whole genome sequences from multiple ethnic groups represented in the United States of America. The TOPMed data combines samples from multiple cohort studies. Abbreviations for each cohort study are in Appendix F. UKBB is a biomedical database containing genotype array data from nearly five hundred thousand participants between 40 and 69 years of age. We use the 711,651 filtered autosomal markers from the UKBB phased data in Browning et al. [18]. The TOPMed and UKBB datasets are kept separate in all analyses.

The selective sweep model we consider assumes no population structure and no familial relatedness. To mitigate the possible confounding effects of population structure and familial relatedness, we form subsets of the TOPMed data and UKBB data representative of one continental ancestry and without first-degree relatives. The following subsections provide details about our subsampling procedures.

7.2.1 TOPMed European and African ancestry samples

We analyze the 38,387 whole genome sequences that are haplotype phased with `Beagle` version 5.2 in Browning et al. [18]. We compute principal components (PCs) using the `GENESIS` [60] and `SNPRelate` [161] software programs. Based on data visualization, we define two European (EUR and EUR2) and African (AFR) inferred ancestry groups by determining bounds on PCs 1-4. Many samples in TOPMed have intermediate admixture proportions belonging to different continental ancestry groups. Moreover, plotting PCs 1 and 2, there are two apparent clusters containing samples that predominantly self-report as White. The visualization-informed clustering aims to create subsets representing broad inferred continental ancestry with low levels of differentiation.

The EUR1 ancestry group consists of 13,778 samples. From self-reporting, this group consists of 4172 White, 9583 Other, and 23 non-White samples. From the specific cohort studies, there are 7682 WHI, 2157 MLOF, 1217 VTE, 1108 BioMe, 976 VUAF, 345 FHS,

287 CCAF, and 1 HyperGen samples in the EUR1 group. Figure S57 displays PCs 1 and 2 for these samples split into their respective cohort studies and the combined group.

The EUR2 ancestry group consists of 1719 samples. Sixty-four percent of these samples come from the BioMe Biobank cohort study at Mt. Sinai School of Medicine in New York City. Thirty-five percent of these samples self-report as Other. Sixty-four percent of these samples self-report as White. For this group, we infer a demographic history that sharply drops to an effective size as small as one thousand in the most recent thirty generations (`IBDNe` using $\geq$ 2.0 cM IBD segments). Tian et al. [150] also infer a severe bottleneck in the past thirty generations for a subset of the samples in the Framingham Heart Study. In an Ashkenazi Jewish sample, Carmi et al. [30] infer a recent bottleneck of effective size a few hundred diploids, which Tian et al. [150] say is consistent with their demographic inference. Carmi et al. [30] state that the Ashkenazi Jewish population is most genetically similar to European and Middle Eastern populations, which is consistent with our principal components analysis.

The inferred AFR ancestry group consists of 1737 samples. Fifty-four percent of these samples self-report as Black or African American. Forty-six percent of these samples self-report as Other. Only samples from the BAGS, JHS, and HyperGen cohorts are represented in this subset. Afro-Caribbeans living in Barbados are in the BAGS study, whereas African-Americans living in the southern continental United States are in the JHS and HyperGen studies.

To validate these continental ancestry clusters, we use the 1000 Genomes [10, 25] and Human Genome Diversity Panel (HGDP) [28, 12] datasets to perform supervised global ancestry inference with `ADMIXTURE` [5]. We use the reference panels CEU (Utah residents with Northern and Western European ancestry), CHB (Han Chinese in Beijing, China), GIH (Gujarati Indians in Houston, TX), and YRI (Yoruba in Ibadan, Nigeria) from 1000 Genomes as representing European, East Asian, South Asian, and African ancestry, respectively, and 61 Indigenous American samples from HGDP. For samples in the EUR1 group, the inferred global ancestry with respect to the CEU reference panel is minimum 73%, interquartile range

(90%, 97%), and mean 93%. For samples in the AFR group, the inferred global ancestry with respect to the YRI reference panel is minimum 88%, interquartile range (91%, 95%), and mean 93%. Other subsets of majority global ancestry with respect to the CHB, GIH, and Indigenous American reference panels have less than one thousand samples. We do not analyze these subsets. We consider two subsets EUR95 and EUR98 of the EUR1 group that comprise individuals with at least 95% and 98% inferred CEU ancestry, respectively.

In each subset, kinship is estimated using `IBDkin` with input IBD segments longer than 2.0 cM [164]. If diploid individuals i, j have a kinship coefficient greater than 0.125 (the expected kinship of first cousins), we add an edge between the diploid individuals in forming a graph of connected components. We take one sample from each connected component, noting that most samples have no pairwise kinship coefficients greater than 0.125.

We performed these pre-processing steps in mid-2023. We have since developed a bioinformatics pipeline to do haplotype phasing, IBD segment detection, and local ancestry inference[1]. The detected IBD segments and local ancestry inference from this automated workflow can be used to select representative subsets as done here. The detected IBD segments from this bioinformatics pipeline can be used to filter on close familial relatedness with `IBDkin`. We can aggregate over the local ancestry inference[2] in the bioinformatics pipeline to filter on global ancestry proportions. This step would be similar to our global ancestry `ADMIXTURE` analysis but distinct from our PC-based clustering of the EUR1 and EUR2 groups.

7.2.2 UK Biobank self-report samples

We analyze a subset of these samples who self-report as various non-White ethnic groups. The first subset includes 5660 individuals who self-report as Indian. The second subset includes 3202 individuals who self-report as Black British. The sample sets are haplotype phased individually with `Beagle` version 5.4. Based on genetic relatedness inference in Cai et al. [26], we remove closely related individuals from both subsets, resulting in 5374 Indian

[1] https://github.com/sdtemple/flare-pipeline

[2] We use `FLARE` for local ancestry inference [23].

samples and 3154 Black British samples. We do not perform a global ancestry analysis for these self-report sample sets.

We also study self-report White British datasets of size two, five, ten, twenty, and fifty thousand samples. These five datasets are randomly selected subsets of the 408,833 White British samples studied in Browning and Browning [20]. The array data is haplotype phased with `Beagle 5.2` as described in Browning et al. [18].

7.3 Scanning statistic thresholds

7.3.1 Detecting identity-by-descent segments

To detect IBD segments in the TOPMed sample sets, we use the 2019 pedigree-based map from deCODE Genetics [70]. This recombination map is aligned to the GRCh38 reference genome. For the TOPMed sequence data aligned to the GRCh38 reference genome, we use the algorithm parameters in Table S5, which are the same parameters we use in our simulation studies (Chapter 5). In the EUR1 group, we perform a preliminary analysis of chromosomes 19 to 22 with `ibd-ends` to get an estimate of the error rate parameter. In the main analysis, we specify the error rate err=1.5e-4 instead of the software default setting.

To detect IBD segments in the UKBB sample sets, we use the Bhérer et al. [15] pedigree-based map. This recombination map is aligned to the GRCh37 reference genome. For the UKBB array data aligned to the GRCh37 reference genome, we modify our `hap-ibd` settings to min-seed=1.8, min-extend=0.5, min-output=1.8, and a minor allele frequency of 0.001 [3]. In the Black British group, we perform a preliminary analysis of chromosomes 19 to 22 with `ibd-ends` to get an estimate of the error rate parameter. In the main analysis, we specify the error rate err=3e-4 instead of the software default setting.

[3] We have not explored these algorithm settings in simulated array data. We show in the following sections that our analyses of array data are consistent with our analyses of sequence data and with the existing literature on some selected loci.

7.3.2 Heuristic scan threshold

Here we describe the heuristic approach to scan for excess IBD rates in our initial work [20, 148]. The detectable IBD rate is the number of IBD segments longer than 2.0 cM divided by the number of haplotype pairs. Every 20 kb we calculate the IBD rate based on detected segments that overlap the focal position. Positions where the IBD rate exceeds the genome-wide median plus four times the standard deviation are considered loci with excess IBD rates. Compared to our scanning procedure in Chapter 6, this scan uses base pair distance, not cM distance. We find that the standard deviations for IBD rates every 20 kb are typically larger than standard deviations for IBD rates every 0.02 cM, but the multiple testing adjusted thresholds based on Chapter 6 are typically larger than four standard deviations. Compared to the selection scan in Albrechtsen et al. [3], our selection scan concerns long IBD segments overlapping a focal position, rather than the probability that alleles are IBD at a marker location.

In Temple et al. [148], we make three additional modifications. First, before computing the genome-wide standard deviation, we exclude positions that are initially three standard deviations below or above the genome-wide median IBD rate around a locus. Second, we ignore positions that are within 0.5 cM of the start and end of the recombination map for each chromosome. These two quality control measures exclude data near telomeres and centromeres where it is challenging to accurately detect long IBD segments. Third, we require that around loci of interest IBD rates must exceed our heuristic threshold for a 1.0 cM contiguous stretch. Similar to our max-min analysis in Chapter 6, loci that marginally exceed our heuristic threshold for a small stretch of a chromosome could be false positives.

7.3.3 Multiple testing significance threshold

For each subset, we compute IBD rates every 0.02 cM and apply the multiple testing adjustments described in Chapter 6, Section 6.2. Figure S58 provides some evidence that the empirical distributions of IBD rates around a locus resemble normal distributions in our sample

sets[4]. Figure 7.1 shows chromosome-specific estimated covariances and the fitted exponential curve with estimates $\hat{\theta}$ for the TOPMed EUR1 group, the TOPMed EUR2 group, and the UKBB Indian self-report group. Alongside the European ancestry and Indian self-report exponential decay curves, we also show the estimated covariances and fitted exponential curve for one of our population bottleneck simulations in Chapter 6 (Figure 7.1)[5]. Figure S59 shows chromosome-specific estimated covariances and the fitted exponential curve with estimates $\hat{\theta}$ for the TOPMed inferred AFR group and the UKBB Black British self-report group. Upon visual inspection, the fitted exponential curves fit the estimated covariances well in all plots, except in the TOPMed EUR2 the covariance decays are slower than implied by $\hat{\theta}$.

The exponential decay parameter estimates $\hat{\theta}$ are 45, 30, 49, 83, and 78 for the TOPMed EUR1, TOPMed EUR2, UKBB Indian, TOPMed AFR, and UKBB Black British groups, respectively. The exponential decay parameter estimates $\hat{\theta}$ are 44 and 44 for the TOPMed EUR95 and EUR98 groups, respectively. The exponential decay parameter estimates $\hat{\theta}$ are 51, 51, and 50 for the UKBB White British 2k, 5k, and 10k groups, respectively [6].

7.4 Genome-wide selection scans

7.4.1 TOPMed European ancestry analysis

Figure 7.2A shows the IBD rates along the autosomes, the genome-wide median, the heuristic four standard deviations threshold, and the multiple testing adjusted threshold for the EUR1 group. The multiple testing adjusted threshold is greater than the heuristic four standard deviations threshold. Figure S60 exhibits similar results among the five study-specific cohorts. The four largest signals at LCT, MHC, TRPM1, and OAS are thirty-five, twenty-six,

[4]Recall that normality is a property of the Ornstein-Uhlenbeck process.

[5]Recall that the population bottleneck is based on recent effective sizes of European-Americans.

[6]Recall that the exponential decay parameter θ depends on the IBD segment detection threshold (Section 6.3), and that population demography influences the IBD segment length distribution (Figures S4 and S5). These two observations may explain the different estimates $\hat{\theta}$ among our ancestry groups.

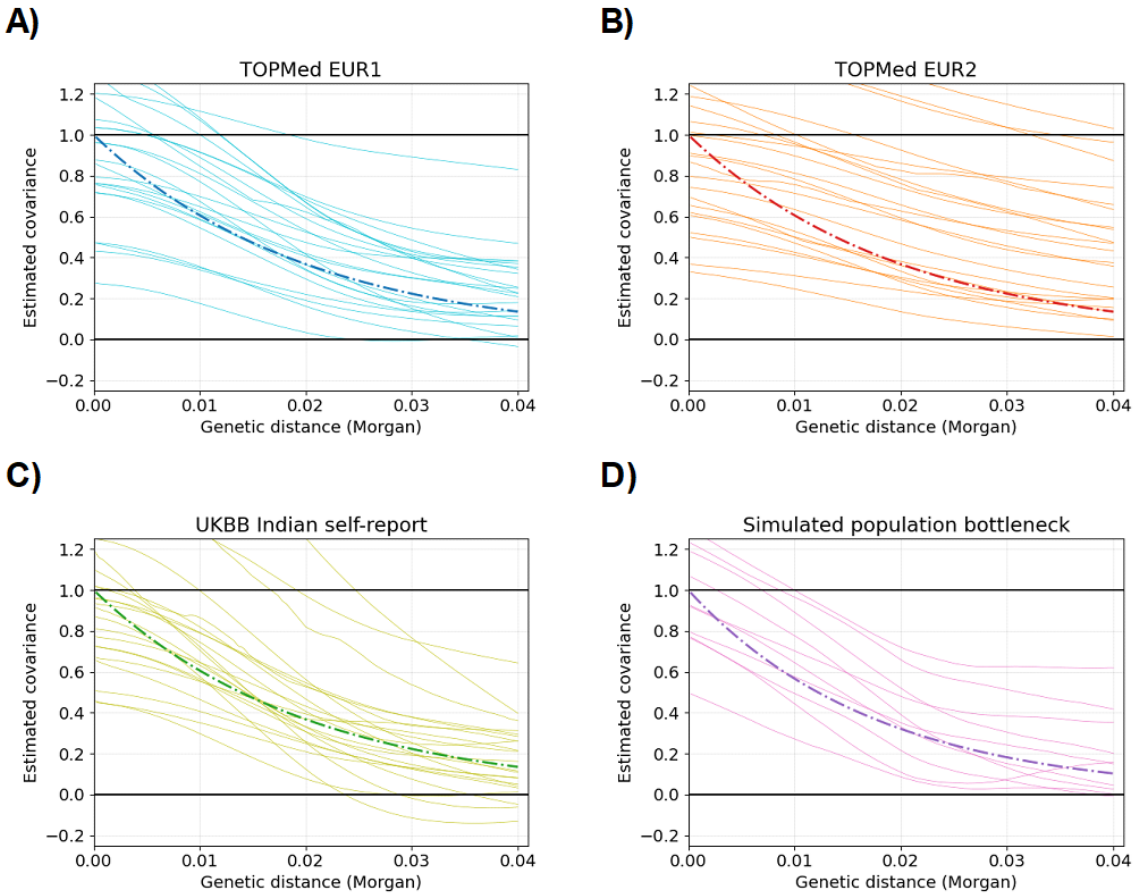

Figure 7.1: Estimating exponential decay parameter θ in real and simulated data. Each faint colored line shows estimated covariances (y-axis) for different cM distances (x-axis) and a specific chromosome. Dark-colored dashed lines show the predicted covariances from estimates $\hat{\theta}$ in fitted Ornstein-Uhlenbeck processes. The title of each subplot gives the names of the sample sets A) TOPMed EUR1, B) TOPMed EUR2, C) UKBB Indian, and D) a simulated population bottleneck. There are twenty-two chromosomes in human autosomal data and ten chromosomes in simulated data of the population bottleneck demography.

ten, and eight standard deviations above the genome-wide median IBD rate, respectively. All six analyses observe excess IBD rates at LCT, MHC, and TRPM1, and five of them observe excess IBD rates at OAS. Figure S61 exhibits similar results as well for our EUR1 subsets with 95% and 98% CEU global ancestry proportion. This observation suggests that our IBD-based selection scan may be robust to mild levels of population admixture.

Table 7.1 summarizes the results of our selection scan in the TOPMed EUR1 group. We find eight regions where IBD rates exceed four standard deviations above the genome-wide median and three or more cohort-specific analyses replicate the signal. In the combined TOPMed EUR1 group, IBD rates at seven of these loci exceed our multiple testing adjusted threshold of IBD rate 1.94e-4 at the 0.05 family-wise significance level, including signals near the LCT, MHC, and TRPM1 genes. The four standard deviations threshold is 1.95e-4 in the TOPMed EUR1 analysis. The selection coefficient in which the expected IBD rate equals the genome-wide significance level is 0.0172 [7] (Figure S62).

Seven of these eight consensus loci appear in prior selection scans on White British individuals in the UK Biobank [20, 104, 111] and pan-European analyses using ancient DNA [101, 100]. The signal between 205 to 207 Mb on chromosome 1 appears in the EUR1, WHI, and MLOF groups, but it does not appear in any of the cited works above. The inconsistencies within our cohort-specific analyses and across separate European ancestry analyses from other authors apply to the selection signals which marginally pass our heuristic four standard deviations threshold.

Figure S63 shows the IBD rates along the autosomes, the genome-wide median, the heuristic four standard deviations threshold, and the multiple testing adjusted threshold for the EUR2 group. The IBD rate near the MHC region narrowly exceeds the multiple testing threshold for a few kb. The genome-wide median IBD rate in the TOPMed EUR2 group is more than an order of magnitude greater than the genome-wide median IBD rates in all of our other ancestry groups. The variance in the IBD rate process is much larger when recent

[7] The expected IBD rate depends on selection coefficient s and allele frequency $p(0)$. For a given s, we average over the expected IBD rates conditional on $p(0) \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$.

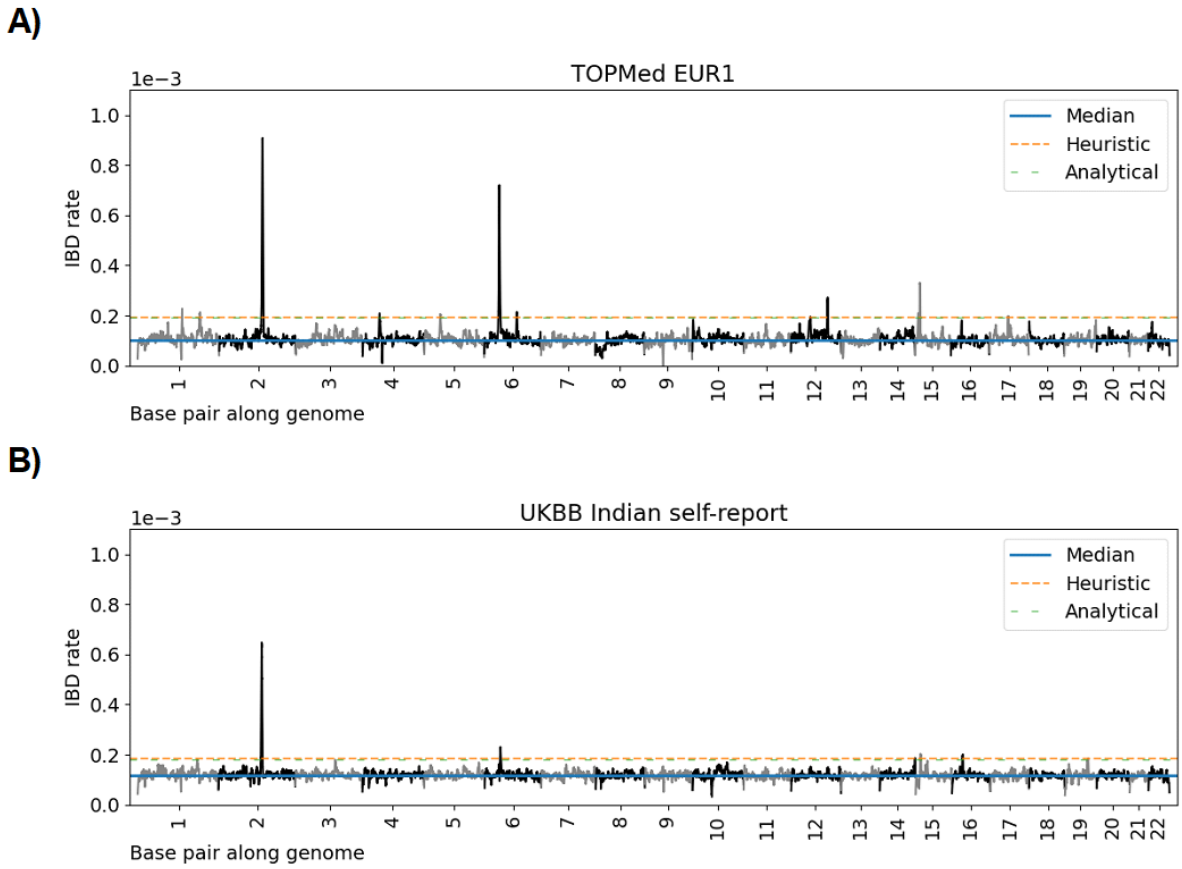

Figure 7.2: Genome-wide IBD rate scans in European and South Asian ancestry data. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on A) inferred European ancestry samples in the TOPMed project and B) self-reported Indian samples in the UK Biobank. The main text describes the EUR1 ancestry group. Each subplot has a different y-axis scale. Horizontal dashed lines show (blue) the genome-wide median IBD rate, (orange) the heuristic threshold of four standard deviations above the median IBD rate, and (green) the Siegmund and Yakir discrete approximation (S&Y). The S&Y method is calculated assuming hypothesis testing every 0.02 cM.

Chr	Max rate (1e-4)	cM	Position (Mb)	# studies	Genes
2	9.09	141.79	134.84 (132.52-139.90)	6 / 6	LCT
6	7.20	49.35	30.80 (24.12-36.12)	6 / 6	MHC
15	3.31	15.42	31.18 (30.34-32.16)	6 / 6	TRPM1
12	2.72	120.54	113.08 (110.90-113.64)	5 / 6	OAS1-2-3
6	2.14	107.11	105.98 (103.76-108.46)	3 / 6	PRDM1
1	2.13	205.05	206.62 (205.50-207)	3 / 6	.
5	2.06	53.34	33.96 (31.00-35.98)	3 / 6	SLC45A2
15	2.22	10.19	28.10 (25.94-30.86)	3 / 6	OCA2

Table 7.1: Eight regions highlighted in selection scan on TOPMed European Americans. Loci where the identity-by-descent (IBD) rate exceeds the multiple testing adjusted threshold 1.94e-4 at the 0.05 family-wise significance level. Physical and genetic positions for the location of maximum IBD rate are shown in megabases (Mb) and centiMorgans (cM). The intervals denote the range over which IBD rates exceed the scan threshold. Locations are aligned to build GRCh38. The 2019 pedigree-based recombination map from deCODE genetics is used when inferring IBD segments [70]. Genes of interest are annotated for regions discussed in the main text. We also count the number of studies in which IBD rates in the region exceed our heuristic four standard deviations threshold. Rows are based on analysis of the 13,778 samples in the EUR1 group, except the separated row for OCA2 is based on the 7,682 samples in the WHI cohort.

effective population sizes are smaller, which may decrease the power of our test.

7.4.2 UK Biobank White self-report analysis

Figure S64 shows the IBD rates along the autosomes, the genome-wide median, the heuristic four standard deviations threshold, and the multiple testing adjusted threshold for the 2k, 5k, and 10k subsets of UKBB White self-report samples. In all three scans, we observe

IBD rates exceeding our genome-wide significance threshold at the LCT, MHC, OAS1-2-3, and TRPM1 loci. Our results are also consistent with the selection scan in Browning and Browning [20], in which more than four hundred thousand White British samples are studied.

We contrast the impact of sample size in association studies versus our IBD-based selection scan. In association studies, the null hypothesis is that a linear effect β is zero, and the alternative hypothesis is that $\beta \neq 0$. The power to detect $\beta \neq 0$ increases with sample size in GWAS because the standard error of $\hat{\beta}$ decreases. In our IBD-based selection scan, we only require enough data to reliably estimate the mean, standard deviation, and exponential decay parameters in the Ornstein-Uhlenbeck process, after which collecting more samples leads to diminishing returns.

For our selection scans on the 5k, 10k, 20k, and 50k subsets, we also observe a significant locus near the CCR9 gene. The maximum IBD rate in this region is the second greatest in Browning and Browning [20], where all UKBB White British samples are analyzed. This chemokine receptor is known to play an important role in the mucosal immune system [113] and has been associated with increased COVID-19 outcome severity, especially in Europeans [133]. At this locus, Browning et al. [17] and Ding et al. [41] suggest that introgressed Neanderthal haplotypes may be selected for in South Asians and East Asians, respectively. These studies observe introgressed haplotypes that are at less than ten percent frequency in the CEU population but at more than thirty-five percent frequency in South and East Asians. Our combined and cohort-specific analyses of the TOPMed EUR1 group do not replicate this significant result, nor does the following analysis of the UKBB Indian self-report samples.

7.4.3 UK Biobank Indian self-report analysis

Figure 7.2B shows the IBD rates along the autosomes, the genome-wide median, the heuristic four standard deviations threshold, and the multiple testing adjusted threshold for the UKBB Indian self-report group. Table S11 reports four loci where IBD rates exceed the multiple testing adjusted threshold 2.07e-4. Similar to the TOPMed EUR1 group analysis, IBD rates near LCT, MHC, and TRPM1 exceed the multiple testing adjusted threshold.

Gallego Romero et al. [55] suggest that northern European haplotypes carrying a putatively selected allele at LCT may be identical by descent to haplotypes in Indian pastoralists that carry the same selected allele. Using methods in Chapter 5, we infer an excess IBD outgroup comprising roughly seventeen percent of the samples, which would be in the range of the selected allele frequency in Gallego Romero et al. [55]. IBD rates surrounding HLA genes are known to be high in all HapMap populations [3], which we also find in our selection scan at MHC. The OCA pigmentation gene does not exceed our significance threshold in our self-report Indian samples, albeit this region is said to be under selection in many human populations [85]. On the other hand, the TRPM1 pigmentation gene a few Mb to the right exceeds our significance threshold in the TOPMed EUR1 group and the UKBB Indian self-report group. IBD rates in the TOPMed EUR1 group narrowly do not exceed four standard deviations at the statistically significant genomic region on chromosome 16 in our UKBB Indian self-report analysis.

7.4.4 TOPMed African ancestry analysis

Figure 7.3A shows the IBD rates along the autosomes, the genome-wide median, the heuristic four standard deviations threshold, and the multiple testing adjusted threshold for the TOPMed AFR group. Twenty-four loci exceed our heuristic four standard deviations threshold. Only five of these twenty-four loci exceed our multiple testing adjusted threshold of IBD rate 2.63e-4 at the 0.05 family-wise significance level, which we report in Table 7.2. Moreover, recall that our multiple testing adjusted threshold is anti-conservative in our simulation studies (Chapter 6). These nineteen loci that pass our heuristic threshold but not our multiple testing threshold could be false positives. The excess IBD rates on chromosome 19 should be interpreted with caution because detecting IBD segments near telomere ends can be sensitive.

At a locus on chromosome 16, we observe an excess IBD rate that is 16.94 standard deviations above the genome-wide median. This locus is near to the excess IBD rates on

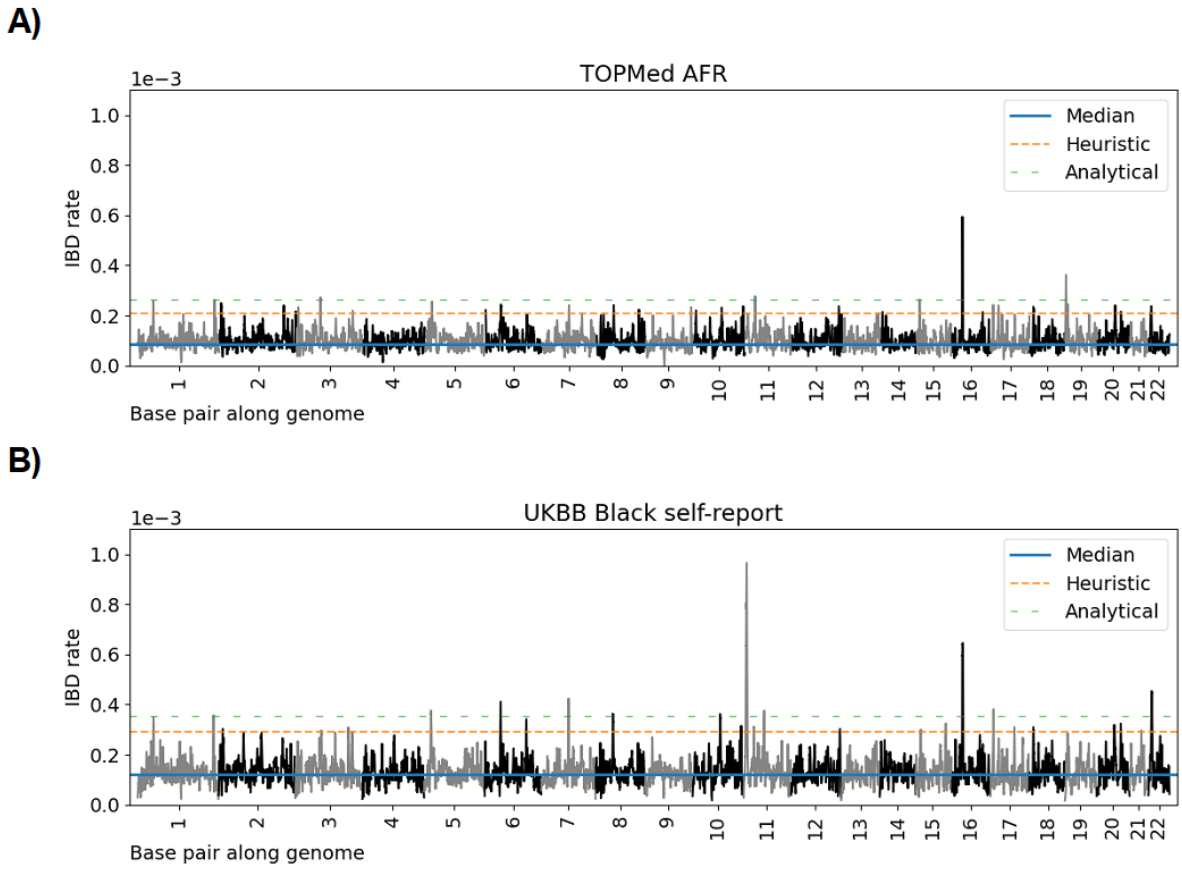

Figure 7.3: Genome-wide IBD rate scans in African ancestry data. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on A) inferred African ancestry samples in the TOPMed project and B) self-reported Black samples in the UK Biobank. Each subplot has a different y-axis scale. Horizontal dashed lines show (blue) the genome-wide median IBD rate, (orange) the heuristic threshold of four standard deviations above the median, and (green) the Siegmund and Yakir discrete approximation (S&Y). The S&Y method is calculated assuming hypothesis testing every 0.02 cM.

Chr	Max IBD rate	Position (cM)	Position (Mb)
16	5.93E-04	34.45	17.00
19	3.62E-04	4.29	1.74
11	2.77E-04	33.19	19.92
3	2.73E-04	79.06	60.64
15	2.64E-04	11.26	28.98

Table 7.2: Regions highlighted in TOPMed African ancestry selection scan. Loci where the identity-by-descent (IBD) rate exceeds the multiple testing adjusted threshold of 2.63e-4 at the 0.05 family-wise significance level. Physical and genetic positions for the location of maximum IBD rate are shown in megabases (Mb) and centiMorgans (cM). Locations are aligned to build GRCh38. The pedigree-based recombination map from Halldorsson et al. [70] is used when inferring IBD segments.

chromosome 16 in our UKBB Indian self-report analysis[8]. Similar to IBD rates at MHC, the IBD rates in this chromosome 16 region are high in all continental ancestry groups, but they are exceptionally high in the TOPMed AFR ancestry group.

The base pair of the highest IBD rate on chromosome 16 is 100 kb from the XYLT1 gene, and excess IBD rates cover this gene entirely. The enzymes xylosyltransferases 1 and 2 initiate a chain reaction in the early maturation of skeletal cells. Linkage analysis in a consanguineous Turkish family associated a recessive missense mutation with a short stature syndrome [130]. Mutagenesis screening of mice also demonstrated disproportionate dwarfism from a recessive missense mutation in XYLT1 [102].

Studying African ancestry samples in the TOPMed data, Taliun et al. [146] indicate a few loci that may be under selection. They use singleton density score [51] as their scanning statistic with the 5e-8 significance threshold. Singleton density score assumes a convenient

[8]After accounting for difference in references GRCh37 versus GRCh38

Gamma family distribution to compute p-values, and the 5e-8 significance threshold may not be appropriate for sequence data. We do not observe excess IBD rates at any of the loci suggested to be under selection by Taliun et al. [146].

The recombination map we use here is based on Icelandic pedigrees, not African ancestry pedigrees. We explore the use of LD-based recombination maps specifically designed for African ancestry populations. We use the maps from Spence and Song [141] based on the ASW and ACB cohorts in 1000 Genomes. Figure S65 shows the IBD rates along the autosomes, the genome-wide median, and the heuristic four standard deviations threshold using these LD-based recombination maps. We infer overall IBD rates that are an order of magnitude smaller than those in our pedigree-based analysis. There appear to be many false positives near the telomeres and centromeres as well, which is consistent with the observations Browning and Browning [20] make in their IBD-based selection scan with the HapMap LD-based genetic map [82]. We also fail to observe excess IBD rates near the XYLT1 gene, which may be a result of lower estimated recombination rates in the LD-based genetic map as opposed to the pedigree-based map (Table S13). Albeit the deCODE genetic map is derived from European ancestry pedigrees, we prefer this pedigree-based map in our African ancestry selection scan over the African ancestry-specific LD-based maps.

7.4.5 UK Biobank Black self-report analysis

Figure 7.3B shows the IBD rates along the autosomes, the genome-wide median, the heuristic four standard deviations threshold, and the multiple testing adjusted threshold for the UKBB Black self-report group. Nineteen loci exceed our heuristic four standard deviations threshold. Only eleven of these nineteen loci exceed our multiple testing adjusted threshold of IBD rate 3.54e-4 at the 0.05 family-wise significance level, which we report in Table S12. The four standard deviations threshold is equivalent to an IBD rate of 2.91e-4. The excess IBD rates on chromosome 11, which are the highest in our UKBB Black self-report analysis, should be interpreted with caution because detecting IBD segments near telomere ends can be sensitive. The inflated IBD rates near the XYLT1 gene are reflected in the UKBB Black

self-report group as with the TOPMed AFR group. Other loci exceeding the genome-wide significance level are not replicated in our analysis of the TOPMed AFR group (Table 7.2).

We explore the use of IBD-based recombination maps specifically designed for African ancestry populations. We use the Jackson Heart Study (JHS) maps from Zhou et al. [162]. Figure S66 shows the IBD rates along the autosomes, the genome-wide median, and the heuristic four standard deviations threshold using IBD-based recombination maps. We infer overall IBD rates of the same magnitude as those in our pedigree-based analysis. However, there appear to be many false positives near the telomeres and centromeres, which is consistent with the observations Browning and Browning [20] make in their IBD-based selection scan with the IBD-based genetic maps. We also fail to observe excess IBD rates near the XYLT1 gene, which may be a result of lower estimated recombination rates in the IBD-based genetic map as opposed to the pedigree-based map (Table S13). Albeit the deCODE genetic map is derived from European ancestry pedigrees, this pedigree-based map is preferred in our African ancestry selection scan over the African ancestry-specific IBD-based maps.

7.5 Putative selective sweeps in Europeans

We apply our suite of methods in Chapter 5 to model recent selective sweeps in Europeans. We do not model any recent selective sweeps in African ancestry populations. The excess IBD rates at XYLT1 are the only signal in those analyses that we have strong confidence in, and, as stated previously, we suspect this signal is not the result of a hard sweep. We do not model any recent selective sweeps in the self-report Indians because we have not evaluated the accuracy of our algorithm settings in simulated array data.

To infer the nonparametric effective sizes $\hat{N}_e(t)$, we use inferred IBD segments ≥ 2.0 cM as input to `IBDNe` with default settings [19]. To estimate the sweeping allele frequency, we run our analysis pipeline with default settings. To estimate the selection coefficient, we use Equation 4.6, conditioning on the recent effective size estimates $\hat{N}_e(t)$ and the allele frequency estimate $\hat{p}(0)$. Unless otherwise specified, we assume the additive genetic model $1{:}(1+s){:}(1+2s)$.

7.5.1 Modeling LCT selection

Lactase persistence is widely believed to be subject to a selective sweep in European ancestry populations [132]. We view it as a positive control to evaluate our method. The putative selected allele is a regulatory mutation -13.910C>T upstream of LCT (OMIM: 603202) in intron 13 of MCM6 (OMIM: 601806) [132]. This allele enhances the promoter of LCT, allowing an alternative path for gene expression.

Here the individual and aggregate analyses show nearly perfect concordance. Gini impurity indices are zero, and there are singular clusters of excess IBD rate that comprise fifty percent or more of the samples. Five of six analyses rank the -13.910C>T mutation as the maximally differentiated SNP between the inferred outgroup and the rest of the sample. We apply `iSAFE` to rank SNPs in this region, and it also ranks lowest the -13.910C>T mutation. Figure S67A depicts scores for the SNPs flanking the causal mutation. Our allele frequency estimate without prior knowledge of the selected allele is $\hat{p}(0) = 0.725$, while the frequency of 13.910C>T is 0.6864. We estimate the LCT selection coefficient using the frequency of the consensus -13.910C>T mutation.

The lactase persistence phenotype is dominant but the enzymatic activity is additive [100]. What phenotype is selected for and which form of genic selection is never known. Given the additive model, we estimate $\hat{s}_a = 0.0325$ (95% CI = (0.0278,0.0373)). Table S14 gives the additive model estimates for BioMe, MLOF, VTE, VUAF, and WHI cohorts. Given the dominance model, we estimate $\hat{s}_d = 0.0488$ (95% CI = (0.0423,0.0552)).

Figure 7.4A shows the historical allele frequencies implied by $\hat{s}_a$ and $\hat{s}_d$. To display historical allele frequencies, we draw parametric bootstraps over selection coefficients and simulate allele frequencies backward in time with selection and genetic drift. This bootstrap represents uncertainty in the IBD process and genetic drift. Mathieson and Terhorst [100] and Vaughn and Nielsen [152] model LCT selection with time-varying coefficients. Their approaches use ancient DNA; our method does not leverage temporal information, and thus estimating time-varying selection coefficients is out of its scope. However, the historical allele

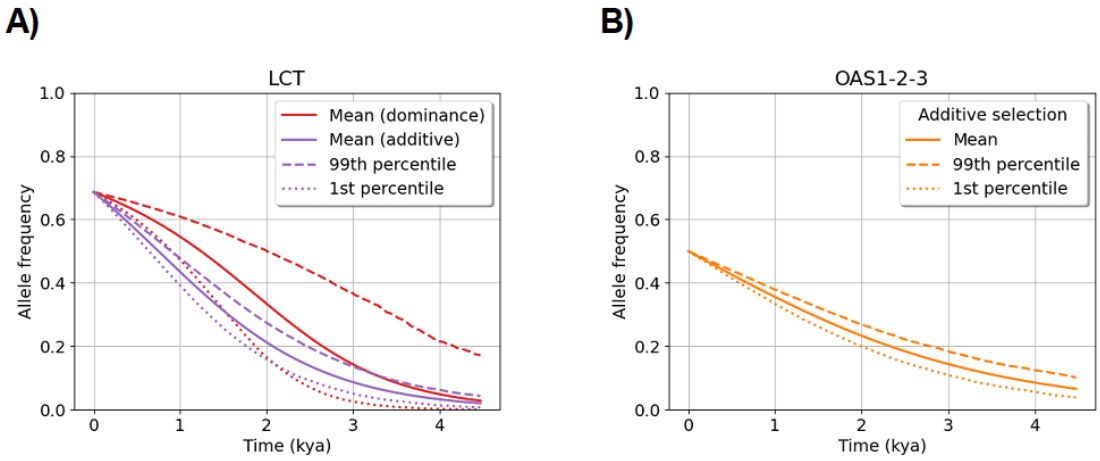

Figure 7.4: Estimated historical allele frequencies in last 150 generations for LCT and OAS1-2-3 loci putatively under selection in Europeans. The plots show the means, the first percentiles, and the ninety-ninth percentiles for allele frequencies $p(t)$ from one thousand parametric bootstraps over identity-by-descent and Wright-Fisher processes for selective sweeps at (A) LCT and (B) OAS 1-2-3 loci. The LCT analysis is based on the best-ranked SNP because there is consensus across cohort analyses. The OAS1-2-3 analysis is based on haplotype-based frequency estimates because there is no consensus on a variant. LCT and OAS1-2-3 analyses are based on IBD segments in the first European ancestry group (EUR1). Additive selection is assumed, but dominance selection is also shown for LCT. Time is measured in thousands of years ago (kya) using one generation equals thirty years.

frequencies from our dominance model closely match the curve in Mathieson and Terhorst [100]. Compared to the other loci, the sweep at LCT may have progressed more rapidly than any other putative sweeps in Europeans.

7.5.2 Possible selective sweeps on introgressed haplotypes in OAS 1-2-3 genes

Studies on archaic introgression have identified a Neanderthal haplotype spanning the OAS genes 1, 2, and 3 (OMIM: 164350, 603350, 603351). These genes transcribe antiviral proteins

and may be an example of selection for immune response [129, 17]. The introgressed haplotype harbors a splice variant that upregulates the expression of a specific isoform [59]. In our six analyses, the locations of maximum IBD rate are 1.5 Mb to the left of OAS1-2-3, instead lying within the LINC02356 (Ensembl: ENSG00000257595), SH2B3 (OMIM: 605093), and ATXN2 (OMIM: 601517) genes. Figure 7.4B shows a long haplotype stretching between these genes all the way to the OAS genes around chr12:113,000,000.

Applying our estimation methods, we arrive at present-day frequency estimates between 0.40 and 0.50. Previous studies have estimated the introgressed Neanderthal haplotype to have a frequency of approximately forty percent among modern Europeans [129, 59, 17]. In Table S14, three of the cohort analyses rank as first the SNP at base pair chr12:111,270,654. This marker is in the CUX2 gene (OMIM: 610648). We could not apply `iSAFE` in this region because the method gives an error flag that marker density is too low. Gini impurity indices are roughly 0.60 across our five analyses. These Gini impurity indices do not rule out OAS1-2-3 as a sweep locus, but there is weaker evidence for a selective sweep here than at LCT. For the EUR1 group, we estimate $\hat{s}$ conditional on the haplotype-based frequency estimate $\hat{p}(0) = 0.50$, which is close to the frequency of consensus best-ranked SNP from three of five studies. Our estimate of additive selection is $\hat{s} = 0.0182$ (95% CI = (0.0156,0.208)). According to our simulation study, this selection coefficient estimate could be inflated.

The evolutionarily conserved haplotypes in the OAS1-2-3 region differ from our modeling assumptions of a single causal allele. We emphasize that the causal allele in our model is not necessarily a SNP variant but could be a haplotype with high LD. For instance, there could have been a soft sweep with multiple advantageous alleles, and then a more advantageous allele is introgressed in whose haplotype subsequently sweeps towards fixation. Our statistical method can estimate selection coefficients in the hardening of such a soft sweep from standing variation (Table S4), but an interpretation of adaptive introgression is out of its scope. The implied historical allele frequencies are in Figure 7.4B. They assume the sweep from *de novo* mutation, but the curve in more recent time periods should not depend on this assumption.

7.5.3 Possible Selective Sweeps around Pigmentation Genes OCA2 and TRPM1

Alleles in pigmentation genes could improve vitamin D levels at northern latitudes. The chromosome 15 region containing TRPM1 (OMIM: 603576) and OCA2 (OMIM: 203200) appears in our IBD-based selection scan as in a related report [20]. In particular, OCA2 has been studied in numerous selection studies [100, 142, 75]. These genes are expressed in retinal cells and melanocytes and have been linked to eye and skin color phenotypes [8, 143].

For TRPM1, our analysis in each cohort ranks first the SNP at base pair 31,437,832 (Table S14). This is the only example in our work where each of the cohorts ranks best the same SNP. This marker is in a splice variant of the KLF13 (OMIM: 605328). Its frequency ranges from 11-13%, and its location is nearly 300 kb to the right of TRPM1. We perform estimation conditional on the allele frequency of this consensus best-ranked SNP and additive selection. We estimate selection coefficients between 0.026 and 0.030 for all of the five cohort-specific analyses. In the EUR1 group, we estimate $\hat{s} = 0.0273$ (95% CI = (0.0200,0.0350)). The implied historical allele frequencies are in Figure 7.5B.

The neighboring gene OCA2 passes our selection scans in only three cohorts. This locus does not appear in the EUR1 group analysis because its excess IBD rates span only 0.95 cM. SNPs in this region are known to inhibit gene regulation and affect the blue eye color phenotype [153, 42]. In Table S14, we compute Gini impurity indices of zero or below 0.60 for these analyses. However, different SNPs are ranked best for each cohort. Figure S67A shows that this region contains around 200 kb of low marker coverage upstream of the rs12913832 variant discussed in Visser et al. [153]. Low marker density as well as small sample size can impact the precision of our methods, instead emphasizing hitchhiking variants further from the selected allele. On the other hand, these intermediate frequency SNPs at base pairs chr15:28,100,878 and chr15:28,141,480 may be functionally important to an unknown selected haplotype. We apply `iSAFE` to study this region as well, but it removes thousands of SNPs where there is low marker density, including most of the variants near the OCA2 gene. Since there is not a consensus best-ranked variant, we use the haplotype-based frequency

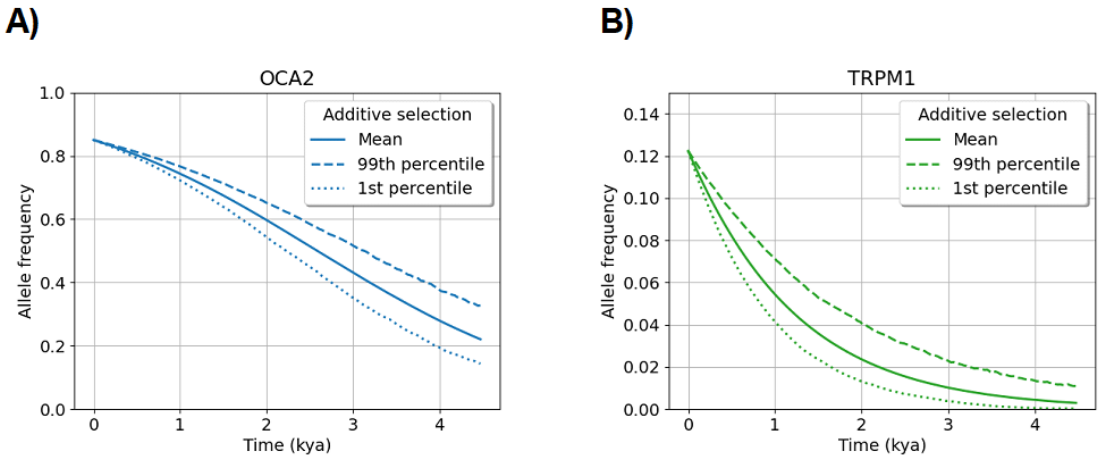

Figure 7.5: Estimated historical allele frequencies in last 150 generations for OCA2 and TRPM1 loci putatively under selection in Europeans. The plots show the means, the first percentiles, and the ninety-ninth percentiles for allele frequencies $p(t)$ from one thousand parametric bootstraps over identity-by-descent and Wright-Fisher processes for selective sweeps at (A) OCA2 and (B) TRPM1 loci. The TRPM1 analysis is based on the best-ranked SNP because there is consensus across cohort analyses. The OCA2 analysis is based on haplotype-based frequency estimates because there is no consensus on a variant. The TRPM1 analysis is based on IBD segments in the first European ancestry group (EUR1); the OCA2 analysis is based on IBD segments in the WHI cohort, the largest cohort where this locus passes the selection scan. Additive selection is assumed. The y-axis scale for the (B) TRPM1 analysis is changed for improved visibility. Time is measured in thousands of years ago (kya) using one generation equals thirty years.

estimate $\hat{p}(0) = 0.85$ from the WHI cohort. (The frequency of the rs12913832 variant is 78% in the WHI sample.) We estimate the additive selection coefficient $\hat{s} = 0.0206$ (95% CI = (0.0176,0.0236)). The implied historical allele frequencies are in Figure 7.5A.

7.5.4 Limited evidence for recent hard sweeps at many European loci

The MHC region contains genes that encode molecules that bind to antigens and mediate the presentation of these to the surface of immune cells. These molecules help cells to recognize themselves from foreign agents, thereby influencing donor compatibility for organ transplant. Gini impurity indices at MHC are below 0.60 for the smaller cohorts but above 0.60 for the larger WHI and the EUR1 group. Additionally, we infer locations for the sweep that differ by as much as seven Mb across studies. These Gini impurity indices and location estimates indicate that there may be multiple sweeps or multiple adaptive haplotypes in the MHC region. These scenarios are a clear violation of our hard sweep from *de novo* mutation model.

The SLC45A2 gene is associated with pigmentation in skin, hair, and eye [49]. We do not estimate selection coefficients for this locus because its Gini impurity indices are greater than 0.6 for all cohorts and there is no consensus best-ranked SNP. Figure S67D illustrates an unusual spectrum of common variation between forty and sixty percent around chr5:34,350,000. This region is 400 kb to the right of SLC45A2 and its putative target SNP rs16891972 [100]. Sabeti et al. [125] observe that a nonsynonymous substitution variant in the SLC45A2 gene is fixed in a European sample but absent in African and Asian samples [125]. Mathieson and Terhorst [100] and Vaughn and Nielsen [152] estimate a selection coefficient at this locus based on an allele frequency change from forty to ninety percent found in ancient DNA. Our IBD-based method is not well-suited for sweeps at or near fixation (Figure 4.1). Analyses that compare between populations or use time series data may be more suitable for modeling sweeps at or near fixation.

Estimating a selection coefficient with our method involves assuming the recent hard sweep model. We do not estimate selection coefficients for other loci in Table 7.1 because they do not have as clear a signal in our analyses. These loci have Gini impurity indices greater than 0.60, no consensus best-ranked SNP, and the best-ranked SNPs have vastly different allele frequencies and locations across studies. These loci could be examples of

older fixed sweeps, sweeps with small selection coefficients, balancing selection, or other evolutionary processes.

7.6 Discussion

We apply our methods to study positive selection in European, Indian, and African ancestry samples. Our two datasets are the TOPMed 38k sequence data [146] and the UKBB array data [24]. The automation built into the analysis pipeline makes it easy to study different sample sets under different configuration settings. For example, we use separate African ancestry sample sets in TOPMed and UKBB to see if our significant results replicate. Additionally, we use cohorts within a combined TOPMed European ancestry group to see if our significant results replicate in smaller sample sets collected from different studies. The IBD-based analyses we conduct on cohorts of less than 2,000 samples are broadly consistent with our analysis on a larger consortium dataset of 13,778 samples.

In our TOPMed EUR1 analysis, only seven loci exceed our multiple testing adjusted significance threshold. We fail to observe excess IBD rates or a lack of heterozygosity among IBD clusters at numerous loci that have been repeatedly suggested to be under selection in Europeans. Many prior studies do not provide significance thresholds [155, 51], use a significance threshold 5e-8 that may not correspond to multiple testing at the 0.05 nominal level [101, 140, 146], or assume some parametric model without theoretical support [111, 51]. Meanwhile, our approximate significance threshold adjusts for multiple testing and is supported by Theorems 3.3.1, 3.3.2, and 3.3.3. Our test may be underpowered to detect anything except strong selection within the past tens of generations. Nevertheless, we encourage modeling selective sweeps only after identifying loci that pass a multiple testing adjusted significance threshold.

For European ancestry samples in TOPMed, we estimate selection coefficients for signals at the LCT, OAS1-2-3, TRPM1, and OCA2 genes. Strictly speaking, the OCA2 and OAS1-2-3 loci do not pass our multiple testing threshold, but we do observe a low Gini impurity index, a large excess IBD outgroup, and replication across cohorts. At LCT, we estimate

ninety-five percent confidence intervals for the selection coefficient that are of width less than 0.01. In comparison, Bersaglieri et al. [13] report a confidence interval that has a width greater than 0.10, and Hejase et al. [75] use a machine learning dropout technique to measure model uncertainty of width less than 0.001, which is likely too small to cover a true selection coefficient. Additionally, we identify variants near the TRPM1 and the OAS1-2-3 loci that are strongly differentiated between an inferred excess IBD outgroup and the rest of the sample. We find limited evidence in our data to fit recent and strong hard selective sweep models at the MHC and SLC45A2 loci. Further studies could explore to what extent balancing selection and population structure confound statistical inference of selective sweeps.

Most human studies on positive selection have focused on European samples. Granka et al. [61] use cross-population extended haplotype homozygosity (XP-EHH) and integrated Haplotype Score (iHS) to study positive selection within and between African ancestry cohorts in 1000 Genomes. They fail to identify any particular loci under selection that could not be distinguished from explanations of ascertainment biases, population structure, differences in linkage disequilibrium, or poor phasing accuracy. We identify one region containing the XYLT1 gene where IBD rates are high in European, Indian, and African ancestry sample sets and especially high in African ancestry sample sets. At this locus, we compute the Gini impurity index greater than 0.6 and detect an excess IBD outgroup of less than ten percent of samples in our TOPMed data. This pattern is similar to our observations at MHC, where there are high IBD rates in all sample sets but especially so in European ancestry sample sets. Garud et al. [58] states that most selection scans are specifically designed for hard sweeps, whereas our selection scan detects regions where excess IBD rates are inconsistent with neutral evolution. Future work could investigate haplotypes in this region for evidence of balancing selection or soft sweeps, especially given that selection scans can indicate genes of biomedical importance.

Using a 24-core Intel Xeon Silver 4214 2.2 GHz compute node, wall clock computing time to detect IBD segments $\geq$1.0 cM on chromosome 2 for 13,778 samples is 4.5 hours

for `hap-ibd` and 12.5 hours for `ibd-ends`. Compute times are much faster with array data. These algorithms scale to the analysis of large sequence datasets which far exceed the sample size limitations of existing approaches. The tool `ibd-ends` may be especially important to adjust for genotyping errors, gene conversion, and regions of low marker density in sequence data [20]. High-quality recombination maps are recommended to address IBD segment detection at telomeres and centromeres. We recommend pedigree-based genetic maps because low estimated recombination rates in LD-based maps tend to confound signals of positive selection [108]. Beyond human genetics, some examples of pedigree or lab crossover-based recombination maps include *Canis familiaris* [27], *Drosophila melanogaster* [36], *Arabidopsis thaliana* [126], and *Caenorhabditis elegans* [123]. Our methods are not designed specifically for human data and could be transferable to selection studies in other species.

Chapter 8

CONCLUSION

In this dissertation, we propose and assess statistical methods to model recent and strong selective sweeps. Our methods and theory are based on how recent and strong positive selection impacts identity-by-descent within a population. Specifically, many clusters of haplotypes may maintain long identity-by-descent segments descendant from an especially advantageous allele sweeping towards fixation. Based on this intuition, we use large identity-by-descent clusters and excess identity-by-descent rates to detect and model selective sweeps.

We start by motivating our methods for selection inference with theoretical contributions. In Chapter 2, we introduce the mathematical model in which long identity-by-descent segments derive from common ancestors within the past few hundred generations. In Chapter 3, we present and prove central limit theorems for the detectable identity-by-descent rate around a locus. In Chapter 4, we give an estimator for the selection coefficient that is an easy-to-interpret one-to-one function of the detectable identity-by-descent rate around a locus. We leverage our central limit theorems and an efficient algorithm in Chapter 2 to offer confidence intervals that are valid under Delta method conditions [31].

The mathematical contributions in the aforementioned chapters apply to long identity-by-descent segments accurately detected from genetic markers in real data [16, 66, 134, 54, 20, 163, 120]. In Chapter 5, we provide heuristic approaches to pinpoint the location and frequency of a sweeping allele, which are necessary inputs to our selection coefficient estimator. In Chapter 6, we adjust for multiple testing in a principled statistical framework, and we indicate that our hypothesis test may be anti-conservative in most real datasets. Finally, in Chapter 7, we apply the statistical methods in Chapters 2, 4, 5, and 6 to analyze positive selection in African, European, and South Asian ancestry samples from the Trans-

Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project [146] and the United Kingdom Biobank (UKBB) [24].

Our work has highlighted several shortcomings in existing methods to study positive selection. We refer to the 5e-8 significance threshold as a nominal significance threshold in existing methods to test for selection [51, 140, 146], not a multiple testing adjustment, because they have not demonstrated control of the FWER in theory nor in simulation studies. Moreover, existing hypothesis tests for selection are not the same as genome-wide association studies in array data, which is the historical precedence of the 5e-8 significance threshold [35]. As case studies, our multiple testing adjustments and those in Grinde et al. [63] indicate that the effective number of tests in population genetics may differ significantly between human populations. On a related note, we question whether existing selection coefficient estimators for $s \leq 0.01$ [142, 152, 103, 75, 151] should be trusted without uncertainty quantification. In our simulation studies, we find that a coalescent likelihood-based approach and a deep learning approach have considerable biases in estimating $s > 0.01$, which is a scenario that should have more not less statistical signal than $s \leq 0.01$. It is difficult to investigate why these complicated estimators perform so poorly in our simulation studies, whereas our estimator is an easy-to-interpret one-to-one function of a sample mean.

We make a few key observations about our methods and their robust performance in extensive simulation studies. First, the detectable identity-by-descent rate around a locus is a sample mean of correlated binary random variables, and the regularity conditions in its central limit theorems are such that its variance is comparable to the variance of the binomial sample mean. Second, the binomial sample mean is a maximum likelihood estimator (MLE) for its mean parameter, and one-to-one functions of MLEs are also MLEs (Theorem 7.2.10 in Casella and Berger [31]). The composite likelihood of the detectable identity-by-descent rate is the binomial likelihood, and this relationship, combined with our central limit theorems, may explain the accuracy of our selection coefficient estimator and confidence intervals[1].

[1]See Casella and Berger [31] for the statistical properties of maximum likelihood estimators and exponential family distributions.

Third, the detectable identity by descent rate around a locus, not the entire identity-by-descent segment length distribution, behaves empirically like a sufficient statistic.

The hard selective sweep model is likely wrong in practice but useful in its simplicity. Our heuristic methods for marker data (Chapter 5) provide ancillary support for modeling selective sweeps once a null model of constant identity-by-descent rate is rejected. A single majority haplotype cluster of excess identity-by-descent rate and a reduction in the diversity of common variants are evidence consistent with a hard selective sweep. Our automated and efficient workflow enables reproducible analyses in cohorts of consortium datasets or in different populations. Findings that replicate within and across datasets are important alongside statistically valid hypothesis tests and estimation. For example, we observe similar results for selection at the LCT locus between our analyses of individual cohorts and a combined sample of inferred European ancestry individuals from the TOPMed data.

We fail to reject a null model of constant identity-by-descent rate at many human loci that have been reported as putative sites of adaptive evolution. Many of these putatively selected loci have high identity-by-descent rates relative to the genome-wide median, but they do not exceed our anti-conservative multiple testing adjusted threshold. Another explanation is that selection at these loci is weaker than $s = 0.01$ or older than a couple hundred generations. We also reject the null model of constant identity-by-descent rate in African and South Asian ancestry samples near a gene important to skeletal cell development. We suggest that this putatively selected locus is likely not due to a hard selective sweep but that the locus could be subject to balancing selection. Future studies on this locus may contribute to medical genetics in non-European populations.

There are many opportunities to extend the methods presented in this thesis. One current limitation is identity-by-descent segment detection near telomeres and centromeres. This limitation may be addressed technologically by increasing the availability of long-read sequencing and a telomere-to-telomere reference genome [106]. This limitation may also be addressed methodologically by adapting our selection scan to identity-by-descent segment detection without a genetic map. The latter opportunity is especially important for non-

human studies. Our analysis of samples from a small founder population suggests that our selection scan could be underpowered when the identity-by-descent rate process has high variance relative to that of a modestly sized population. Recurrent sweeps and strong selective sweeps across the genome may also make it difficult to perform anomaly detection in our hypothesis testing framework. Human interventions in the environment could accelerate genome-wide selective sweeps in *Plasmodium falciparum*, *Drosophila* species, *Anopheles gambiae*, and SARS-CoV-2 populations, to name a few examples [57, 65, 6, 7]. Lastly, Skov et al. [138] suggest that there may have been many strong selective sweeps on the human X chromosome, but our methods are currently limited to autosomes. Combining information from autosomes may be important for selection scans on sex chromosomes, where there is less sequence data to measure genome-wide tendencies. Lessons from demographic inference and local ancestry inference on the X chromosome [22, 62, 26] may provide some motivation for methods development.

The overarching aim of this dissertation is to offer statistical inference of recent positive selection that has strong theoretical support and robust performance in real data. We hope that its statistical contribution to population genetics helps foster valid and reproducible results in ongoing genetics research on non-European and non-human populations.

BIBLIOGRAPHY

[1] Jeffrey R Adrion, Christopher B Cole, Noah Dukler, Jared G Galloway, Ariella L Gladstein, Graham Gower, Christopher C Kyriazis, Aaron P Ragsdale, Georgia Tsambos, Franz Baumdicker, Jedidiah Carlson, Reed A Cartwright, Arun Durvasula, Ilan Gronau, Bernard Y Kim, Patrick McKenzie, Philipp W Messer, Ekaterina Noskova, Diego Ortega-Del Vecchyo, Fernando Racimo, Travis J Struck, Simon Gravel, Ryan N Gutenkunst, Kirk E Lohmueller, Peter L Ralph, Daniel R Schrider, Adam Siepel, Jerome Kelleher, and Andrew D Kern. A community-maintained standard library of population genetic models. *Elife*, 9, June 2020.

[2] Ali Akbari, Joseph J Vitti, Arya Iranmehr, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. Identifying the favored mutation in a positive selective sweep. *Nat. Methods*, 15(4):279–282, April 2018.

[3] Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, September 2010.

[4] David Aldous. *Probability Approximations via the Poisson Clumping Heuristic*. Springer Science & Business Media, March 2013.

[5] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664, September 2009.

[6] Anopheles gambiae 1000 Genomes Consortium. Genetic diversity of the African malaria vector Anopheles gambiae. *Nature*, 552(7683):96–100, December 2017.

[7] Anopheles gambiae 1000 Genomes Consortium. Genome variation and population

structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii. *Genome Res.*, 30(10):1533–1546, October 2020.

[8] Vivek K Bajpai, Tomek Swigut, Jaaved Mohammed, Sahin Naqvi, Martin Arreola, Josh Tycko, Tayne C Kim, Jonathan K Pritchard, Michael C Bassik, and Joanna Wysocka. A genome-wide genetic screen uncovers determinants of human pigmentation. *Science*, 381(6658):eade6289, August 2023.

[9] Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, Ariella L Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S Pope, Consuelo D Quinto-Cortés, Murillo F Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W Wohns, Yan Wong, Simon Gravel, Andrew D Kern, Jere Koskela, Peter L Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3), March 2022.

[10] John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan Yang Ch'Ang, Wei Huang, Bin Liu, Yan Shen, Paul Kwong Hang Tam, Lap Chee Tsui, Mary Miu Yee Waye, Jeffrey Tze Fei Wong, Changqing Zeng, Qingrun Zhang, Mark S Chee, Luana M Galver, Semyon Kruglyak, Sarah S Murray, Arnold R Oliphant, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Michael S Phillips, Andrei Verner, Shenghui Duan, Denise L Lind, Raymond D Miller, John Rice, Nancy L Saccone, Patricia Taillon-Miller, Ming Xiao, Akihiro Sekine, Koki Sorimachi, Yoichi Tanaka, Tatsuhiko Tsunoda, Eiji Yoshino, David R Bentley, Sarah Hunt, Don Powell, Houcan Zhang, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles Rotimi, Clement A Adebamowo, Toyin Aniagwu, Patricia A Marshall, Olayemi Matthew, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Fiona Cunningham, Ardavan Kanani, Gudmundur A Thorisson, Peter E Chen, David J Cutler, Carl S Kashuk, Peter Donnelly, Jonathan Marchini,

Gilean A T McVean, Simon R Myers, Lon R Cardon, Andrew Morris, Bruce S Weir, James C Mullikin, Michael Feolo, Mark J Daly, Renzong Qiu, Alastair Kent, Georgia M Dunston, Kazuto Kato, Norio Niikawa, Jessica Watkin, Richard A Gibbs, Erica Sodergren, George M Weinstock, Richard K Wilson, Lucinda L Fulton, Jane Rogers, Bruce W Birren, Hua Han, Hongguang Wang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Kazuo Todani, Takashi Fujita, Satoshi Tanaka, Arthur L Holden, Francis S Collins, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Elke Jordan, Jane L Peterson, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Karen Kennedy, Michael G Dunn, Richard Seabrook, Mark Shillito, Barbara Skene, John G Stewart, David L Valle, Ellen Wright Clayton, Lynn B Jorde, Aravinda Chakravarti, Mildred K Cho, Troy Duster, Morris W Foster, Maria Jasperse, Bartha M Knoppers, Pui Yan Kwok, Julio Licinio, Jeffrey C Long, Pilar Ossorio, Vivian Ota Wang, Charles N Rotimi, Patricia Spallone, Sharon F Terry, Eric S Lander, Eric H Lai, Deborah A Nickerson, Gonçalo R Abecasis, David Altshuler, Michael Boehnke, Panos Deloukas, Julie A Douglas, Stacey B Gabriel, Richard R Hudson, Thomas J Hudson, Leonid Kruglyak, Yusuke Nakamura, Robert L Nussbaum, Stephen F Schaffner, Stephen T Sherry, Lincoln D Stein, and Toshihiro Tanaka. The international HapMap project. *Nature*, 426(6968):789–796, December 2003.

[11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.

[12] Anders Bergström, Shane A McCarthy, Ruoyun Hui, Mohamed A Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Hélène Blanché, Jean-François Deleuze, Howard Cann, Swapan Mallick, David Reich, Manjinder S Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue, Richard Durbin, and Chris Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484), March 2020.

[13] Todd Bersaglieri, Pardis C Sabeti, Nick Patterson, Trisha Vanderploeg, Steve F Schaffner, Jared A Drake, Matthew Rhodes, David E Reich, and Joel N Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.*, 74(6):1111–1120, June 2004.

[14] Anand Bhaskar, Andrew G Clark, and Yun S Song. Distortion of genealogical properties when the sample is very large. *Proc. Natl. Acad. Sci. U. S. A.*, 111(6):2385–2390, February 2014.

[15] Claude Bhérer, Christopher L Campbell, and Adam Auton. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.*, 8:14994, April 2017.

[16] Brian L Browning and Sharon R Browning. A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.*, 88(2):173–182, February 2011.

[17] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, 103(3):338–348, September 2018.

[18] Brian L Browning, Xiaowen Tian, Ying Zhou, and Sharon R Browning. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.*, 108(10):1880–1890, October 2021.

[19] Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.*, 97(3):404–418, September 2015.

[20] Sharon R Browning and Brian L Browning. Probabilistic estimation of identity by descent segment endpoints and detection of recent selection. *Am. J. Hum. Genet.*, 107 (5):895–910, November 2020.

[21] Sharon R Browning and Elizabeth A Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, April 2012.

[22] Sharon R Browning, Kelsey Grinde, Anna Plantinga, Stephanie M Gogarten, Adrienne M Stilp, Robert C Kaplan, M Larissa Avilés-Santa, Brian L Browning, and Cathy C Laurie. Local ancestry inference in a large US-Based Hispanic/Latino study: Hispanic community health study/study of Latinos (HCHS/SOL). *G3*, 6(6):1525–1534, June 2016.

[23] Sharon R Browning, Ryan K Waples, and Brian L Browning. Fast, accurate local ancestry inference with FLARE. *Am. J. Hum. Genet.*, 110(2):326–335, February 2023.

[24] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[25] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, Susan Fairley, Alexi Runnels, Lara Winterkorn, Ernesto Lowy, Human Genome Structural Variation Consortium, Paul Flicek, Soren Germer, Harrison Brand, Ira M Hall, Michael E Talkowski, Giuseppe Narzisi, and Michael C Zody. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, September 2022.

[26] Ruoyi Cai, Brian L Browning, and Sharon R Browning. Identity-by-descent-based estimation of the X chromosome effective population size with application to sex-specific demographic history. *G3*, 13(10), September 2023.

[27] Christopher L Campbell, Claude Bhérer, Bernice E Morrow, Adam R Boyko, and Adam Auton. A pedigree-based map of recombination in the domestic dog genome. *G3 (Bethesda)*, 6(11):3517–3524, November 2016.

[28] Howard M Cann, Claudia de Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, Zhu Chen, J Chu, Carlo Carcassi, Licinio Contu, Ruofu Du, Laurent Excoffier, G B Ferrara, Jonathan S Friedlaender, Helena Groot, David Gurwitz, Trefor Jenkins, Rene J Herrera, Xiaoyi Huang, Judith Kidd, Kenneth K Kidd, Andre Langaney, Alice A Lin, S Qasim Mehdi, Peter Parham, Alberto Piazza, Maria Pia Pistillo, Yaping Qian, Qunfang Shu, Jiujin Xu, S Zhu, James L Weber, Henry T Greely, Marcus W Feldman, Gilles Thomas, Jean Dausset, and L Luca Cavalli-Sforza. A human genome diversity cell line panel. *Science*, 296(5566):261–262, April 2002.

[29] Shai Carmi, Pier Francesco Palamara, Vladimir Vacic, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. The variance of identity-by-descent sharing in the Wright-Fisher model. *Genetics*, 193(3):911–928, March 2013.

[30] Shai Carmi, Ken Y Hui, Ethan Kochav, Xinmin Liu, James Xue, Fillan Grady, Saurav Guha, Kinnari Upadhyay, Dan Ben-Avraham, Semanti Mukherjee, B Monica Bowen, Tinu Thomas, Joseph Vijai, Marc Cruts, Guy Froyen, Diether Lambrechts, Stéphane Plaisance, Christine Van Broeckhoven, Philip Van Damme, Herwig Van Marck, Nir Barzilai, Ariel Darvasi, Kenneth Offit, Susan Bressman, Laurie J Ozelius, Inga Peter, Judy H Cho, Harry Ostrer, Gil Atzmon, Lorraine N Clark, Todd Lencz, and Itsik Pe'er. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.*, 5:4835, September 2014.

[31] George Casella and Roger L Berger. Statistical inference. *Biometrics*, 49(1):320, March 1993.

[32] Arun G Chandrasekhar and Matthew O Jackson. A network formation model based on subgraphs. *arXiv*, November 2016.

[33] Arun G Chandrasekhar, Matthew O Jackson, Tyler H McCormick, and Vydhourie Thiyageswaran. General covariance-based conditions for central limit theorems with dependent triangular arrays. *arXiv*, August 2023.

[34] Han Chen, Ardalan Naseri, and Degui Zhi. FiMAP: a fast identity-by-descent mapping test for biobank-scale cohorts. *PLoS Genet.*, 19(12):e1011057, December 2023.

[35] Zhongsheng Chen, Michael Boehnke, Xiaoquan Wen, and Bhramar Mukherjee. Revisiting the genome-wide significance threshold for common variant GWAS. *G3*, 11(2), February 2021.

[36] Josep M Comeron, Ramesh Ratnappan, and Samuel Bailin. The many landscapes of recombination in Drosophila melanogaster. *PLoS Genet.*, 8(10):e1002905, October 2012.

[37] Karen N Conneely and Michael Boehnke. "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests". *Am. J. Hum. Genet.*, 81 (6):1158–1168, December 2007.

[38] James F Crow and Motoo Kimura. *An Introduction to Population Genetics Theory.* Harper & Row, New York, NY, 1970.

[39] Iulia Dahmer and Götz Kersting. The internal branch lengths of the Kingman coalescent. *aoap*, 25(3):1325–1348, June 2015.

[40] Charles Darwin. On the origin of species: facsimile of the first edition. 1859.

[41] Qiliang Ding, Ya Hu, Shuhua Xu, Jiucun Wang, and Li Jin. Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in East Asians. *Mol. Biol. Evol.*, 31(3):683–695, March 2014.

[42] Hans Eiberg, Jesper Troelsen, Mette Nielsen, Annemette Mikkelsen, Jonas Mengel-From, Klaus W Kjaer, and Lars Hansen. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.*, 123(2):177–187, March 2008.

[43] Paul Erdős and Alfréd Rényi. On random graphs I. *Publ. Math. Debrecen*, 6(290-297): 18, 1959.

[44] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

[45] Gregory Ewing and Joachim Hermisson. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, August 2010.

[46] Justin C Fay and Chung I Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413, July 2000.

[47] Eleanor Feingold, Patrick O Brown, and David Siegmund. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.*, 53(1):234–251, July 1993.

[48] Joseph Felsenstein. *Theoretical Population Genetics*. University of Washington, 2005.

[49] L P Fernandez, Roger L Milne, G Pita, J A Avilés, P Lázaro, J Benítez, and Gloria Ribas. SLC45A2: a novel malignant melanoma-associated gene. *Hum. Mutat.*, 29(9): 1161–1167, September 2008.

[50] Anna Ferrer-Admetlla, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.*, 31(5):1275–1291, May 2014.

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

[52] Ronald Fisher. XXI.—on the dominance ratio. *Proceedings of the royal society of Edinburgh*, 42:321–341, 1923.

[53] Romain Fournier, Zoi Tsangalidou, David Reich, and Pier Francesco Palamara. Haplotype-based inference of recent effective population size in modern and ancient DNA samples. *Nat. Commun.*, 14(1):7945, December 2023.

[54] William A Freyman, Kimberly F Mcmanus, Suyash S Shringarpure, Ethan M Jewett, Katarzyna Bryc, and Adam Auton. Fast and robust identity-by-descent inference with the templated positional Burrows–Wheeler transform. *Mol. Biol. Evol.*, 38(5):2131–2151, May 2021.

[55] Irene Gallego Romero, Chandana Basu Mallick, Anke Liebert, Federica Crivellaro, Gyaneshwer Chaubey, Yuval Itan, Mait Metspalu, Muthukrishnan Eaaswarkhanth, Ramasamy Pitchappan, Richard Villems, David Reich, Lalji Singh, Kumarasamy Thangaraj, Mark G Thomas, Dallas M Swallow, Marta Mirazón Lahr, and Toomas Kivisild. Herders of Indian and European cattle share their predominant allele for lactase persistence. *Mol. Biol. Evol.*, 29(1):249–260, January 2012.

[56] Daniel Garrigan and Philip W Hedrick. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution*, 57(8):1707–1722, August 2003.

[57] Nandita R Garud. Understanding soft sweeps: a signature of rapid adaptation. *Nat. Rev. Genet.*, page 1, February 2023.

[58] Nandita R Garud, Philipp W Messer, Erkan O Buzbas, and Dmitri A Petrov. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet.*, 11(2):e1005004, February 2015.

[59] Rachel M Gittelman, Joshua G Schraiber, Benjamin Vernot, Carmen Mikacenic, Mark M Wurfel, and Joshua M Akey. Archaic hominin admixture facilitated adaptation to Out-of-Africa environments. *Curr. Biol.*, 26(24):3375–3382, December 2016.

[60] Stephanie M Gogarten, Tamar Sofer, Han Chen, Chaoyu Yu, Jennifer A Brody, Timothy A Thornton, Kenneth M Rice, and Matthew P Conomos. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, 35(24):5346–5348, July 2019.

[61] Julie M Granka, Brenna M Henn, Christopher R Gignoux, Jeffrey M Kidd, Carlos D Bustamante, and Marcus W Feldman. Limited evidence for classic selective sweeps in African populations. *Genetics*, 192(3):1049–1064, November 2012.

[62] Kelsey Grinde. Statistical inference in admixed populations. 2019.

[63] Kelsey E Grinde, Lisa A Brown, Alexander P Reiner, Timothy A Thornton, and Sharon R Browning. Genome-wide significance thresholds for admixture mapping studies. *Am. J. Hum. Genet.*, 104(3):454–465, March 2019.

[64] Sharon R Grossman, Ilya Shlyakhter, Elinor K Karlsson, Elizabeth H Byrne, Shannon Morales, Gabriel Frieden, Elizabeth Hostetter, Elaine Angelino, Manuel Garber, Or Zuk, Eric S Lander, Stephen F Schaffner, and Pardis C Sabeti. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327(5967):883–886, February 2010.

[65] Bing Guo, Victor Borda, Roland Laboulaye, Michele D Spring, Mariusz Wojnarski, Brian A Vesely, Joana C Silva, Norman C Waters, Timothy D O'Connor, and Shannon Takala-Harrison. Strong positive selection biases identity-by-descent-based inferences of recent demography and population structure in Plasmodium falciparum. *Nat. Commun.*, 15(1):2499, March 2024.

[66] Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, 19(2):318–326, February 2009.

[67] John B S Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.*, 8(29):299–309, 1919.

[68] John B S Haldane. A mathematical theory of natural and artificial selection. Part I. *Math. Proc. Camb. Philos. Soc.*, 23:19–41, 1924.

[69] John Burdon Haldane. *The causes of evolution*. Harper & Row, New York, NY, 1932.

[70] Bjarni V Halldorsson, Gunnar Palsson, Olafur A Stefansson, Hakon Jonsson, Marteinn T Hardarson, Hannes P Eggertsson, Bjarni Gunnarsson, Asmundur Oddsson, Gisli H Halldorsson, Florian Zink, Sigurjon A Gudjonsson, Michael L Frigge, Gudmar Thorleifsson, Asgeir Sigurdsson, Simon N Stacey, Patrick Sulem, Gisli Masson, Agnar Helgason, Daniel F Gudbjartsson, Unnur Thorsteinsdottir, and Kari Stefansson. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363(6425), January 2019.

[71] Benjamin C Haller and Philipp W Messer. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Mol. Biol. Evol.*, 36(3):632–637, March 2019.

[72] Benjamin C Haller, Jared Galloway, Jerome Kelleher, Philipp W Messer, and Peter L Ralph. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Mol. Ecol. Resour.*, 19(2):552–566, March 2019.

[73] Kelley Harris. Using enormous genealogies to map causal variants in space and time. *Nat. Genet.*, 55(5):730–731, May 2023.

[74] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: a primer in coalescent theory*. Oxford University Press, USA, December 2004.

[75] Hussein A Hejase, Ziyi Mo, Leonardo Campagna, and Adam Siepel. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Mol. Biol. Evol.*, 39(1), January 2022.

[76] Joachim Hermisson and Pleuni S Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–2352, April 2005.

[77] Joachim Hermisson and Pleuni S Pennings. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.*, 8(6):700–716, June 2017.

[78] Margarita Hernandez and George H Perry. Scanning the human genome for "signatures" of positive selection: transformative opportunities and ethical obligations. *Evol. Anthropol.*, 30(2):113–121, March 2021.

[79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scand. Stat. Theory Appl.*, 6(2):65–70, 1979.

[80] Yilei Huang, Shai Carmi, and Harald Ringbauer. Estimating effective population size trajectories from time-series identity-by-descent (IBD) segments. *bioRxiv*, page 2024.05.06.592728, May 2024.

[81] Richard R Hudson and Norman L Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1): 147–164, September 1985.

[82] International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Wei-

hua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.

[83] Jerome Kelleher, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nat. Genet.*, 51(9):1330–1338, September 2019.

[84] Andrew D Kern and Daniel R Schrider. diploS/HIC: an updated approach to classifying

selective sweeps. *G3: Genes, Genomes, Genetics*, 2018.

[85] Kenneth K Kidd, Andrew J Pakstis, Michael P Donnelly, Ozlem Bulbul, Lotfi Cherni, Cemal Gurkan, Longli Kang, Hui Li, Libing Yun, Peristera Paschou, Kelly A Meiklejohn, Eva Haigh, and William C Speed. The distinctive geographic patterns of common pigmentation variants at the OCA2 gene. *Sci. Rep.*, 10(1):15433, September 2020.

[86] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[87] John F C Kingman. On the genealogy of large populations. *J. Appl. Probab.*, 19(A): 27–43, 1982.

[88] John F C Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, September 1982.

[89] Johannes Kö Ster and Sven Rahmann. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28:2520–2522, 2012.

[90] Martin Kreitman and Hiroshi Akashi. Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.*, 26:403–422, 1995.

[91] Martin Kuhlwilm, Sojung Han, Vitor C Sousa, Laurent Excoffier, and Tomas Marques-Bonet. Ancient admixture from an extinct ape lineage into bonobos. *Nat Ecol Evol*, 3 (6):957–965, June 2019.

[92] Mary K Kuhner. LAMARC 2.0: maximum likelihood and bayesian estimation of population parameters. *Bioinformatics*, 22(6):768–770, March 2006.

[93] Mary K Kuhner, Jon Yamato, and Joseph Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156(3):1393–1401, November 2000.

[94] Fabrice Larribe and Paul Fearnhead. On composite likelihoods in statistical genetics. *Stat. Sin.*, 21(1):43–69, 2011.

[95] Elise M Lauterbur, Maria Izabel A Cavassim, Ariella L Gladstein, Graham Gower, Nathaniel S Pope, Georgia Tsambos, Jeffrey Adrion, Saurabh Belsare, Arjun Biddanda, Victoria Caudill, Jean Cury, Ignacio Echevarria, Benjamin C Haller, Ahmed R Hasan, Xin Huang, Leonardo Nicola Martin Iasi, Ekaterina Noskova, Jana Obsteter, Vitor Antonio Correa Pavinato, Alice Pearson, David Peede, Manolo F Perez, Murillo F Rodrigues, Chris C R Smith, Jeffrey P Spence, Anastasia Teterina, Silas Tittes, Per Unneberg, Juan Manuel Vazquez, Ryan K Waples, Anthony Wilder Wohns, Yan Wong, Franz Baumdicker, Reed A Cartwright, Gregor Gorjanc, Ryan N Gutenkunst, Jerome Kelleher, Andrew D Kern, Aaron P Ragsdale, Peter L Ralph, Daniel R Schrider, and Ilan Gronau. Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *Elife*, 12, June 2023.

[96] Kerstin Lindblad-Toh, Claire M Wade, Tarjei S Mikkelsen, Elinor K Karlsson, David B Jaffe, Michael Kamal, Michele Clamp, Jean L Chang, Edward J Kulbokas, 3rd, Michael C Zody, Evan Mauceli, Xiaohui Xie, Matthew Breen, Robert K Wayne, Elaine A Ostrander, Chris P Ponting, Francis Galibert, Douglas R Smith, Pieter J DeJong, Ewen Kirkness, Pablo Alvarez, Tara Biagi, William Brockman, Jonathan Butler, Chee-Wye Chin, April Cook, James Cuff, Mark J Daly, David DeCaprio, Sante Gnerre, Manfred Grabherr, Manolis Kellis, Michael Kleber, Carolyne Bardeleben, Leo Goodstadt, Andreas Heger, Christophe Hitte, Lisa Kim, Klaus-Peter Koepfli, Heidi G Parker, John P Pollinger, Stephen M J Searle, Nathan B Sutter, Rachael Thomas, Caleb Webber, Jennifer Baldwin, Adal Abebe, Amr Abouelleil, Lynne Aftuck, Mostafa Ait-Zahra, Tyler Aldredge, Nicole Allen, Peter An, Scott Anderson, Claudel Antoine, Harindra Arachchi, Ali Aslam, Laura Ayotte, Pasang Bachantsang, Andrew Barry, Tashi Bayul, Mostafa Benamara, Aaron Berlin, Daniel Bessette, Berta Blitshteyn, Toby Bloom, Jason Blye, Leonid Boguslavskiy, Claude Bonnet, Boris Boukhgalter,

Adam Brown, Patrick Cahill, Nadia Calixte, Jody Camarata, Yama Cheshatsang, Jeffrey Chu, Mieke Citroen, Alville Collymore, Patrick Cooke, Tenzin Dawoe, Riza Daza, Karin Decktor, Stuart DeGray, Norbu Dhargay, Kimberly Dooley, Kathleen Dooley, Passang Dorje, Kunsang Dorjee, Lester Dorris, Noah Duffey, Alan Dupes, Osebhajajeme Egbiremolen, Richard Elong, Jill Falk, Abderrahim Farina, Susan Faro, Diallo Ferguson, Patricia Ferreira, Sheila Fisher, Mike FitzGerald, Karen Foley, Chelsea Foley, Alicia Franke, Dennis Friedrich, Diane Gage, Manuel Garber, Gary Gearin, Georgia Giannoukos, Tina Goode, Audra Goyette, Joseph Graham, Edward Grandbois, Kunsang Gyaltsen, Nabil Hafez, Daniel Hagopian, Birhane Hagos, Jennifer Hall, Claire Healy, Ryan Hegarty, Tracey Honan, Andrea Horn, Nathan Houde, Leanne Hughes, Leigh Hunnicutt, M Husby, Benjamin Jester, Charlien Jones, Asha Kamat, Ben Kanga, Cristyn Kells, Dmitry Khazanovich, Alix Chinh Kieu, Peter Kisner, Mayank Kumar, Krista Lance, Thomas Landers, Marcia Lara, William Lee, Jean-Pierre Leger, Niall Lennon, Lisa Leuper, Sarah LeVine, Jinlei Liu, Xiaohong Liu, Yeshi Lokyitsang, Tashi Lokyitsang, Annie Lui, Jan Macdonald, John Major, Richard Marabella, Kebede Maru, Charles Matthews, Susan McDonough, Teena Mehta, James Meldrim, Alexandre Melnikov, Louis Meneus, Atanas Mihalev, Tanya Mihova, Karen Miller, Rachel Mittelman, Valentine Mlenga, Leonidas Mulrain, Glen Munson, Adam Navidi, Jerome Naylor, Tuyen Nguyen, Nga Nguyen, Cindy Nguyen, Thu Nguyen, Robert Nicol, Nyima Norbu, Choe Norbu, Nathaniel Novod, Tenchoe Nyima, Peter Olandt, Barry O'Neill, Keith O'Neill, Sahal Osman, Lucien Oyono, Christopher Patti, Danielle Perrin, Pema Phunkhang, Fritz Pierre, Margaret Priest, Anthony Rachupka, Sujaa Raghuraman, Rayale Rameau, Verneda Ray, Christina Raymond, Filip Rege, Cecil Rise, Julie Rogers, Peter Rogov, Julie Sahalie, Sampath Settipalli, Theodore Sharpe, Terrance Shea, Mechele Sheehan, Ngawang Sherpa, Jianying Shi, Diana Shih, Jessie Sloan, Cherylyn Smith, Todd Sparrow, John Stalker, Nicole Stange-Thomann, Sharon Stavropoulos, Catherine Stone, Sabrina Stone, Sean Sykes, Pierre Tchuinga, Pema Tenzing, Senait Tesfaye, Dawa Thoulutsang, Yama Thoulutsang, Kerri Topham, Ira

Topping, Tsamla Tsamla, Helen Vassiliev, Vijay Venkataraman, Andy Vo, Tsering Wangchuk, Tsering Wangdi, Michael Weiand, Jane Wilkinson, Adam Wilson, Shailendra Yadav, Shuli Yang, Xiaoping Yang, Geneva Young, Qing Yu, Joanne Zainoun, Lisa Zembek, Andrew Zimmer, and Eric S Lander. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, December 2005.

[97] Devin P Locke, Ladeana W Hillier, Wesley C Warren, Kim C Worley, Lynne V Nazareth, Donna M Muzny, Shiaw-Pyng Yang, Zhengyuan Wang, Asif T Chinwalla, Pat Minx, Makedonka Mitreva, Lisa Cook, Kim D Delehaunty, Catrina Fronick, Heather Schmidt, Lucinda A Fulton, Robert S Fulton, Joanne O Nelson, Vincent Magrini, Craig Pohl, Tina A Graves, Chris Markovic, Andy Cree, Huyen H Dinh, Jennifer Hume, Christie L Kovar, Gerald R Fowler, Gerton Lunter, Stephen Meader, Andreas Heger, Chris P Ponting, Tomas Marques-Bonet, Can Alkan, Lin Chen, Ze Cheng, Jeffrey M Kidd, Evan E Eichler, Simon White, Stephen Searle, Albert J Vilella, Yuan Chen, Paul Flicek, Jian Ma, Brian Raney, Bernard Suh, Richard Burhans, Javier Herrero, David Haussler, Rui Faria, Olga Fernando, Fleur Darré, Domènec Farré, Elodie Gazave, Meritxell Oliva, Arcadi Navarro, Roberta Roberto, Oronzo Capozzi, Nicoletta Archidiacono, Giuliano Della Valle, Stefania Purgato, Mariano Rocchi, Miriam K Konkel, Jerilyn A Walker, Brygg Ullmer, Mark A Batzer, Arian F A Smit, Robert Hubley, Claudio Casola, Daniel R Schrider, Matthew W Hahn, Victor Quesada, Xose S Puente, Gonzalo R Ordoñez, Carlos López-Otín, Tomas Vinar, Brona Brejova, Aakrosh Ratan, Robert S Harris, Webb Miller, Carolin Kosiol, Heather A Lawson, Vikas Taliwal, André L Martins, Adam Siepel, Arindam Roychoudhury, Xin Ma, Jeremiah Degenhardt, Carlos D Bustamante, Ryan N Gutenkunst, Thomas Mailund, Julien Y Dutheil, Asger Hobolth, Mikkel H Schierup, Oliver A Ryder, Yuko Yoshinaga, Pieter J de Jong, George M Weinstock, Jeffrey Rogers, Elaine R Mardis, Richard A Gibbs, and Richard K Wilson. Comparative and demographic analysis of orang-utan genomes.

Nature, 469(7331):529–533, January 2011.

[98] Bryan F J Manly. *Randomization, Bootstrap and Monte Carlo Methods in Biology: Texts in Statistical Science*. chapman and hall/CRC, 2018.

[99] Iain Mathieson and Gil McVean. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, March 2013.

[100] Iain Mathieson and Jonathan Terhorst. Direct detection of natural selection in Bronze Age Britain. *Genome Res.*, 32(11-12):2057–2067, October 2022.

[101] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, Kendra Sirak, Cristina Gamba, Eppie R Jones, Bastien Llamas, Stanislav Dryomov, Joseph Pickrell, Juan Luís Arsuaga, José María Bermúdez de Castro, Eudald Carbonell, Fokke Gerritsen, Aleksandr Khokhlov, Pavel Kuznetsov, Marina Lozano, Harald Meller, Oleg Mochalov, Vyacheslav Moiseyev, Manuel A Rojo Guerra, Jacob Roodenberg, Josep Maria Vergès, Johannes Krause, Alan Cooper, Kurt W Alt, Dorcas Brown, David Anthony, Carles Lalueza-Fox, Wolfgang Haak, Ron Pinhasi, and David Reich. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528 (7583):499–503, December 2015.

[102] Emily K Mis, Karel F Liem, Jr, Yong Kong, Nancy B Schwartz, Miriam Domowicz, and Scott D Weatherbee. Forward genetics defines XYLT1 as a key, conserved regulator of early chondrocyte maturation and skeletal length. *Dev. Biol.*, 385(1):67–82, January 2014.

[103] Ziyi Mo and Adam Siepel. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLoS Genet.*, 19(11): e1011032, November 2023.

[104] Juba Nait Saada, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev, and Pier Francesco Palamara. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.*, 11(1):1–15, November 2020.

[105] Ardalan Naseri, Xiaoming Liu, Kecong Tang, Shaojie Zhang, and Degui Zhi. RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biol.*, 20(1):143, July 2019.

[106] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J Hoyt, Mark Diekhans, Glennis A Logsdon, Michael Alonge, Stylianos E Antonarakis, Matthew Borchers, Gerard G Bouffard, Shelise Y Brooks, Gina V Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G de Lima, Philip C Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T Fiddes, Giulio Formenti, Robert S Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G S Grady, Tina A Graves-Lindsay, Ira M Hall, Nancy F Hansen, Gabrielle A Hartley, Marina Haukness, Kerstin Howe, Michael W Hunkapiller, Chirag Jain, Miten Jain, Erich D Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V Maduro, Tobias Marschall, Ann M McCartney, Jennifer McDaniel, Danny E Miller, James C Mullikin, Eugene W Myers, Nathan D Olson, Benedict Paten, Paul Peluso, Pavel A Pevzner, David Porubsky, Tamara Potapova, Evgeny I Rogaev, Jeffrey A Rosenfeld, Steven L Salzberg, Valerie A Schneider, Fritz J Sedlazeck, Kishwar Shafin, Colin J Shew, Alaina Shumate, Ying Sims, Arian F A Smit, Daniela C Soto, Ivan Sović, Jessica M Storer, Aaron Streets, Beth A Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P Walenz, Aaron Wenger, Jonathan M D Wood, Chunlin Xiao, Stephanie M Yan, Alice C Young, Samantha Zarate, Urvashi Surti, Rajiv C McCoy, Megan Y Dennis, Ivan A Alexandrov, Jennifer L Gerton, Rachel J O'Neill, Winston Timp, Justin M

Zook, Michael C Schatz, Evan E Eichler, Karen H Miga, and Adam M Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, April 2022.

[107] Tomoko Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246 (5428):96–98, November 1973.

[108] Paul F O'Reilly, Ewan Birney, and David J Balding. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.*, 18(8):1304–1313, August 2008.

[109] Pier Francesco Palamara. ARGON: fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics*, 32(19):3032–3034, October 2016.

[110] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.*, 91(5):809–822, November 2012.

[111] Pier Francesco Palamara, Jonathan Terhorst, Yun S Song, and Alkes L Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.*, 50(9):1311–1317, September 2018.

[112] Vasili Pankratov, Milyausha Yunusbaeva, Sergei Ryakhovsky, Maksym Zarodniuk, Estonian Biobank Research Team, and Bayazit Yunusbayev. Prioritizing autoimmunity risk variants for functional analyses by fine-mapping mutations under natural selection. *Nat. Commun.*, 13(1):7069, November 2022.

[113] K A Papadakis, J Prehn, V Nelson, L Cheng, S W Binder, P D Ponath, D P Andrew, and S R Targan. The role of thymus-expressed chemokine and its receptor CCR9 on lymphocytes in the regional specialization of the mucosal immune system. *J. Immunol.*, 165(9):5069–5076, November 2000.

[114] Pleuni S Pennings and Joachim Hermisson. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.*, 2(12):e186, December 2006.

[115] Pleuni S Pennings and Joachim Hermisson. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.*, 23(5): 1076–1084, March 2006.

[116] Benjamin M Peter, Emilia Huerta-Sanchez, and Rasmus Nielsen. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.*, 8(10):e1003011, October 2012.

[117] Joseph K Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman, and Jonathan K Pritchard. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.*, 19(5):826–837, May 2009.

[118] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, August 2006.

[119] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, June 2000.

[120] Monica D Ramstetter, Thomas D Dyer, Donna M Lehman, Joanne E Curran, Ravindranath Duggirala, John Blangero, Jason G Mezey, and Amy L Williams. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 207(1):75–82, 2017.

[121] Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.*, 10(5):e1004342, May 2014.

[122] Rebecca Riley, Iain Mathieson, and Sara Mathieson. Interpreting generative adversarial networks to infer natural selection from genetic data. *Genetics*, 226(4), February 2024.

[123] Matthew V Rockman and Leonid Kruglyak. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet.*, 5(3):e1000419, March 2009.

[124] Pardis C Sabeti, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, Hans C Ackerman, Sarah J Campbell, David Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, October 2002.

[125] Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Steven A McCarroll, Rachelle Gaudet, Stephen F Schaffner, Eric S Lander, International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Lud-

mila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, Todd A Johnson, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu

Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913–918, October 2007.

[126] Patrice A Salomé, Kirsten Bomblies, Joffrey Fitz, Roosa A E Laitinen, Norman Warthmann, Levi Yant, and Detlef Weigel. The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity (Edinb.)*, 108(4):447–455, April 2012.

[127] Michael Salter-Townshend and Simon Myers. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, 212(3):869–889, July 2019.

[128] Daria Salyakina, Shaun R Seaman, Brian L Browning, Frank Dudbridge, and Bertram Müller-Myhsok. Evaluation of Nyholt's procedure for multiple testing correction. *Hum. Hered.*, 60(1):19–25, 2005.

[129] Aaron J Sams, Anne Dumaine, Yohann Nédélec, Vania Yotova, Carolina Alfieri, Jerome E Tanner, Philipp W Messer, and Luis B Barreiro. Adaptively introgressed neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.*, 17(1):246, November 2016.

[130] Julia Schreml, Burak Durmaz, Ozgur Cogulu, Katharina Keupp, Filippo Beleggia, Esther Pohl, Esther Milz, Mahmut Coker, Sema Kalkan Ucar, Gudrun Nürnberg, Peter Nürnberg, Joachim Kuhn, and Ferda Ozkinay. The missing "link": an autosomal recessive short stature syndrome caused by a hypofunctional XYLT1 mutation. *Hum. Genet.*, 133(1):29–39, January 2014.

[131] Shaun R Seaman and Bertram Müller-Myhsok. Rapid simulation of P values for prod-

uct methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.*, 76(3):399–408, March 2005.

[132] Laure Ségurel and Céline Bon. On the evolution of lactase persistence in humans. *Annu. Rev. Genomics Hum. Genet.*, 18:297–319, August 2017.

[133] Janie F Shelton, Anjali J Shastri, Chelsea Ye, Catherine H Weldon, Teresa Filshtein-Sonmez, Daniella Coker, Antony Symons, Jorge Esparza-Gordillo, 23andMe COVID-19 Team, Stella Aslibekyan, and Adam Auton. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat. Genet.*, 53(6):801–808, June 2021.

[134] Ruhollah Shemirani, Gillian M Belbin, Christy L Avery, Eimear E Kenny, Christopher R Gignoux, and José Luis Ambite. Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nat. Commun.*, 12(1):3546, June 2021.

[135] Ruhollah Shemirani, Gillian M Belbin, Keith Burghardt, Kristina Lerman, Christy L Avery, Eimear E Kenny, Christopher R Gignoux, and José Luis Ambite. Selecting clustering algorithms for Identity-By-Descent mapping. In *Pacific Symposium on Biocomputing 2023*, pages 121–132, November 2022.

[136] Zbynek Sidak. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.*, 62(318):626–633, 1967.

[137] David Siegmund and Benjamin Yakir. *The Statistics of Gene Mapping*. Springer New York, 2007.

[138] Laurits Skov, Moisès Coll Macià, Elise Anne Lucotte, Maria Izabel Alvez Cavassim, David Castellano, Mikkel Heide Schierup, and Kasper Munch. Extraordinary selection on the human X chromosome associated with archaic admixture. *Cell Genom*, 3(3): 100274, March 2023.

[139] John M Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genet. Res.*, 23(1):23–35, February 1974.

[140] Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.*, 51(9):1321–1329, September 2019.

[141] Jeffrey P Spence and Yun S Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv*, 5(10): eaaw9206, October 2019.

[142] 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. *PLoS Genet.*, 15(9):e1008384, September 2019.

[143] Richard A Sturm, David L Duffy, Zhen Zhen Zhao, Fabio P N Leite, Mitchell S Stark, Nicholas K Hayward, Nicholas G Martin, and Grant W Montgomery. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet.*, 82(2):424–431, February 2008.

[144] Zachary A Szpiech and Ryan D Hernandez. Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.*, 31(10): 2824–2827, October 2014.

[145] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, November 1989.

[146] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, Achilleas N Pitsillides, Jonathon LeFaive, Seung Been Lee, Xiaowen Tian, Brian L Browning, Sayantan Das, Anne Katrin Emde, Wayne E Clarke, Douglas P Loesch, Amol C Shetty, Thomas W Blackwell, Albert V Smith, Quenna Wong,

Xiaoming Liu, Matthew P Conomos, Dean M Bobo, François Aguet, Christine Albert, Alvaro Alonso, Kristin G Ardlie, Dan E Arking, Stella Aslibekyan, Paul L Auer, John Barnard, R Graham Barr, Lucas Barwick, Lewis C Becker, Rebecca L Beer, Emelia J Benjamin, Lawrence F Bielak, John Blangero, Michael Boehnke, Donald W Bowden, Jennifer A Brody, Esteban G Burchard, Brian E Cade, James F Casella, Brandon Chalazan, Daniel I Chasman, Yii Der Ida Chen, Michael H Cho, Seung Hoan Choi, Mina K Chung, Clary B Clish, Adolfo Correa, Joanne E Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L DeMeo, Susan K Dutcher, Patrick T Ellinor, Leslie S Emery, Celeste Eng, Diane Fatkin, Tasha Fingerlin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M Fullerton, Soren Germer, Mark T Gladwin, Daniel J Gottlieb, Xiuqing Guo, Michael E Hall, Jiang He, Nancy L Heard-Costa, Susan R Heckbert, Marguerite R Irvin, Jill M Johnsen, Andrew D Johnson, Robert Kaplan, Sharon L R Kardia, Tanika Kelly, Shannon Kelly, Eimear E Kenny, Douglas P Kiel, Robert Klemmer, Barbara A Konkle, Charles Kooperberg, Anna Köttgen, Leslie A Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng Han Lin, Chunyu Liu, Ruth J F Loos, Lori Garman, Robert Gerszten, Steven A Lubitz, Kathryn L Lunetta, Angel C Y Mak, Ani Manichaikul, Alisa K Manning, Rasika A Mathias, David D McManus, Stephen T McGarvey, James B Meigs, Deborah A Meyers, Julie L Mikulla, Mollie A Minear, Braxton D Mitchell, Sanghamitra Mohanty, May E Montasser, Courtney Montgomery, Alanna C Morrison, Joanne M Murabito, Andrea Natale, Pradeep Natarajan, Sarah C Nelson, Kari E North, Jeffrey R O'Connell, Nicholette D Palmer, Nathan Pankratz, Gina M Peloso, Patricia A Peyser, Jacob Pleiness, Wendy S Post, Bruce M Psaty, D C Rao, Susan Redline, Alexander P Reiner, Dan Roden, Jerome I Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, David A Schwartz, Jeong Sun Seo, Sudha Seshadri, Vivien A Sheehan, Wayne H Sheu, M Benjamin Shoemaker, Nicholas L Smith, Jennifer A Smith, Nona Sotoodehnia, Adrienne M Stilp, Weihong Tang, Kent D Taylor, Marilyn Telen, Timothy A Thornton, Russell P Tracy, David J Van Den Berg,

Ramachandran S Vasan, Karine A Viaud-Martinez, Scott Vrieze, Daniel E Weeks, Bruce S Weir, Scott T Weiss, Lu Chen Weng, Cristen J Willer, Yingze Zhang, Xutong Zhao, Donna K Arnett, Allison E Ashley-Koch, Kathleen C Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M Rice, Stephen S Rich, Edwin K Silverman, Pankaj Qasba, Weiniu Gan, Namiko Abe, Laura Almasy, Seth Ament, Peter Anderson, Pramod Anugu, Deborah Applebaum-Bowden, Tim Assimes, Dimitrios Avramopoulos, Emily Barron-Casella, Terri Beaty, Gerald Beck, Diane Becker, Amber Beitelshees, Takis Benos, Marcos Bezerra, Joshua Bis, Russell Bowler, Ulrich Broeckel, Jai Broome, Karen Bunting, Carlos Bustamante, Erin Buth, Jonathan Cardwell, Vincent Carey, Cara Carty, Richard Casaburi, Peter Castaldi, Mark Chaffin, Christy Chang, Yi Cheng Chang, Sameer Chavan, Bo Juen Chen, Wei Min Chen, Lee Ming Chuang, Ren Hua Chung, Suzy Comhair, Elaine Cornell, Carolyn Crandall, James Crapo, Jeffrey Curtis, Coleen Damcott, Sean David, Colleen Davis, Lisa de las Fuentes, Michael DeBaun, Ranjan Deka, Scott Devine, Qing Duan, Ravi Duggirala, Jon Peter Durda, Charles Eaton, Lynette Ekunwe, Adel El Boueiz, Serpil Erzurum, Charles Farber, Matthew Flickinger, Myriam Fornage, Chris Frazar, Mao Fu, Lucinda Fulton, Shanshan Gao, Yan Gao, Margery Gass, Bruce Gelb, Xiaoqi Priscilla Geng, Mark Geraci, Auyon Ghosh, Chris Gignoux, David Glahn, Da Wei Gong, Harald Goring, Sharon Graw, Daniel Grine, C Charles Gu, Yue Guan, Namrata Gupta, Jeff Haessler, Nicola L Hawley, Ben Heavner, David Herrington, Craig Hersh, Bertha Hidalgo, James Hixson, Brian Hobbs, John Hokanson, Elliott Hong, Karin Hoth, Chao Agnes Hsiung, Yi Jen Hung, Haley Huston, Chii Min Hwu, Rebecca Jackson, Deepti Jain, Min A Jhun, Craig Johnson, Rich Johnston, Kimberly Jones, Sekar Kathiresan, Alyna Khan, Wonji Kim, Greg Kinney, Holly Kramer, Christoph Lange, Ethan Lange, Leslie Lange, Cecelia Laurie, Meryl LeBoff, Jiwon Lee, Seunggeun Shawn Lee, Wen Jane Lee, David Levine, Joshua Lewis, Xiaohui Li, Yun Li, Henry Lin, Honghuang Lin, Keng Han Lin, Simin Liu, Yongmei Liu, Yu Liu, James Luo, Michael Mahaney, Barry Make, Jo Ann Manson, Lauren Margolin, Lisa Martin, Susan Mathai, Susanne May, Patrick McArdle, Merry Lynn

McDonald, Sean McFarland, Daniel McGoldrick, Caitlin McHugh, Hao Mei, Luisa Mestroni, Nancy Min, Ryan L Minster, Matt Moll, Arden Moscati, Solomon Musani, Stanford Mwasongwe, Josyf C Mychaleckyj, Girish Nadkarni, Rakhi Naik, Take Naseri, Sergei Nekhai, Bonnie Neltner, Heather Ochs-Balcom, David Paik, James Pankow, Afshin Parsa, Juan Manuel Peralta, Marco Perez, James Perry, Ulrike Peters, Lawrence S Phillips, Toni Pollin, Julia Powers Becker, Meher Preethi Boorgula, Michael Preuss, Dandi Qiao, Zhaohui Qin, Nicholas Rafaels, Laura Raffield, Laura Rasmussen-Torvik, Aakrosh Ratan, Robert Reed, Elizabeth Regan, Muagututi'a Sefuiva Reupena, Carolina Roselli, Pamela Russell, Sarah Ruuska, Kathleen Ryan, Ester Cerdeira Sabino, Danish Saleheen, Shabnam Salimi, Steven Salzberg, Kevin Sandow, Vijay G Sankaran, Christopher Scheller, Ellen Schmidt, Karen Schwander, Frank Sciurba, Christine Seidman, Jonathan Seidman, Stephanie L Sherman, Aniket Shetty, Wayne Hui Heng Sheu, Brian Silver, Josh Smith, Tanja Smith, Sylvia Smoller, Beverly Snively, Michael Snyder, Tamar Sofer, Garrett Storm, Elizabeth Streeten, Yun Ju Sung, Jody Sylvia, Adam Szpiro, Carole Sztalryd, Hua Tang, Margaret Taub, Matthew Taylor, Simeon Taylor, Machiko Threlkeld, Lesley Tinker, David Tirschwell, Sarah Tishkoff, Hemant Tiwari, Catherine Tong, Michael Tsai, Dhananjay Vaidya, Peter VandeHaar, Tarik Walker, Robert Wallace, Avram Walts, Fei Fei Wang, Heming Wang, Karol Watson, Jennifer Wessel, Kayleen Williams, L Keoki Williams, Carla Wilson, Joseph Wu, Huichun Xu, Lisa Yanek, Ivana Yang, Rongze Yang, Norann Zaghloul, Maryam Zekavat, Snow Xueyan Zhao, Wei Zhao, Degui Zhi, Xiang Zhou, Xiaofeng Zhu, George J Papanicolaou, Deborah A Nickerson, Sharon R Browning, Michael C Zody, Sebastian Zöllner, James G Wilson, L Adrienne Cupples, Cathy C Laurie, Cashell E Jaquish, Ryan D Hernandez, Timothy D O'Connor, and Gonçalo R Abecasis. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*, 590(7845):290–299, February 2021.

[147] Seth D Temple and Elizabeth A Thompson. Identity-by-descent in large samples.

bioRxiv, page 10.1101/2024.06.05.597656, June 2024.

[148] Seth D Temple, Ryan D Waples, and Sharon R Browning. Modeling recent positive selection in Americans of European ancestry. *bioRxiv*, page 2023.11.13.566947, November 2023.

[149] Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, June 2013.

[150] Xiaowen Tian, Brian L Browning, and Sharon R Browning. Estimating the genome-wide mutation rate with three-way identity by descent. *Am. J. Hum. Genet.*, 105(5): 883–893, November 2019.

[151] Luis Torada, Lucrezia Lorenzon, Alice Beddis, Ulas Isildak, Linda Pattini, Sara Mathieson, and Matteo Fumagalli. ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, 20(Suppl 9):337, November 2019.

[152] Andrew H Vaughn and Rasmus Nielsen. Fast and accurate estimation of selection coefficients and allele histories from ancient and modern DNA. *Molecular Biology and Evolution*, 41(8):msae156, August 2024.

[153] Mijke Visser, Manfred Kayser, and Robert-Jan Palstra. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res.*, 22(3):446–455, March 2012.

[154] Joseph J Vitti, Sharon R Grossman, and Pardis C Sabeti. Detecting natural selection in genomic data. *Annu. Rev. Genet.*, 47:97–120, 2013.

[155] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLoS Biol.*, 4(3):e72, 2006.

[156] John Wakeley and Tsuyoshi Takahashi. Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.*, 20(2):208–213, February 2003.

[157] Zhanpeng Wang, Jiaping Wang, Michael Kourakos, Nhung Hoang, Hyong Hark Lee, Iain Mathieson, and Sara Mathieson. Automatic inference of demographic parameters using generative adversarial networks. *Mol. Ecol. Resour.*, 21(8):2689–2705, November 2021.

[158] Carsten Wiuf and Peter Donnelly. Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.*, 56(2):183–201, October 1999.

[159] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97–159, March 1931.

[160] Brian C Zhang, Arjun Biddanda, Árni Freyr Gunnarsson, Fergus Cooper, and Pier Francesco Palamara. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat. Genet.*, May 2023.

[161] Xiuwen Zheng, David Levine, Jess Shen, Stephanie M Gogarten, Cathy Laurie, and Bruce S Weir. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24):3326–3328, December 2012.

[162] Ying Zhou, Brian L Browning, and Sharon R Browning. Population-specific recombination maps from segments of identity by descent. *Am. J. Hum. Genet.*, 107(1): 137–148, June 2020.

[163] Ying Zhou, Sharon R Browning, and Brian L Browning. A fast and simple method for detecting identity-by-descent segments in large-scale data. *Am. J. Hum. Genet.*, 106 (4):426–437, April 2020.

[164] Ying Zhou, Sharon R Browning, and Brian L Browning. IBDkin: fast estimation of kinship coefficients from identity by descent segments. *Bioinformatics*, 36(16):4519–4520, August 2020.

Appendix A

CENTRAL LIMIT THEOREMS

A.1 Theoretical derivations

Lemma A.1. $\mathbb{E}_2[X_{a,b}] \to 0$ *uniformly as* $Nw \to \infty$.

Proof. Let $f(N) = (Nw)^{-1}$, and recall that $E_2[X_{a,b}] = (2Nw+1)^{-1}$. If $Nw > (1/\varepsilon - 1)/2$, then $|f(N) - 0| < \varepsilon$. Choose integer M such that $Mw \geq (1/\varepsilon - 1)/2$. Thus, for $\varepsilon > 0$, there exists M such that $|f(N)| = (2Nw+1)^{-1} < \varepsilon$ for all $N \geq M$. □

Lemma A.2. *Let* $X \sim$ *Bernoulli*(q) *and* $q \in (0,1)$. $\mathbb{E}[|Z|^3]/\mathbb{E}[|Z|^2]^{3/2}$ *is bounded above where* $Z = X - \mathbb{E}[X]$.

Proof.

$$\begin{aligned} \mathbb{E}[|Z|^3] &= |1-q|^3 q + |q|^3(1-q) \\ &= q(1-q)((1-q)^2 + q^2) \\ &< 1. \end{aligned} \tag{A.1}$$

$$\begin{aligned} \mathbb{E}[|Z|^2]^{3/2} &= (|1-q|^2 q + |q|^2(1-q))^{3/2} \\ &= (q(1-q)(1-q+q))^{3/2} \\ &= (q(1-q))^{3/2} \\ &> 0. \end{aligned} \tag{A.2}$$

□

Lemma A.3. $Cov_3(Z_{a,b}, Z_{a,c}) \equiv Cov_3(X_{a,b}, X_{a,c}) = O((Nw)^{-2})$.

Proof. Up to the reordering of three sample haplotypes, there is one possible bifurcating tree (Figure S12). Sample haplotypes a and b coalesce to a common ancestor and their

common ancestor coalesces to a common ancestor with sample haplotype c. We integrate over coalescent time and haplotype segment lengths to bound the covariance.

$$\begin{aligned}\mathbb{E}_3[X_{a,b}] &= 3\int \exp(-2Nt_3w)\exp(-3t_3)dt_3 \\ &= 3(2Nw+3)^{-1}.\end{aligned} \tag{A.3}$$

$$\begin{aligned}\mathbb{E}_3[X_{a,c}] &= 3\int\int \exp(-2Nt_3w)\exp(-2Nt_2w)\exp(-3t_3))\exp(-t_2)dt_3dt_2 \\ &= 3(2Nw+1)^{-1}(2Nw+3)^{-1}.\end{aligned} \tag{A.4}$$

$$\begin{aligned}\mathbb{E}_3[X_{a,b}X_{a,c}] &= 3\int\int\int \exp(-3Nt_3w)\exp(-2Nt_2w)\exp(-3t_3)\exp(-t_2)dt_3dt_2 \\ &= (2Nw+1)^{-1}(Nw+1)^{-1}.\end{aligned} \tag{A.5}$$

$$\begin{aligned}\mathrm{Cov}_3(X_{a,b},X_{a,c}) &= \mathbb{E}_3[X_{a,b}X_{a,c}] - \mathbb{E}_3[X_{a,b}]\cdot\mathbb{E}_3[X_{a,c}] \\ &= (2Nw+1)^{-1}((Nw+1)^{-1} - 9(2Nw+1)^{-1}(2Nw+3)^{-1}) \\ &\leq (2Nw+1)^{-1}(Nw)^{-1} \\ &= O((Nw)^{-2}).\end{aligned} \tag{A.6}$$

□

Lemma A.4. *$Cov_4(Z_{a,b},Z_{c,d}) \equiv Cov_4(X_{a,b},X_{c,d}) = O((Nw)^{-3})$.*

Proof. Up to the reordering of four sample haplotypes, there are two possible bifurcating trees (Figure S13). The first tree is as follows: sample haplotypes a' and b' coalesce to a common ancestor, then sample haplotypes c' and d' coalesce to a common ancestor, and finally those common ancestors coalesce. The covariance of $X_{a',b'}$ and $X_{c',d'}$ is zero because of independent meioses. We focus instead on the covariance of $X_{a',c'}$ and $X_{b',d'}$. We integrate over coalescent time and haplotype segment lengths to bound the covariance.

$$\begin{aligned}\mathbb{E}_4[X_{a',c'}] &= \mathbb{E}_4[X_{b',d'}] \\ &= 6\cdot 3\int\int \exp(-2N(t_4+t_3+t_2)w)\exp(-(6t_4+3t_3+t_2))\,dt_4dt_3dt_2 \\ &= 18(2Nw+6)^{-1}(2Nw+3)^{-1}(2Nw+1)^{-1}.\end{aligned} \tag{A.7}$$

$$\begin{aligned}\mathbb{E}_4[X_{a',c'}X_{b',d'}] &= 6\cdot 3\int\int\int \exp(-(4Nt_4+3Nt_3+2Nt_2)w)\\ &\qquad \exp(-(6t_4+3t_3+t_2))dt_4dt_3dt_2\\ &= 18(4Nw+6)^{-1}(3Nw+3)^{-1}(2Nw+1)^{-1}.\end{aligned} \tag{A.8}$$

$$\mathrm{Cov}_4(X_{a',c'},X_{b',d'}) \le 3(4Nw+6)^{-1}(Nw+1)^{-1}(2Nw+1)^{-1} = O((Nw)^{-3}). \tag{A.9}$$

The second tree is as follows: a' and b' coalesce to a common ancestor, then their common ancestor coalesces with c', and finally, the common ancestor of a', b', and c' coalesces with d'. It is easy to verify that $\mathbb{E}_4[X_{a',c'}X_{b',d'}]$ is the exact same as in Equation A.8. Next,

$$\begin{aligned}\mathbb{E}_4[X_{a',c'}] &= 6\cdot 3\int\int \exp(-2N(t_4+t_3)w)\exp(-(6t_4+3t_3+t_2))\,dt_4dt_3dt_2\\ &= 18(2Nw+6)^{-1}(2Nw+3)^{-1}.\end{aligned} \tag{A.10}$$

$$\begin{aligned}\mathbb{E}_4[X_{b',d'}] &= 6\cdot 3\int\int \exp(-2N(t_4+t_3+t_2)w)\exp(-(6t_4+3t_3+t_2))\,dt_4dt_3dt_2\\ &= 18(2Nw+6)^{-1}(2Nw+3)^{-1}(2Nw+1)^{-1}.\end{aligned} \tag{A.11}$$

Because Equations A.10 and A.11 are nonnegative, the covariance upper bound is the same as in Equation A.9. □

Lemma A.5. *The following are true*

- $Cov_2(\tilde{Z}_{a,b}) = O((Nw)^{-1})$;
- $Cov_3(\tilde{Z}_{a,b},\tilde{Z}_{a,c}) \equiv Cov_3(Y_{a,b},Y_{a,c}) = O((Nw)^{-2})$;
- $Cov_4(\tilde{Z}_{a,c},\tilde{Z}_{b,d}) \equiv Cov_4(Y_{a,c},Y_{b,d}) = O((Nw)^{-3})$.

Proof. We take the same approach as in Lemmas A.3 and A.4, except the survival function

is that of an Erlang random variable with shape parameter 2.

$$\begin{aligned}
\mathbb{E}_2[Y_{a,b}] &= \int (\exp(-2Nt_2w) + 2Nt_2w\exp(-2Nt_2w))\exp(-t_2)dt_2 \\
&= (2Nw+1)^{-1} + \int 2Nt_2w\exp(-(2Nw+1)t_2)dt_2 \\
&= (2Nw+1)^{-1} + 2Nw\int t_2\exp(-(2Nw+1)t_2)dt_2 \\
&= (2Nw+1)^{-1} + 2Nw(2Nw+1)^{-2} \\
&= (2Nw+1)^{-1}(1+2Nw(2Nw+1)^{-1}).
\end{aligned} \tag{A.12}$$

$$\begin{aligned}
\mathbb{E}_3[Y_{a,b}] &= 3\int (\exp(-2Nt_3w) + 2Nt_3w\exp(-2Nt_3w))\exp(-3t_3)dt_2 \\
&= 3((2Nw+3)^{-1} + 2Nw\int t_3\exp(-(2Nw+3)t_3)) \\
&= 3((2Nw+3)^{-1} + 2Nw(2Nw+3)^{-2}) \\
&= 3(2Nw+3)^{-1}(1+2Nw(2Nw+3)^{-1}).
\end{aligned} \tag{A.13}$$

$$\begin{aligned}
\mathbb{E}_3[Y_{a,c}] &= 3(2Nw+3)^{-1}(2Nw+1)^{-1} \\
&\quad + 6Nw\int (t_3+t_2)\exp(-(2Nw+3)t_3)\exp(-(2Nw+1)t_2)dt_3dt_2 \\
&= 3((2Nw+3)^{-1}(2Nw+1)^{-1} + 2Nw(2Nw+3)^{-2}(2Nw+3)^{-2}) \\
&= 3(2Nw+3)^{-1}(2Nw+1)^{-1}(1+2Nw(2Nw+3)^{-1}(2Nw+3)^{-1}).
\end{aligned} \tag{A.14}$$

From Equations A.12, A.13, and A.14, the pattern emerges that the effect of the convolution of crossover points is to multiply $O(1)$ terms to the expected values in Equation 3.4 and Lemmas A.3 and A.4.

Calculating $\mathbb{E}_3[Y_{a,b}Y_{a,c}]$ is more involved. Up to the reordering of three sample haplotypes, we consider sample haplotypes a and c that coalesce at the most recent common ancestor of a, b, and c. Then, $\mathbb{E}_3[Y_{a,c}] \geq \mathbb{E}_3[Y_{a,b}Y_{a,c}]$, and

$$\begin{aligned}
\mathrm{Cov}_3(Y_{a,b}, Y_{a,c}) &= \mathbb{E}_3[Y_{a,b}Y_{a,c}] - \mathbb{E}_3[Y_{a,b}]\mathbb{E}_3[Y_{a,c}] \\
&\leq \mathbb{E}_3[Y_{a,b}Y_{a,c}] \\
&\leq \mathbb{E}_3[Y_{a,c}] \\
&= O((Nw)^{-2}).
\end{aligned} \tag{A.15}$$

Using the same techniques, it is easy to calculate $\mathbb{E}_4[Y_{a,c}]$ and $\mathbb{E}_4[Y_{b,d}]$ for the two different tree shapes and derive the $O((Nw)^{-3})$ bound for $\mathrm{Cov}_4(Y_{a,c}, Y_{b,d})$.

□

Lemma A.6. *For a sample of three haplotypes a, b, and c, when $\mathbb{E}_2[X_{a,c}] < 1/2$, the conditional expectation $\mathbb{E}[Z_{a,c} \times \mathbf{Z}_{-a,c} | \mathbf{Z}_{-a,c}] \not\geq 0$ for all $\mathbf{Z}_{-a,c}$.*

Proof. Define $q =: \mathbb{E}_2[X_{a,c}]$, and fix $\mathbf{X}_{-a,c} = 1$.

$$\begin{aligned}\mathbb{E}[Z_{a,c} \times \mathbf{Z}_{-a,c} | \mathbf{Z}_{-a,c}] &= \mathbb{E}[(X_{a,c} - q) \times (X_{a,b} + X_{b,c} - 2q) | X_{a,b} + X_{b,c} = 1] \\ &= \mathbb{E}[X_{a,c} \times (1 - 2q) | X_{a,b} + X_{a,c} = 1] - q + 2q^2\end{aligned}$$

Because of IBD transitivity, $X_{a,c} = 0$ with probability 1. Then, the equation simplifies to $-q(1 - 2q) < 0$. □

A.2 Verifying an assumption empirically

The third condition in Theorems 3.3.1, 3.3.2, and 3.3.2 amounts to a notion of joint non-negative correlation among IBD segment indicators, which seems reasonable given that any pair of IBD segment indicators has nonnegative covariance. A conceptual interpretation is that conditioning on the sum of all indicators except one should provide information about the height of the unobserved coalescent tree.

Calculating $\mathbb{E}[Y_{a,b}|\mathbf{Y}_{-a,b}]$ and $\mathbb{E}[Y_{a,b}|\mathbf{Y}_{-a,b}]$ involves integrating over the coalescent tree for 2^{n-1} vectors of 0's and 1's, which is analytically intractable. We instead take a Monte Carlo approach to examine this assumption. We fix the identities of two sample haplotypes a and b. We run Algorithm 1 one hundred and twenty million times, recording the value of $y_{a,b}$ and the sum $\mathbf{y}_{-a,b}$. Then, we calculate the difference between the empirical average $\bar{y}_{a,b}$ and $\mathbb{E}[Y_{a,b}]$, stratified into bins depending on the sum $\mathbf{y}_{-a,b}$.

Figure A.1 shows the results of this simulation study. The sample sizes are limited to two and four hundred individuals to keep runtime modest. We split the $y_{a,b}|\mathbf{y}_{-a,b}$ observations into eight quantile bins because $\mathbb{E}[Y_{a,b}|\mathbf{Y}_{-a,b}]$ can be exceptionally small for some $\mathbf{Y}_{-a,b}$. For each bin, the average count $\bar{y}_{a,b}$ is less than or greater than $\mathbb{E}[Y_{a,b}]$ when the sum $\mathbf{y}_{-a,b}$ is less than or greater than $\mathbb{E}[\mathbf{Y}_{-a,b}]$, respectively. This trend is especially apparent for $\mathbf{y}_{-a,b}$ far from the expected value IBD count $(\binom{n}{2} - 1) \times \mathbb{E}[Y_{a,b}]$. These findings provide some empirical evidence that the theorem assumption may be true for moderate sample sizes.

A.3 Interpreting the Ornstein-Uhlenbeck analytical approximation

The Siegmund and Yakir [137] analytical approximation is presented to a non-technical audience without explanation. This analytical approximation is also not elaborated on in Grinde [62], Grinde et al. [63], or Feingold et al. [47]. Here we offer an interpretation that relates to an effective number of hypothesis tests.

Let $\alpha = 1 - \Phi(z)$ be the nominal significance level for a one-sided test, where $Z = z \sim N(0, 1)$ and $\Phi(\cdot)$ is the cumulative distribution function. A conservative Bonferroni approach

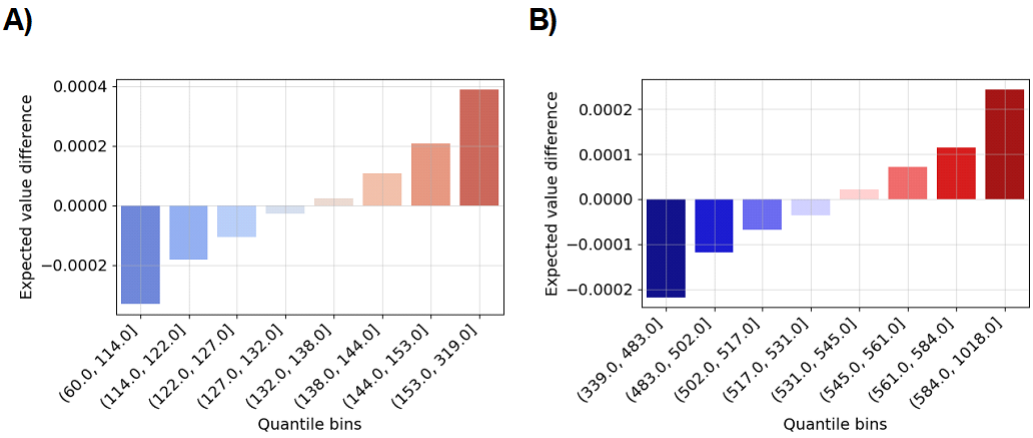

Figure A.1: Monte Carlo verification of the conditional expectation condition in our central limit theorem. Bar charts show the difference between the proportion of simulations where two specific haplotypes share an IBD segment longer than 0.03 Morgans and the true success probability (y-axis). This statistic is stratified into eight quantile bins based on the total number of long IBD segments (x¬-axis). Sample sizes are A) two hundred and B) four hundred diploids. The population size is ten thousand diploids. The expected IBD segment count is 132.78 in A) and 531.78 in B).

is to use the multiple testing correction of α/M_0, where M_0 is the number of hypothesis tests per chromosome. The probability that a marginal hypothesis test is not significant at this multiple testing level is

$$P(Z \leq z) = 1 - \alpha/M_0 \equiv 1 - \frac{1 - \Phi(z)}{M_0}. \tag{A.16}$$

The probability that M_0 independent tests are not significant is

$$\prod_{m=1}^{M_0} P(Z_m \leq z) = (1 - \alpha/M_0)^{M_0} \equiv \left(1 - \frac{1 - \Phi(z)}{M_0}\right)^{M_0}. \tag{A.17}$$

The probability that at least one of M_0 independent tests is significant is

$$P(\max_{1 \leq m \leq M_0} Z_m \geq z) = 1 - \left(1 - \frac{\alpha}{M_0}\right)^{M_0} \equiv 1 - \left(1 - \frac{1 - \Phi(z)}{M_0}\right)^{M_0}. \tag{A.18}$$

Now, let there be C chromosomes, each with M_0 independent tests. The probability that at least one of $C \times M_0$ independent tests is significant is

$$P(\max_{1\leq c\leq C}\max_{1\leq m\leq M_0} Z_{m,c} \geq z) = 1 - \left(1 - \frac{\alpha}{M_0}\right)^{M_0\times C} \equiv 1 - \left(1 - \frac{1-\Phi(z)}{M_0}\right)^{M_0\times C}. \quad \text{(A.19)}$$

As the number of chromosome-specific tests $M_0 \to \infty$ gets large, the left term is approximately $\exp(-C \cdot [1 - \Phi(z)])$:

$$P(\max_{1\leq c\leq C}\max_{1\leq m\leq M_0} Z_{m,C} \geq z) \approx 1 - \exp(-C \times \alpha) \equiv 1 - \exp(-C \times [1 - \Phi(z)]). \quad \text{(A.20)}$$

In other words, the term $\exp(-C[1-\Phi(z)])$ in Equation 6.7 may come from a scenario of many independent tests on each chromosome. This sequence of equations can be generalized for varying chromosome sizes $M_1, \cdots, M_C$ as long as $M_c \to \infty$ for all chromosomes $1 \leq c \leq C$.

The other term $\theta \cdot L \cdot z \cdot \phi(z) \cdot \nu(z\{\theta\Delta\}^{1/2})$ in Equation 6.7's exponent is the probability of the first hitting time of a standardized Ornstein-Uhlenbeck process [4, 47]. The hitting time is when the process exceeds z. The genome length is L, discrete hypothesis tests happen every Δ, and θ is the exponential decay parameter. Feingold et al. [47] suggests setting $\alpha = 1 - \Phi(z) - \theta \cdot L \cdot z \cdot \phi(z) \cdot \nu(z\{\theta\Delta\}^{1/2})$. If we work through the same sequence of equations above, we arrive at the Siegmund and Yakir [137] analytical approximation. Thus, the term $\exp(-\theta \cdot L \cdot z \cdot \phi(z) \cdot \nu(z\{\theta\Delta\}^{1/2}))$ may reflect an adjustment for correlated tests.

We compute the analytical approximation for varying Δ and θ, fixing $C = 10$ chromosomes each of length 1 Morgan. The total number of tests M is thus $10/\Delta$. Let $\hat{z}$ be the estimated threshold for confidence level $\alpha = 0.05$. We find that $\theta \cdot L \cdot \hat{z} \cdot \phi(\hat{z}) \cdot \nu(\hat{z}\{\theta\Delta\}^{1/2})$ is close to 0.05 in all cases and $(1 - \Phi(\hat{z}))/M$ is small but greater than the Bonferroni significance level α/M. We refer to M_{eff}, the solution of $M_{eff} = \alpha/(1 - \Phi(\hat{z}))$, as the effective number of tests. Figure S56 shows M_{eff}/M for varying θ and Δ. As expected, M_{eff}/M decreases as Δ decreases because many of the scanning statistics become strongly correlated. We also display a vertical line for $\theta = 50$, which is close to our estimates of this parameter in simulated and real human data. For $\theta = 50$, the effective ratio of tests is roughly 60% when testing every 0.2 cM, and it is roughly 15% when testing every 0.02 cM. If the recombination

rate is uniform along the genome (no hotspots), the step size $\Delta = 0.02$ is close to the base pair step size every 20 kb that we use in our TOPMed and UKBB selection scans.

Appendix B

ADDITIONAL SIMULATIONS CONCERNING SELECTION COEFFICIENT ESTIMATION

B.1 "Sufficiency" of the selection coefficient estimator

We implement optimization of s in Equation 4.6 over any discretized length distribution, not just the rate of detectable IBD segments greater than some length threshold. We also extend the optimization to allow for length distributions over subsamples with and without a sweeping allele. If the estimator based on the unlabelled detectable IBD rate is a sufficient statistic, then the number of detectable IBD segments captures all the relevant signal, without knowledge of the specific segment lengths or allele labels.

Without analytical results for sufficiency, we conduct simulation studies with the following parameter options: population bottleneck or three phases of exponential growth demographic scenarios, two or five thousand diploid samples, sweeping allele frequency twenty-five or fifty percent, and selection coefficient $0.005 \leq s \leq 0.035$. For each combination of parameters, we perform twenty thousand simulations. We report the mean selection coefficient estimate and the mean absolute deviation between the true selection coefficient and estimates.

Figure S20 shows nearly identical estimation results between the unlabelled rate and length distribution approaches, regardless of the true selection coefficient, the sweeping allele frequency, and the demographic scenario. This empirical result suggests that the IBD length distribution may not give additional signal beyond the detectable IBD rate for the purposes of estimating $s \geq 0.005$. Figure S21 shows smaller mean absolute deviations in estimating $s \leq 0.015$ when given allele labels. The labels do not improve estimation when $s > 0.015$. We do not report results for other parameter combinations because our findings were similar.

B.2 Plasmodium falciparum example

Our simulation studies in Chapter 4 concern selection coefficients $s \leq 0.05$. There is limited to no evidence for selection coefficients $s > 0.05$ in human populations. Examples of dramatic adaptive evolution may have occurred in other organisms like *Plasmodium falciparum* parasites exposed to anti-malaria drugs [65] or *Anopheles gambiae* mosquitos [6, 7, 57]. Figure S22 shows fifty bootstraps of the WF process near fixation and $0.10 \leq s \leq 0.40$. Compared to sweeps near fixation $p(0) = 0.98$ and $s \leq 0.05$ (Figure 4.1), it is apparent the impact of $s \geq 0.10$ within the past twenty generations. Our IBD-based estimator may be especially suitable for modeling selective sweeps with such rapid allele frequency changes. When $s > 0.01$, we observe poor estimation performance among the methods we compare to in Chapter 5. When $s > 0.05$, we observe poor estimation performance for a time-series approach in Section B.3.

In the following simulation study, we consider a demographic scenario that closely mirrors recent effective population sizes inferred for *Plasmodium falciparum* [65]. The effective population size has been exponentially decreasing at a four percent rate since three hundred generations ago. Its current effective population size is three thousand haploids. We refer to this example as the *Pf.* demography. The present-day samples are haploids as well. There are two hundred estimates for each combination of parameters.

First, we assess the effect of sample size when estimating $0.10 \leq s \leq 0.40$ and $p(0) = 0.5$. Figure S23 shows that selection coefficient estimates increase as the true selection coefficient increases, which we expect given the monotone relationship between the detectable IBD rate and the selection coefficient (Figure S15). The variance of point estimates increases considerably as the true selection coefficient increases[1]. Between samples of size fifty and five hundred, we observe similar results. Consistent with Figure 4.2A, estimation with smaller sample sizes may be feasible for large s and recent effective population sizes in the thousands.

[1]For selection coefficients $s \geq 0.10$, the sweeping allele could have arisen within the past one hundred generations. When there is a limited number of recombinations separating the original ancestor of the adaptive allele from its current descendants, there is high variance.

Second, we assess the effect of the present-day allele frequency when estimating $0.10 \leq s \leq 0.40$ with fifty haploid samples. Figure S24 shows that estimation of $s \geq 0.20$ is more accurate for $p(0) = 0.98$ fixation than for $p(0) \leq 0.75$. Recall that Figure 4.2C shows poor estimation of $s \leq 0.04$ when $p(0) = 0.9$. Even though the trajectory of selective sweeps is logistic in shape, we still observe large excess IBD rates because the sweeping allele frequency changes rapidly in recent generations with such strong selection coefficients (Figure S22). The trends in estimation are otherwise the same as in Figure S23.

Lastly, in these examples of very large $s \geq 0.10$ and the *Pf.* demography, we estimate coverage probabilities for our selection coefficient confidence intervals. Figure S25 shows coverage estimates for standard normal and percentile-based confidence intervals of $s = 0.10, 0.20, 0.30, 0.40$. Regardless of sample size, the standard normal-based confidence intervals tend to undercover the true selection coefficients $s \geq 0.20$. We observe coverage estimates of 95% standard normal-based confidence intervals that are less than 90% when $s \geq 0.30$. In comparison, we observe coverage estimates of percentile-based confidence intervals that are closer to 95% than those of the standard normal-based confidence intervals.

B.3 Performance relative to a time series-based method

Our IBD-based selection coefficient estimator exploits genetic relatedness among samples in a single present-day generation to learn about the evolution of a sweeping allele in the past. The length of an IBD segment can serve to calibrate the time depth of common ancestry around a locus [149]. For example, when two specific haplotypes share an IBD segment longer than 2.0 cM overlapping a locus, Figure 2.4 shows that the coalescent time of their common ancestor can be different around a selected locus as opposed to around a neutral locus. On the other other, some methods to estimate selection coefficients explicitly use allele frequency data from past generations [99, 100, 152].

Here we modify the estimation approach in Mathieson and McVean [99] to compare our estimator versus an approximate MLE over changes in allele frequency. Mathieson and McVean [99] define their formulas in forward time, so, in the following derivations, $t = 0$

in $p(0)$ is the time of *de novo* mutation and time $t = T$ in $p(T)$ is the time T generations after *de novo* mutation. First, we use the haploid model. Now, Mathieson and McVean [99] express the binomial probability mass

$$\begin{aligned} P(p(t+1) = p \,|\, p(t), N_e(t+1)) = & \binom{N_e(t+1)}{p \cdot N_e(t+1)} \\ & \times \left(\frac{p(t) + s \cdot p(t)}{1 + s \cdot p(t)} \right)^{p \times N_e(t+1)} \\ & \times \left(\frac{1 - p(t)}{1 + s \cdot p(t)} \right)^{(1-p) \times N_e(t+1)}, \end{aligned} \tag{B.1}$$

where $N_e(t)$ and $p(t)$ are the effective population size and sweeping allele frequency at t generations after *de novo* mutation. Second, for constant effective population size N_e, they simplify the log-likelihood of s given $p(t)$ for all t,

$$\ell(s) \propto N_e \sum_{t=1}^{T} p(t+1) \cdot \log(1+s) - \log(1 + s \cdot p(t)), \tag{B.2}$$

where terms not depending on s are dropped. Third, they solve for s that maximizes the log-likelihood. We do this by taking the first derivative, setting the equation to zero, and taking a first-order Taylor series expansion when $s \approx 0$. The Mathieson and McVean [99] selection coefficient estimator becomes

$$\hat{s}_{M\&M} = \frac{p(T) - p(0)}{\sum_{t=0}^{T-1} p(t) \cdot (1 - p(t))}. \tag{B.3}$$

Equation B.3 works the same for the multiplicative diploid model: $p(t+1)$ in terms of $p(t)$ and s is the same formula, and the factors of 2 cancel out.

Mathieson and McVean [99] do not consider non-constant effective population sizes in their study. Following the same derivations, we modify their first-order approximation for non-constant effective population sizes $N_e(t)$. Our derivations lead to the modified estimator

$$\hat{s}_{M\&M+} = \frac{\sum_{t=0}^{T-1} (p(t+1) - p(t)) \times N_e(t+1)}{\sum_{t=0}^{T-1} p(t) \cdot (1 - p(t)) \times N_e(t+1)}. \tag{B.4}$$

In Mathieson and McVean [99], the authors handle the lack of samples at certain generations by fitting a hidden Markov model. Given regular sampling of historical allele frequencies, we

instead interpolate allele frequencies between two adjacent data points and compute $\hat{s}_{M\&M+}$ from Equation B.4.

In simulations, we compare this modified approach versus our IBD-based selection coefficient estimator by estimating $0.01 \leq s \leq 0.05$ and $0.1 \leq s \leq 0.4$. We explore the following parameter choices: demographic scenarios population bottleneck or a population of constant size ten thousand diploids, sampling allele frequencies every five, ten, or twenty generations, historical sample sizes of size five, ten, twenty, forty, or sixty diploids, and times of the final data point being fifty, one hundred, or two hundred generations ago. Given a combination of these four parameters, we simulate WF processes as input to the $\hat{s}_{M\&M+}$ estimator. We use Algorithm 1 to simulate detectable IBD segments ≥ 3.0 cM as input to our IBD-based estimator. We use the deterministic formula for $p(t)$, so the following results do not reflect variance due to genetic drift, which should be small in our demographic scenarios of uniformly large effective population sizes. The sample size in the current generation is one thousand diploids. The sweeping allele frequency is fifty percent. For each parameter configuration, we perform two hundred simulations. Unless otherwise specified, the default configuration for the modified time series approach is ancient sample sizes of twenty diploids, the last sampling of ancient data was one hundred generations ago, and sampling is done every ten generations. Figure S26 illustrates the WF processes for different sampling designs in the modified time series approach.

In Figure S27A, the exact Mathieson and McVean [99] estimator $\hat{s}_{M\&M}$ (Equation B.3) shows near perfect accuracy. On the other hand, Figure S28 shows that the approximation estimator $\hat{s}_{M\&M}$ can have moderate bias when $s \geq 0.10$. Figure S27B shows that the variance in point estimates can be similar between the modified time series-based estimator and the IBD-based estimator when historical sample sizes are as small as ten diploids. Figure S27C shows that the variance in point estimates increases for the modified time series-based estimator as sampling frequency decreases, but the variance in point estimates remains uniformly smaller than those of the IBD-based estimator. Figure S27D shows that the variance in point estimates can be similar between the modified time series-based estimator

and the IBD-based estimator when the last historical samples come from fifty generations ago. Figure S29A shows that the point estimates in the modified time series-based approach appear to resemble normal distributions, so parametric bootstrapping with standard normal quantiles may provide sensible symmetric confidence intervals using the $\hat{s}_{M\&M}$ estimator [2]. Simulating the WF process with selection is very fast, so parametric bootstrapping with percentile-based confidence intervals is also an option.

Next, we consider estimation in scenarios of non-constant effective population sizes. First, we study the population bottleneck demographic scenario. In Figure S30A-C, when $s \leq 0.02$, we observe uniformly greater variance in point estimates between our IBD-based estimator versus the $\hat{s}_{M\&M+}$ time series-based estimator. On the other hand, when $s \geq 0.03$, in most cases the variance in point estimates is smaller for our IBD-based estimator versus $\hat{s}_{M\&M+}$. In particular, when $s \geq 0.03$, the variance in point estimates is comparable between the IBD and time series-based approaches only once historical sample sizes exceed twenty diploids. For the three phases of exponential growth scenario, we observe uniformly smaller or comparable variance in point estimates between our IBD-based estimator versus the time series-based $\hat{s}_{M\&M+}$. Figures S30D and S31D indicate that the $\hat{s}_{M\&M+}$ we derive may be unbiased for non-constant effective population sizes. For simulations of non-constant effective population sizes, the point estimates in the modified time series approach appear to resemble normal distributions, so parametric bootstrapping with standard normal quantiles may provide sensible symmetric confidence intervals using the $\hat{s}_{M\&M}$ estimator (Figure S29B-C).

To model hard and recent selective sweeps, designing an experiment to detect many long IBD segments in a present-day sample may sometimes be preferable to sampling allele frequencies over time. One consideration is the cost to sequence many present-day samples versus relatively fewer samples but interspersed across multiple generations. Another consideration is that IBD-based approaches require accurate genotype calls, genetic maps, and haplotype phasing, which may be challenging to obtain for non-model organisms. On the

[2]Mathieson and McVean [99] also provide plots of point estimates that resemble normal distributions in simulations of $s \leq 0.06$.

other hand, DNA from samples in prior generations could be degraded, affecting the accuracy of allele frequency estimates and possibly introducing lots of missing data. The Mathieson and McVean [99] approach also applies to binary phenotypes, which could be cost-effective to measure phenotypic adaptation in species with short generation times. Overall, our IBD-based methods to model recent positive selection complement time series-based methods [99, 100] as a more or less suitable approach depending on the population being studied.

Appendix C

THE NUMBER OF RECOMBINATION ENDPOINT COMPARISONS

C.1 The expected subtree sizes near the root of a tree

Here we argue that a large proportion of the $\binom{n}{2}$ total recombination endpoint comparisons are unnecessary and/or redundant when simulating IBD segments around a locus that are longer than a couple of centiMorgans. Before the final coalescent event, there are two subtrees of degree j and $(n-j)$. There are thus $j \cdot (n-j)$ comparisons of recombination endpoints to be made across the two subtrees. The worst case number of recombination endpoint comparisons is when the two subtrees are of equal size $\lfloor n/2 \rfloor$ and $\lceil n/2 \rceil$.

At *and* near the root of a random bifurcating tree, Theorem C.1 says that the expected number of recombination endpoint comparisons (plus some noise) divided by the worst case number of recombination endpoint comparisons approaches 1 as the sample size gets large. The Binomial(n, 1/2) random variable's coefficient of variation may provide some intuition for this result, where one-half is the probability a randomly chosen haplotype is assigned to one of two subtrees. The coefficient of variation $n^{-1/2}$ means that the average is much larger than the standard deviation. For bounded k, if sizes of subtrees at coalescent time $T^{+}_{n:k}$ are expected to be of order n, then we expect there to be recombination endpoint comparisons of the order n^2 at the final $k-2$ coalescent events in the simulation.

Without loss of generality, we consider a rooted tree with the MRCA at the top of the tree. We assume that there are no multiple merger events. First, we work downward from the root of a coalescent tree of n sample haplotypes. Throughout we assume that n equals a power of 2 to simplify the floor and ceiling functions $\lfloor n/2^j \rfloor = \lceil n/2^j \rceil$ for $j \in \mathbb{N}$.

At the coalescent event T_2, the tree bifurcates into two subtrees. At the coalescent event

T_3, the scenario with the worst case number of comparisons is subtrees of size $n/2, n/4$, and $n/4$. In general, at each coalescent event, the worst case is to split in half the largest subtree.

We provide a result for the expected number of comparisons after j splits along a path from the root. Specifically, if B_j is the size of one subtree at the j^{th} split from a subtree of size B_{j-1}, the number of comparisons $B_j(B_{j-1} - B_j)$ is expected to be of the same magnitude as in the worst case tree. Our result concerns a bounded number of standard deviations from the expected value, which is a stronger notion than the average number of computations $\Theta(\cdot)$. The proof is long, but the general strategy is to recursively apply the law of total covariance and identify the exponents in the dominating terms.

Theorem C.1. *Let $B_j \sim$ Binomial$(B_{j-1}, 1/2)$ for $j \geq 1$ and $B_0 = n$. The index j is such that $2^{2j} = O(1)$.*

$$\lim_{n\to\infty} \frac{\mathbb{E}[B_j(B_{j-1} - B_j)] + O(1) \cdot Cov^{1/2}(B_j(B_{j-1} - B_j))}{n^2/2^{2j}} = 1. \tag{C.1}$$

Proof. We must calculate the expected value and the covariance in the numerator. Let $B \sim$ Binomial(m, 1/2).

$$\begin{aligned} \mathbb{E}[B(B-1)] &= \mathbb{E}[B^2] - \mathbb{E}[B] \\ &= m/4 + m^2/4 - m/2 \\ &= m(m-1)/4 \\ &= m(m-1) \cdot 2^{-2 \cdot 1}. \end{aligned} \tag{C.2}$$

Using the law of total expectation, we solve the expected value for $j = 2$.

$$\begin{aligned} \mathbb{E}[B_2(B_1 - B_2)] &= \mathbb{E}[\mathbb{E}[B_2(B_1 - B_2)|B_1]] \\ &= \mathbb{E}[B_1(B_1 - 1) \cdot 2^{-2 \cdot 1}] \\ &= n(n-1) \cdot 2^{-2 \cdot 1} \cdot 2^{-2 \cdot 1} = n(n-1) \times 2^{-2 \cdot 2}. \end{aligned} \tag{C.3}$$

Applying Equation C.2 recursively, we derive the general formula

$$\mathbb{E}[B_j(B_{j-1} - B_j)] = n(n-1) \cdot 2^{-2j}. \tag{C.4}$$

The limit of Equation C.4 divided by $n^2 \cdot 2^{-2j}$ is one. Next, we require that the standard deviation is of order less than n^2. Using the law of total covariance, we derive in Lemma C.2 that $\text{Cov}(B_j(B_{j-1} - B_j)) \sim n^3$, where $\sim$ means asymptotically equivalent. Consequently,

$$\lim_{n\to\infty} n^{-2} \cdot \text{Cov}^{1/2}(B_j(B_{j-1} - B_j)) = 0.$$

□

Lemma C.2. *Recall that* $B_j \sim$ *Binomial*$(B_{j-1}, 1/2)$ *for* $j \geq 1$ *and* $B_0 = n$. *Then,*

$$\textit{Cov}(B_j(B_{j-1} - B_j)) \sim n^3. \tag{C.5}$$

Proof. We apply the laws of total covariance and expectation in a recursive fashion. Overall, we must control three terms:

$$\begin{aligned}\text{Cov}(B_j(B_{j-1} - B_j)) = &\text{Cov}(B_j^2, B_j^2) \\ &+ \text{Cov}(B_jB_{j-1}, B_jB_{j-1}) \\ &- 2 \cdot \text{Cov}(B_jB_{j-1}, B_j^2).\end{aligned} \tag{C.6}$$

The first four moments of the conditional binomial random variable are useful:

$$\begin{aligned}\mathbb{E}[B_j|B_{j-1}] &= 0.5 \cdot B_{j-1}; \\ \mathbb{E}[B_j^2|B_{j-1}] &= 0.5^2(B_{j-1} + B_{j-1}^2); \\ \mathbb{E}[B_j^3|B_{j-1}] &= 1 \cdot 0.5 \cdot B_{j-1} \\ &\quad + 3 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\ &\quad + 1 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2); \\ \mathbb{E}[B_j^4|B_{j-1}] &= 1 \cdot 0.5 \cdot B_{j-1} \\ &\quad + 7 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\ &\quad + 6 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2) \\ &\quad + 1 \cdot 0.5^4 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2)(B_{j-1} - 3).\end{aligned} \tag{C.7}$$

The following conditional covariances are also useful.

$$\text{Cov}(B_j, B_j|B_{j-1}) = 0.5^2 \cdot B_{j-1}^1 = O(B_{j-1}^1). \tag{C.8}$$

$$\begin{aligned}
\mathrm{Cov}(B_j, B_j^2|B_{j-1}) &= \mathbb{E}[B_j^3|B_{j-1}] - \mathbb{E}[B_j|B_{j-1}] \cdot \mathbb{E}[B_j^2|B_{j-1}] \\
&= 1 \cdot 0.5 \cdot B_{j-1} \\
&+ 3 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\
&+ 1 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2) \\
&- 0.5^3 \cdot B_{j-1}(B_{j-1} + B_{j-1}^2) \\
&= 0.5^2 \cdot B_{j-1}^2 \\
&= O(B_{j-1}^2).
\end{aligned} \tag{C.9}$$

$$\begin{aligned}
\mathrm{Cov}(B_j^2, B_j^2|B_{j-1}) &= \mathbb{E}[B_j^4|B_{j-1}] - \mathbb{E}[B_j^2|B_{j-1}] \cdot \mathbb{E}[B_j^2|B_{j-1}] \\
&= 1 \cdot 0.5 \cdot B_{j-1} \\
&+ 7 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\
&+ 6 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2) \\
&+ 1 \cdot 0.5^4 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2)(B_{j-1} - 3) \\
&- 0.5^4(B_{j-1}^2 + 2 \cdot B_{j-1}^3 + B_{j-1}^4) \\
&= -0.5^3 \cdot B_{j-1} - 0.5^3 \cdot B_{j-1}^2 + 0.5 \cdot B_{j-1}^3 \\
&= O(B_{j-1}^3).
\end{aligned} \tag{C.10}$$

Notice that all of the conditional covariances are of an order of three or less. By recursively applying the law of total expectation, we derive $\mathbb{E}[B_j^3] \sim n^3$. Another important unconditional covariance term is

$$\begin{aligned}
\mathrm{Cov}(B_j, B_j^2) &= \mathbb{E}[\mathrm{Cov}(B_j, B_j^2|B_{j-1})] + \mathrm{Cov}(\mathbb{E}[B_j|B_{j-1}], \mathbb{E}[B_j^2|B_{j-1}]) \\
&= \mathbb{E}[O(B_{j-1}^2)] + \mathrm{Cov}(O(B_{j-1}), O(B_{j-1}^2)),
\end{aligned} \tag{C.11}$$

which is asymptotically equivalent to n^2 when the total laws of expectation and covariance are applied recursively. Finally, we evaluate the asymptotic behavior of the three unconditional covariances in Equation C.6.

$$\begin{aligned}
\mathrm{Cov}(B_j^2, B_j^2) &= \mathbb{E}[\mathrm{Cov}(B_j^2, B_j^2|B_{j-1})] + \mathrm{Cov}(\mathbb{E}[B_j^2|B_{j-1}], \mathbb{E}[B_j^2|B_{j-1}]) \\
&= \mathbb{E}[O(B_{j-1}^3)] + 0.5^4 \cdot \mathrm{Cov}(B_{j-1} + B_{j-1}^2, B_{j-1} + B_{j-1}^2)
\end{aligned} \tag{C.12}$$

$$\begin{aligned}\mathrm{Cov}(B_{j-1}B_j, B_{j-1}B_j) &= \mathbb{E}[B_{j-1}^2 \mathrm{Cov}(B_j, B_j|B_{j-1})] \\ &\quad + \mathrm{Cov}(B_{j-1} \cdot \mathbb{E}[B_j|B_{j-1}], B_{j-1} \cdot \mathbb{E}[B_j|B_{j-1}]) \\ &= 0.5^2 \cdot (\mathbb{E}[B_{j-1}^3] + \mathrm{Cov}(B_{j-1}^2, B_{j-1}^2))\end{aligned} \tag{C.13}$$

$$\begin{aligned}\mathrm{Cov}(B_{j-1}B_j, B_j^2) &= \mathbb{E}[B_{j-1} \cdot \mathrm{Cov}(B_j, B_j^2|B_{j-1})] \\ &\quad + \mathrm{Cov}(B_{j-1} \cdot \mathbb{E}[B_j|B_{j-1}], \mathbb{E}[B_j^2|B_{j-1}]) \\ &= \mathbb{E}[O(B_{j-1}^3)] + 0.5^3 \cdot \mathrm{Cov}(B_{j-1}^2, B_{j-1} + B_{j-1}^2)\end{aligned} \tag{C.14}$$

By recursively applying the total laws of expectation and covariance, we conclude that Equations C.12, C.13, and C.14 are asymptotically equivalent to n^3. □

C.2 The expected subtree sizes near the leaves of a tree

A complementary perspective on joint subtree sizes we take from Dahmer and Kersting [39]. Now, we work upward from the leaves to the root. Dahmer and Kersting [39] provide a lemma for the expected number of subtrees containing r sample haplotypes at the $(n-k)^{\text{th}}$ coalescent event. We can use their moment calculations together with the expected value of the hypoexponential random variable $T_{n:k}^+$ to build intuition for the average subtree sizes at a specified generation. In a toy example, Figure C.1 shows the average number of sample haplotypes under subtrees of a given size at generations $N \cdot \mathbb{E}[T_{n:k}^+]$. Our main observation is that before the final coalescent events most sample haplotypes are expected to be under a subtree of size an order of magnitude smaller than sample size. In Figure C.1, the coalescent times are a few hundred generations. From Chapter 2 and Figure S1, we know that the probability of a long IBD segment around a locus from an ancestor at such coalescent times is very rare. These expectations give additional support to the idea that many of the $n(n-1)/2$ comparisons happen near the root and that these computations might be avoided using the pruning and merging rules in Algorithm 1.

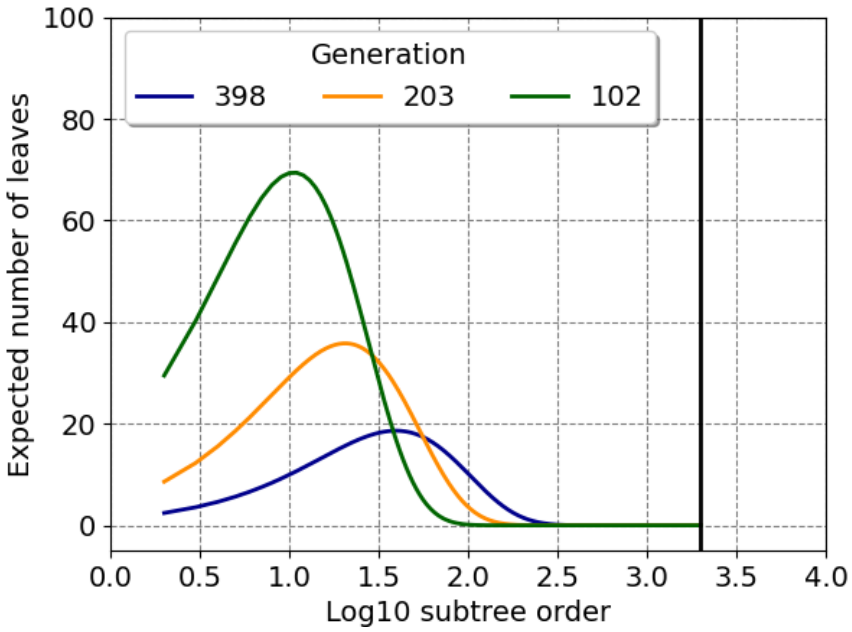

Figure C.1: The expected cardinality of subtree sizes at different coalescent times. Using Lemma 1 in Dahmer and Kersting [39], we compute the expected number of subtrees containing r samples (x-axis) at the $(n-k)^{\text{th}}$ coalescent event. We multiply these moments by r to get the expected number of leaves under such subtrees (y-axis). There are two thousand samples. Dark blue, orange, and green lines correspond to k =50, 95, and 180. We compute the expected time of the $(n-k)^{\text{th}}$ coalescent event [74] and multiply by a population size of ten thousand to get generations (legend). The vertical line is logarithm 10 of sample size.

Appendix D

SUPPLEMENTARY FIGURES

Shared haplotypes overlapping a focal point

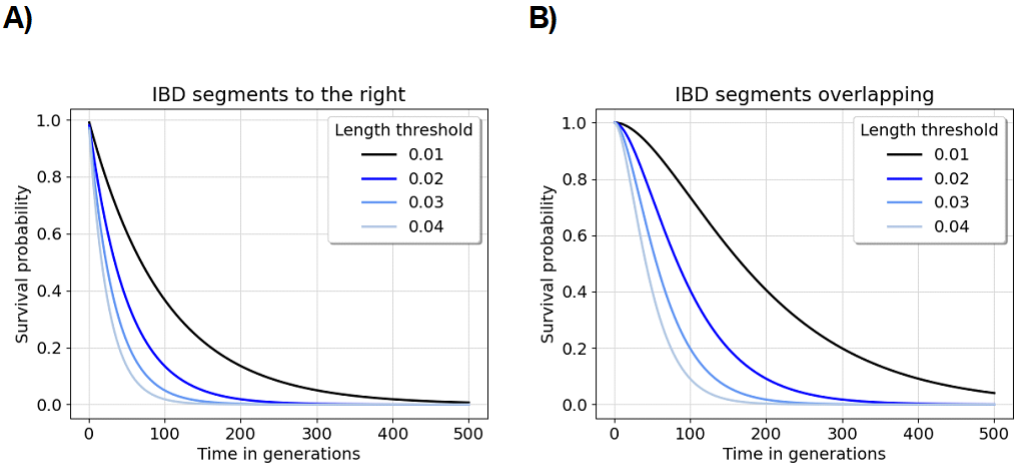

Figure S1: The survival probabilities of Erlang random variables. Subplots A) and B) show the survival probabilities for shape parameters 1 and 2, respectively. The rate of the random variables is the coalescent time in generations (x-axis). The survival probability (y-axis) comes from Equations 2.6 and 2.7. The length thresholds are denoted by different colors and line styles, defined in the legend.

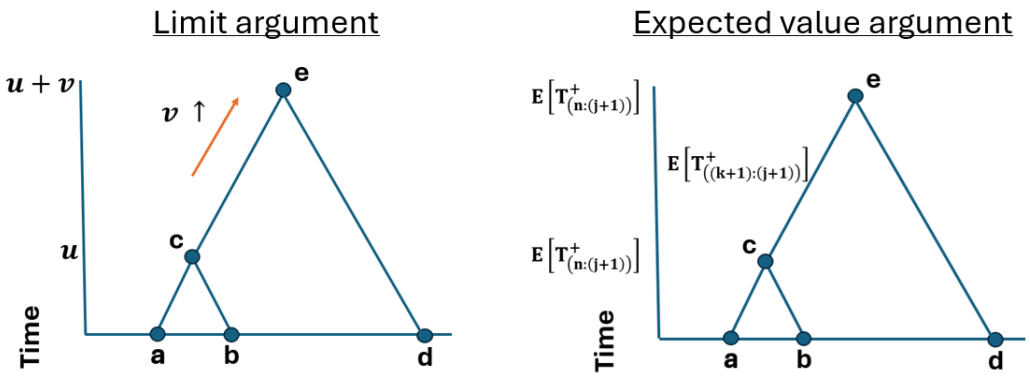

Figure S2: Illustration of coalescent tree branch lengths for merging arguments. (Left) The branch length connecting common ancestor c to its common ancestor e increases to infinity. (Right) The branch lengths of the tree are the expected values of time after $(k+1)$, $(j+1)$ coalescent events.

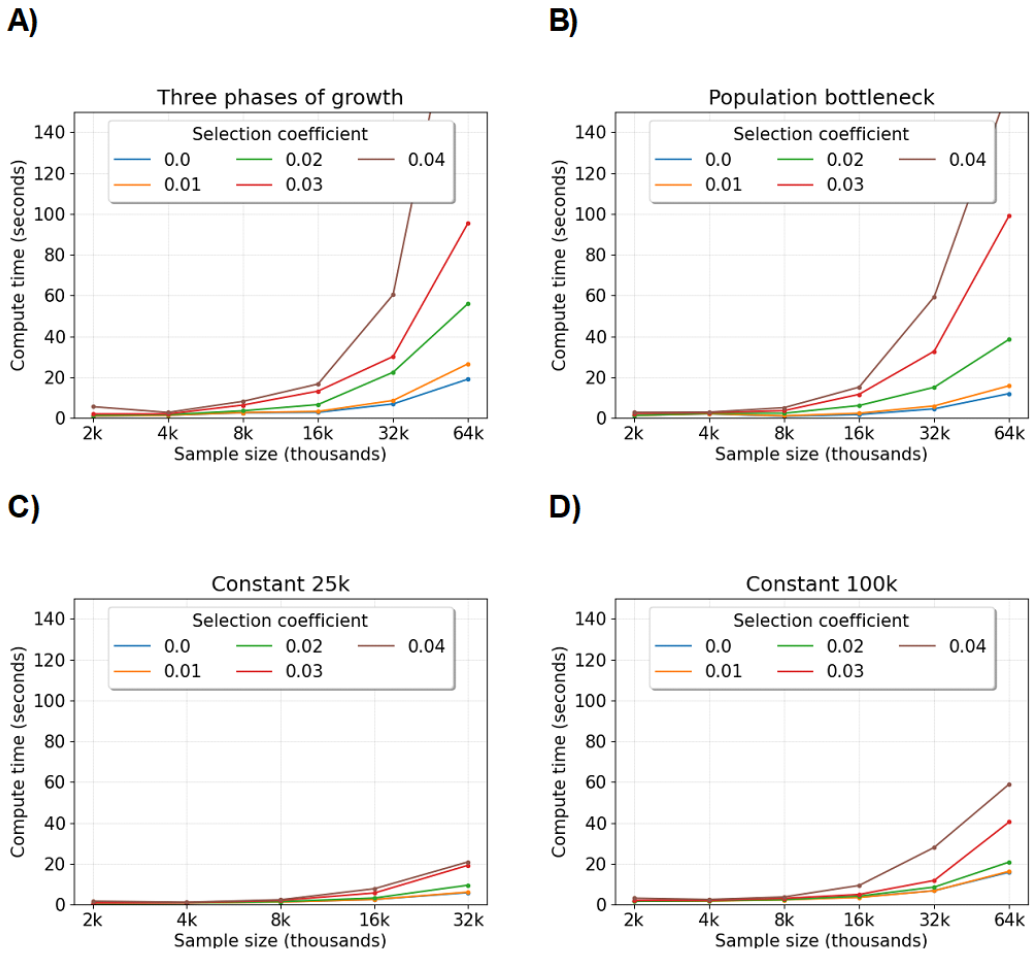

Figure S3: Compute time to simulate IBD segment lengths around a locus depending demography and selection. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legends denote colored line styles for different selection coefficients. A), B), C), and D) show results for demographic scenarios of three phases of exponential growth, a population bottleneck, and constant population sizes of twenty-five and one hundred thousand diploids, respectively. The Morgan length threshold is 0.01.

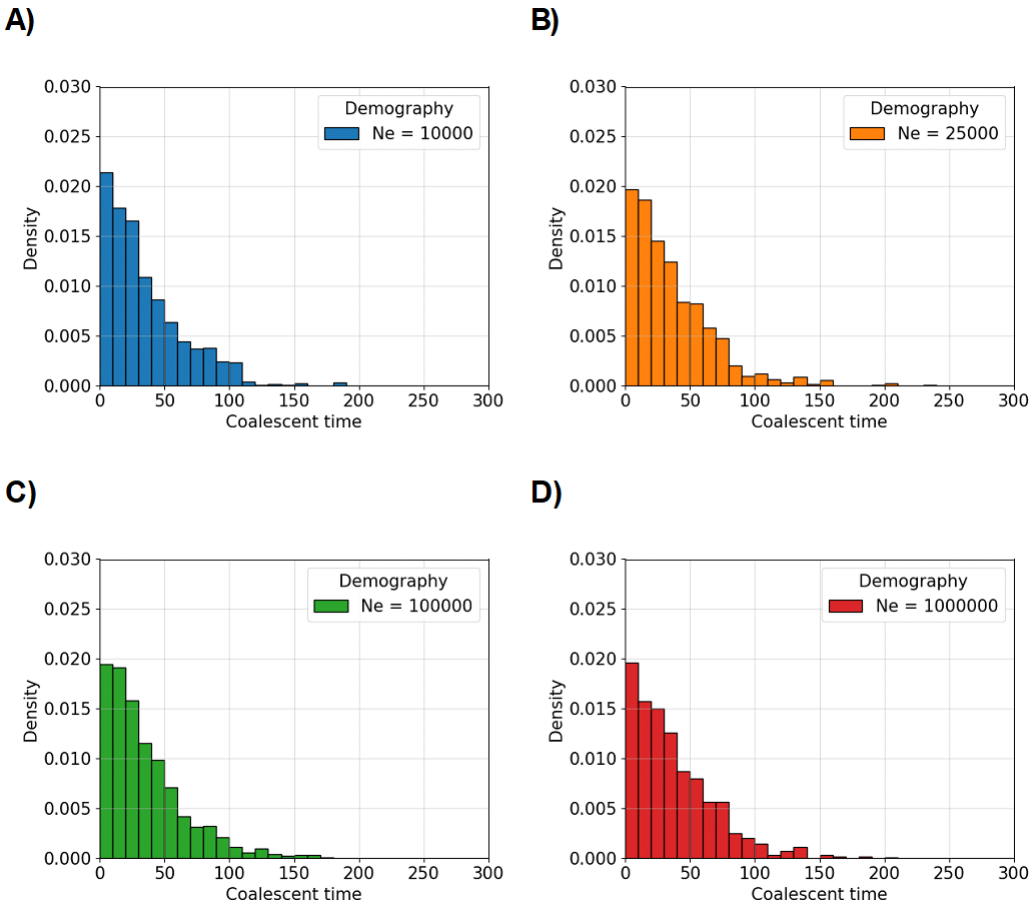

Figure S4: Empirical distributions of coalescent times conditional on detectable IBD segments as a locus for varying population size. Histograms show the density of coalescent times split into fifty bins. Each panel represents one simulation. The population size is set to A) ten thousand, B) one hundred thousand, C) one million, and D) ten million diploids. The Morgan length threshold is 0.02. The sample size is five thousand diploids. There is no selection.

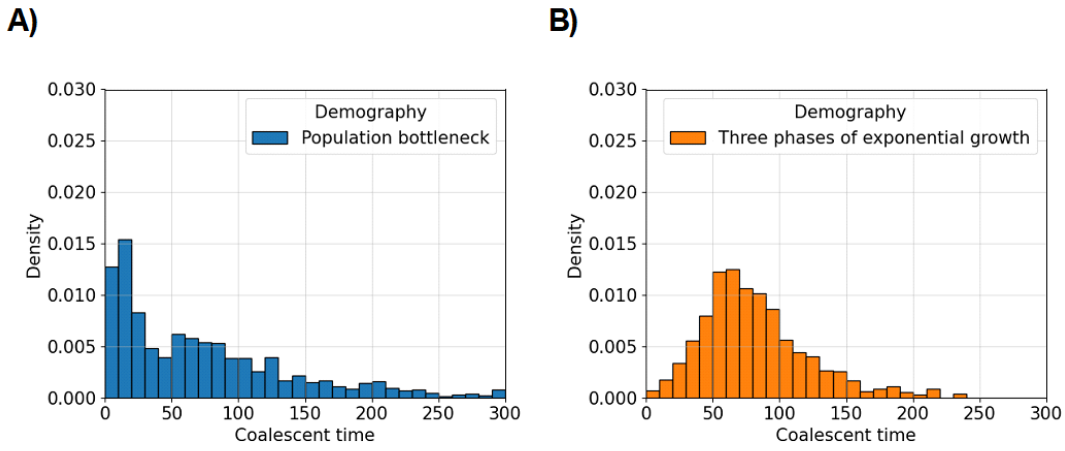

Figure S5: Empirical distributions of coalescent times conditional on detectable IBD segments around a locus for different demographic scenarios. Histograms show the density of coalescent times split into fifty bins. Each panel represents one simulation. The demographic scenario is set to A) a population bottleneck and B) three phases of exponential growth. The Morgan length threshold is 0.02. The sample size is five thousand diploids. There is no selection.

Identity-by-descent in large samples

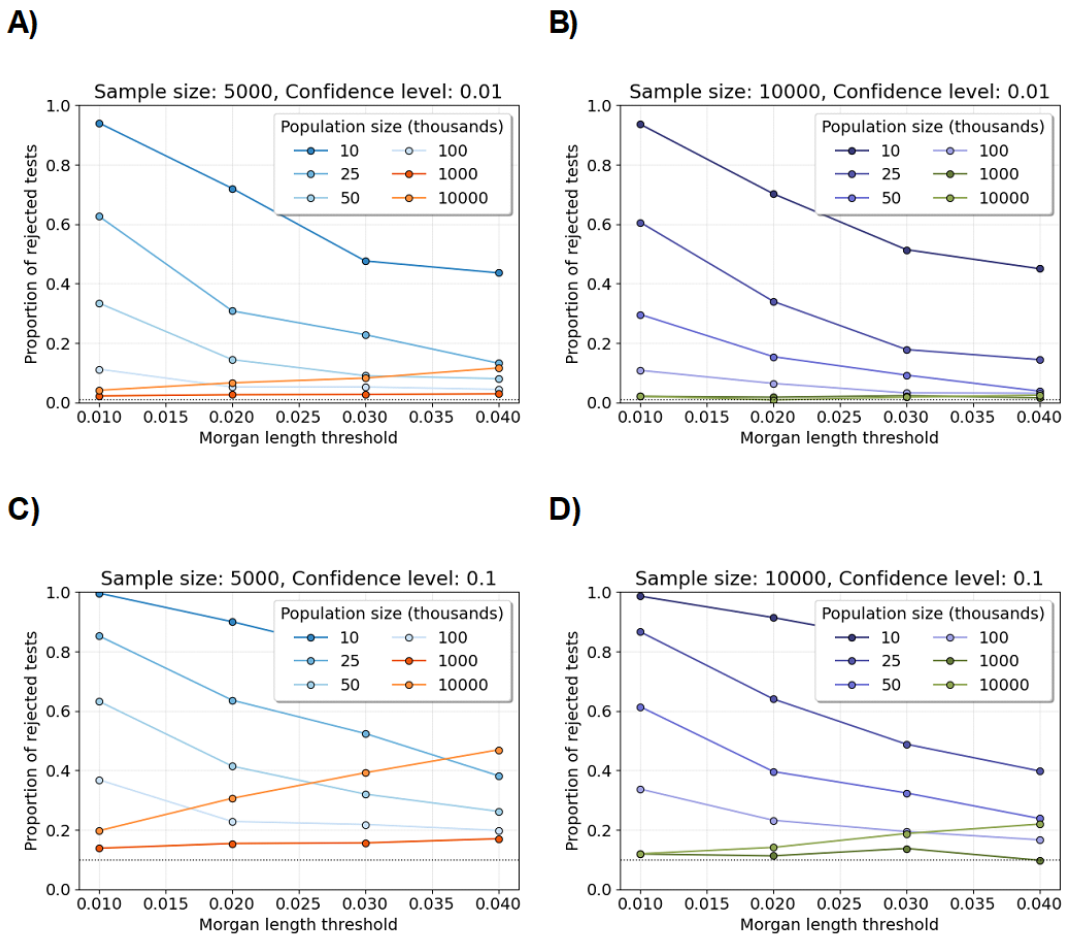

Figure S6: Shapiro-Wilk tests for varying population sizes and confidence levels. Line plots show the proportions of Shapiro-Wilk tests rejected at confidence levels A,B) 0.01 and C,D) 0.1 (y-axis) for varying population size and fixed sample size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). A,C) The sample size is five thousand diploids. B,D) The sample size is ten thousand diploids. The legends give colors assigned to different population sizes. The horizontal dotted lines are confidence levels.

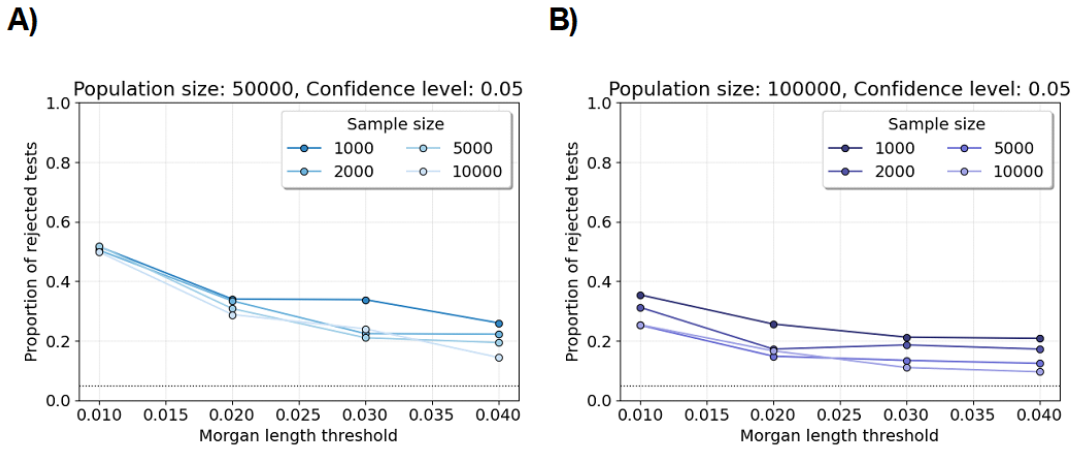

Figure S7: Shapiro-Wilk tests for varying sample sizes. Line plots show the proportions of Shapiro-Wilk tests rejected at a confidence level of 0.05 (y-axis) for varying sample size and fixed population size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). A) The population size is fifty thousand diploids. B) The population size is one hundred thousand diploids. The legends give colors assigned to different sample sizes. The horizontal dotted line is at 0.05.

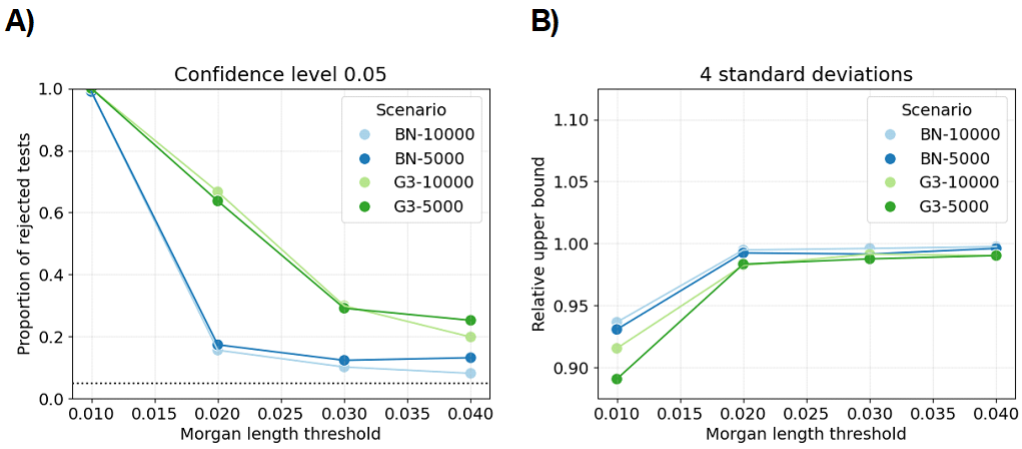

Figure S8: Shapiro-Wilk tests and relative upper tail bounds for complex demography scenarios. A) Line plots show the proportions of Shapiro-Wilk tests rejected at confidence level 0.05 (y-axis) for the population bottleneck (BN) or three phases of exponential growth (G3) demographic scenarios and sample sizes of five or ten thousand diploids. Each proportion is computed over at least six hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). B) Line plots show the average mean plus four standard deviations divided by the 99.99683 percentile over two million simulations (y-axis). Plot designs are identical to Figures 3.3 and 3.4.

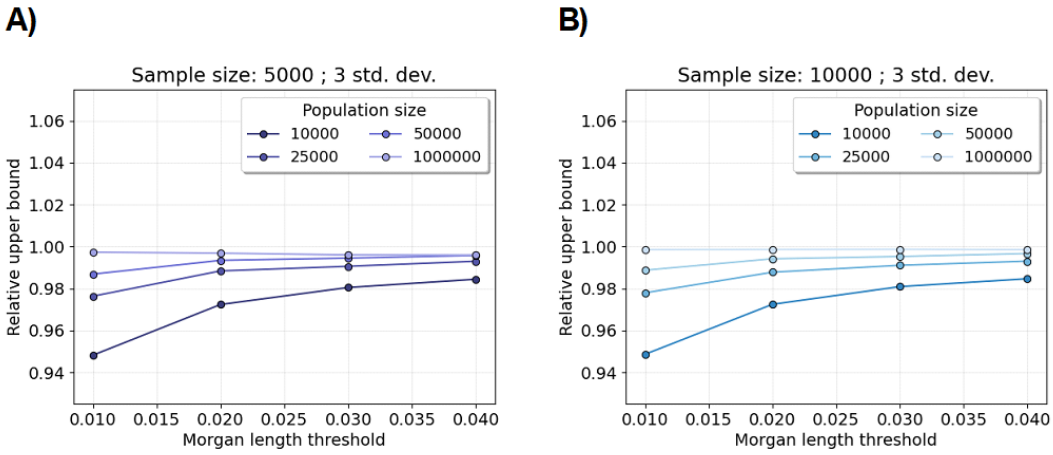

Figure S9: Relative upper bound for excess IBD scan. Line plots show the average mean plus three standard deviations divided by the 99.86501 percentile over two million simulations (y-axis). (The standard normal survival function of three is 0.9986501.) Each average relative upper bound is computed over one thousand tests. Each test is based on two thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). A) The sample size is five thousand diploids. B) The sample size is ten thousand diploids. The legends give colors assigned to different constant population sizes.

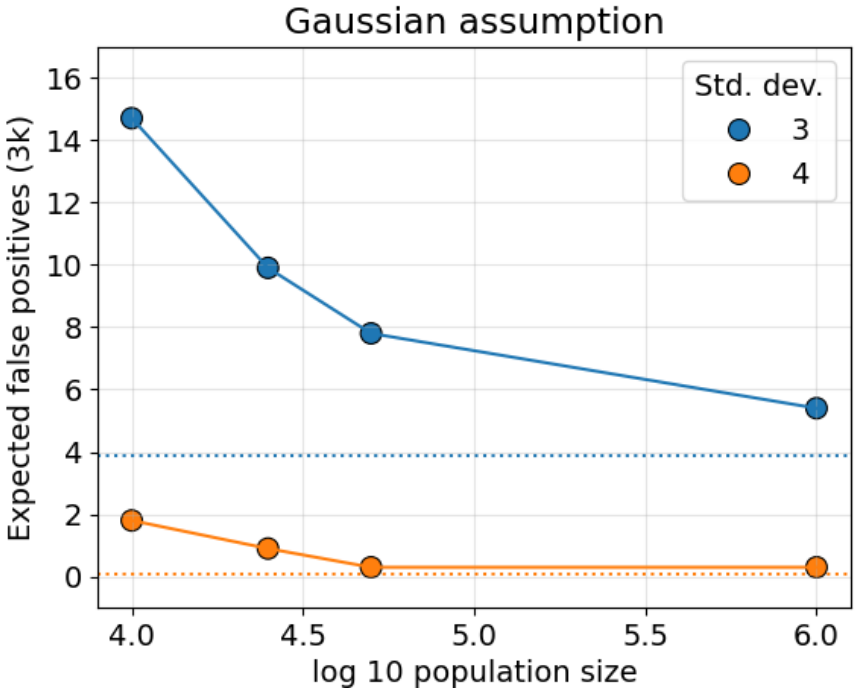

Figure S10: Expected false positives for excess IBD scan. Line plots show the expected false positive tests out of three thousand (x-axis) by the log 10 constant population size (y-axis). The z-tests use three (blue) or four (orange) estimated standard deviations above the estimated mean. The horizontal dotted blue and orange lines correspond to $1-\Phi(3)$ and $1-\Phi(4)$ times three thousand. The legend gives colors assigned to the different quantiles.

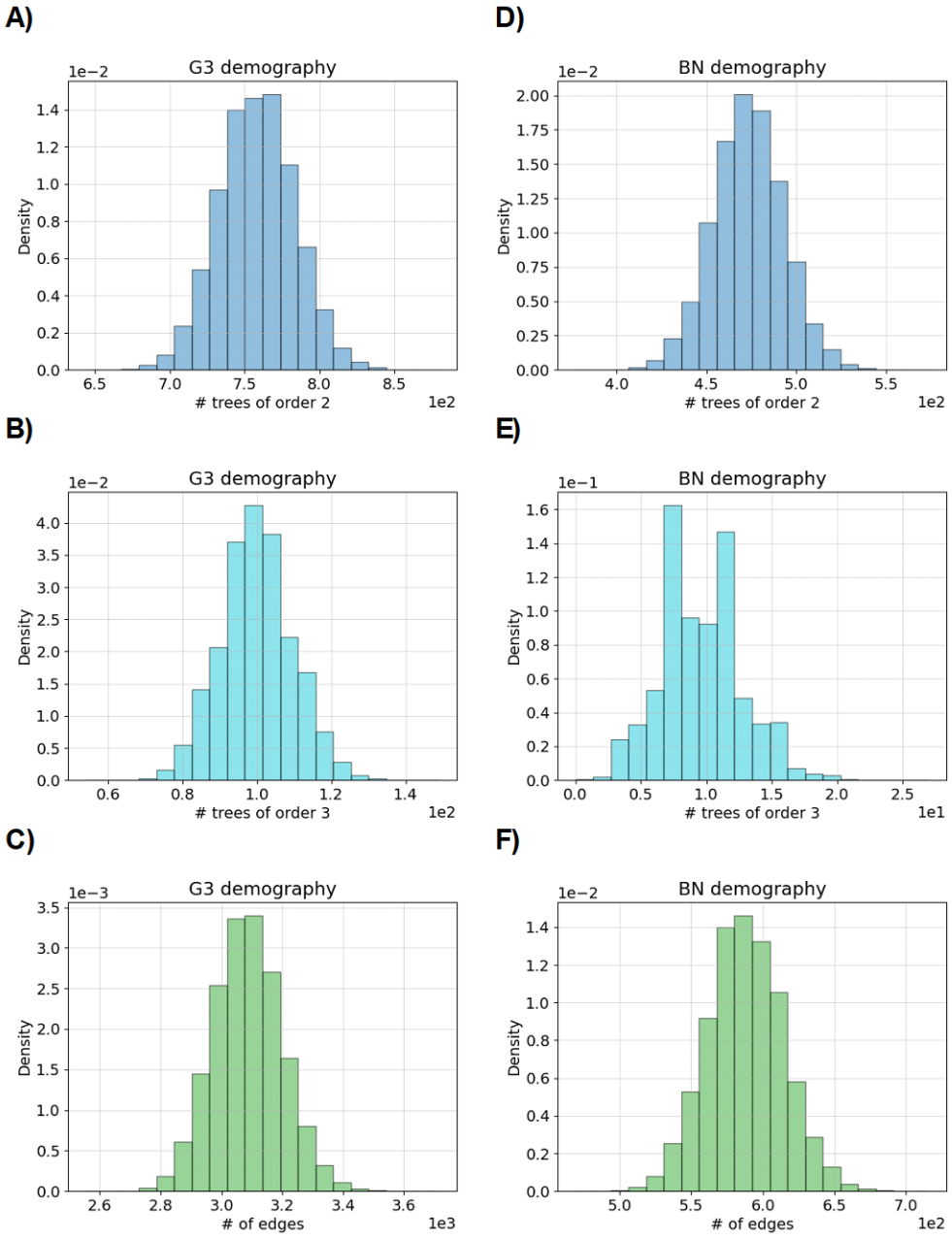

Figure S11: Comparing features between IBD graphs for complex demographic scenarios. Histograms show the density of IBD graph features between A-C) the three phases of exponential growth (G3) and D-F) the population bottleneck (BN) demographic scenarios. Each histogram is based on at least six hundred thousand simulations. A,D), B,D), and C,F) show the number of trees of order 2, the number of trees of order 3, and the total number of edges, respectively. The Morgan length threshold is 0.03. The sample size is five thousand diploids.

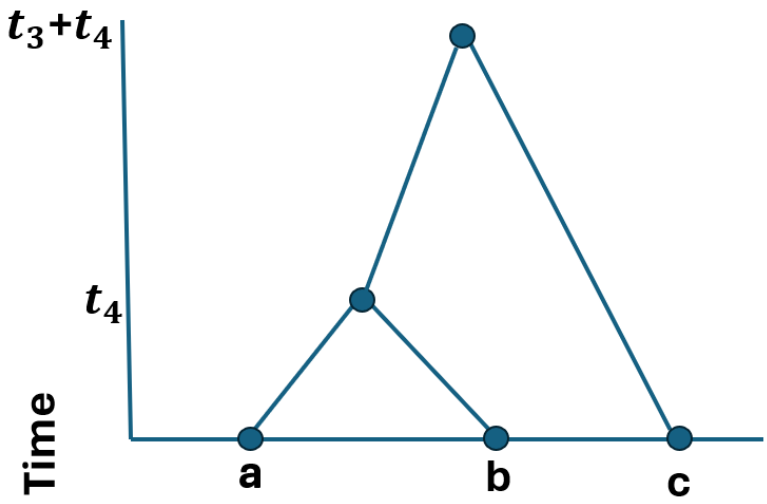

Figure S12: Illustration of the one possible coalescent tree used to calculate Cov_3 terms in Appendix A.

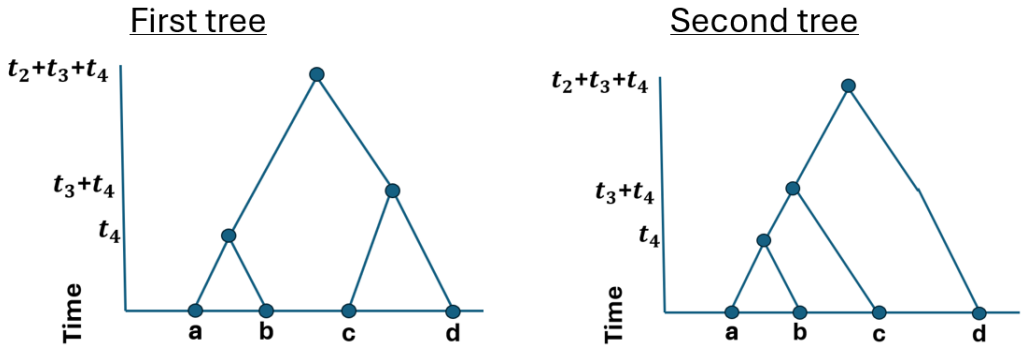

Figure S13: Illustration of the two possible coalescent trees used to calculate Cov_4 terms in Appendix A.

Selection coefficient estimation

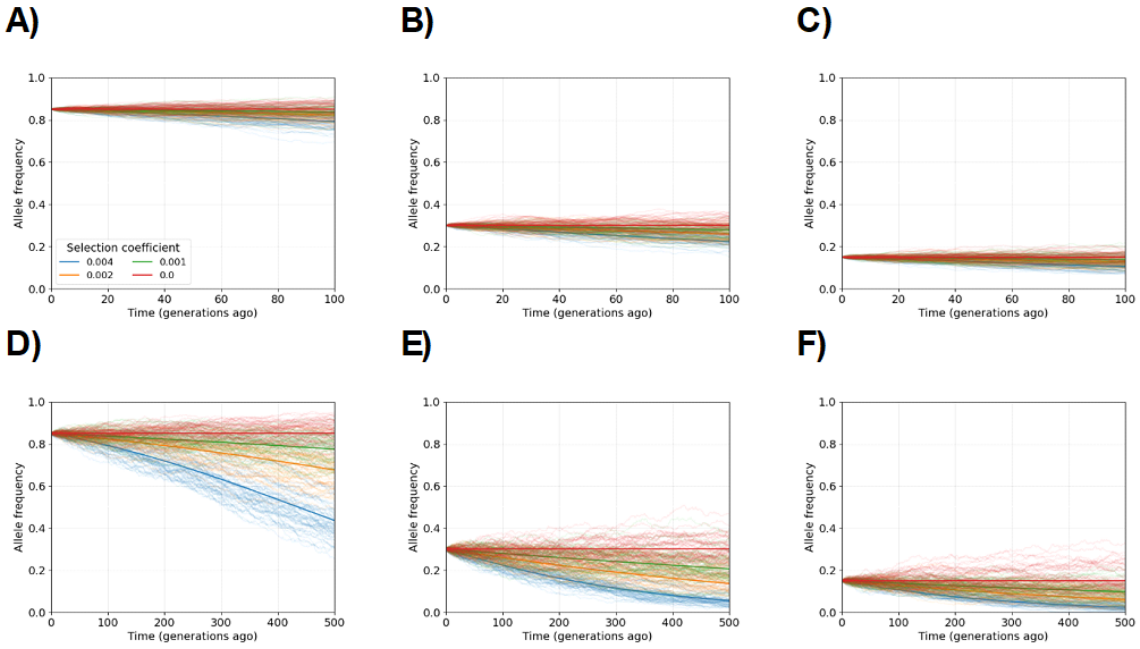

Figure S14: Wright-Fisher simulations with small selection coefficients. Fifty simulations of the Wright-Fisher process for $s \in [0.000, 0.001, 0.002, 0.004]$. Allele frequency ($y$-axis) by time ($x$-axis) is shown for the most recent (A-C) one hundred and (D-F) five hundred generations. The allele frequencies in the present day are (A,D) 0.85, (B,E) 0.3, (C,F) and 0.15. The legend in A) defines colors for different selection coefficients and applies to all subplots.

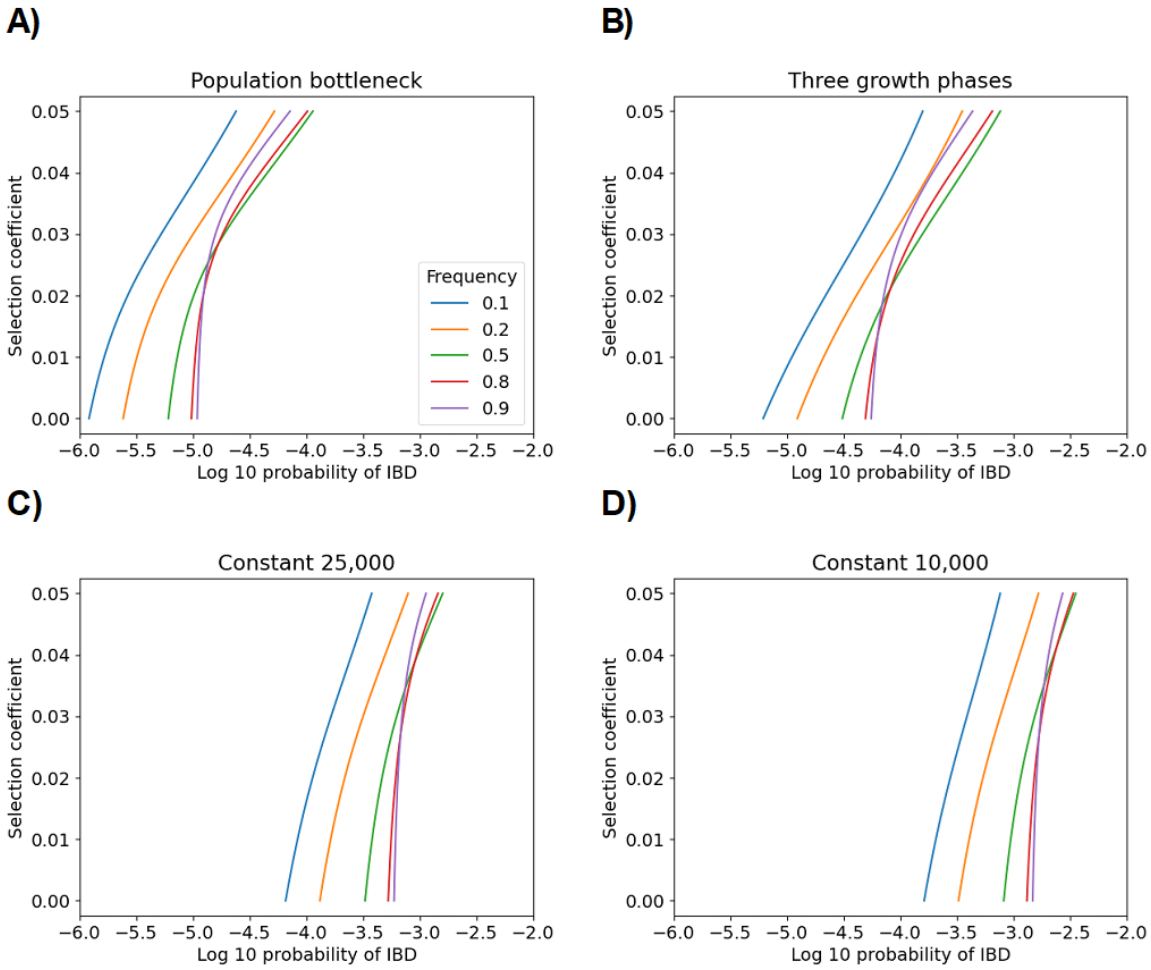

Figure S15: Differentiable maps between the probability of a detectable IBD segment and the selection coefficient. Line plots for selection coefficients (y-axis) by the $\log_{10}$ probability of an IBD segment $\geq$ 2.0 cM (x-axis) for different demographic scenarios: A) population bottleneck, B) three phases of exponential growth, C) constant size twenty-five thousand diploids, and D) constant size ten thousand diploids. Selection coefficients are discretized every 0.001. Calculation of the detectable IBD probability is based on the Palamara et al. [110] and Browning and Browning [19] approximations, extended to include selection sweeps. Colors in the legend correspond to estimation conditional on different present-day sweeping allele frequencies.

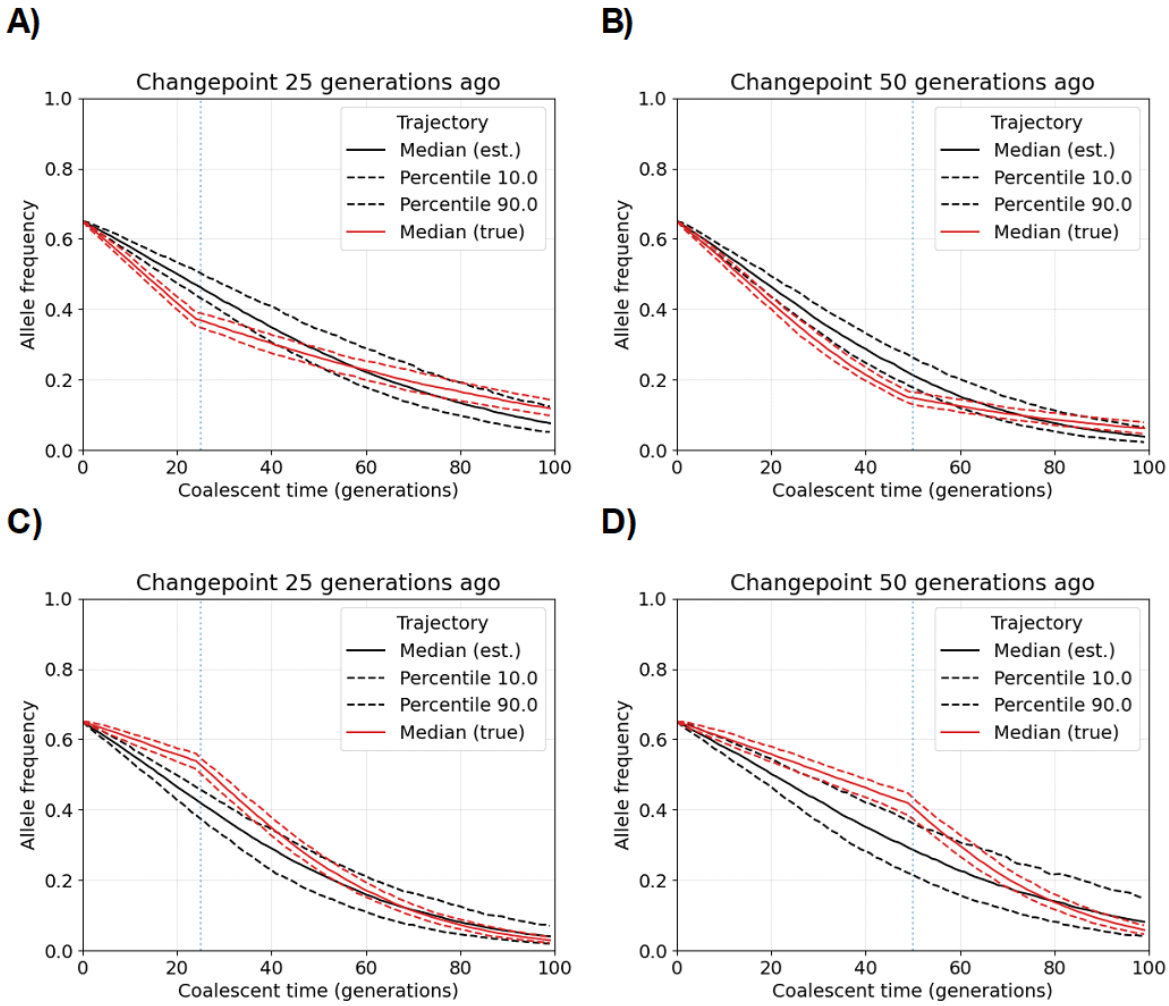

Figure S16: Estimated historical allele frequencies when selection coefficients vary over time. The plots show the medians, tenth percentiles, and ninetieth percentiles for allele frequencies $p(t)$ from two hundred parametric bootstraps over identity-by-descent and Wright-Fisher processes for selective sweeps. The selection coefficient changes A,B) from 0.05 to 0.02 and C,D) from 0.02 to 0.05. The selection coefficients change at A,C) twenty-five generations ago and B,D) fifty generations ago, denoted by the vertical blue dotted lines and annotated as subplot titles. The sweeping allele frequency is sixty-five percent in the present generation. The population size is ten thousand diploids. The sample size is one thousand diploids. Estimates are based on true IBD segment lengths greater than 3.0 cM.

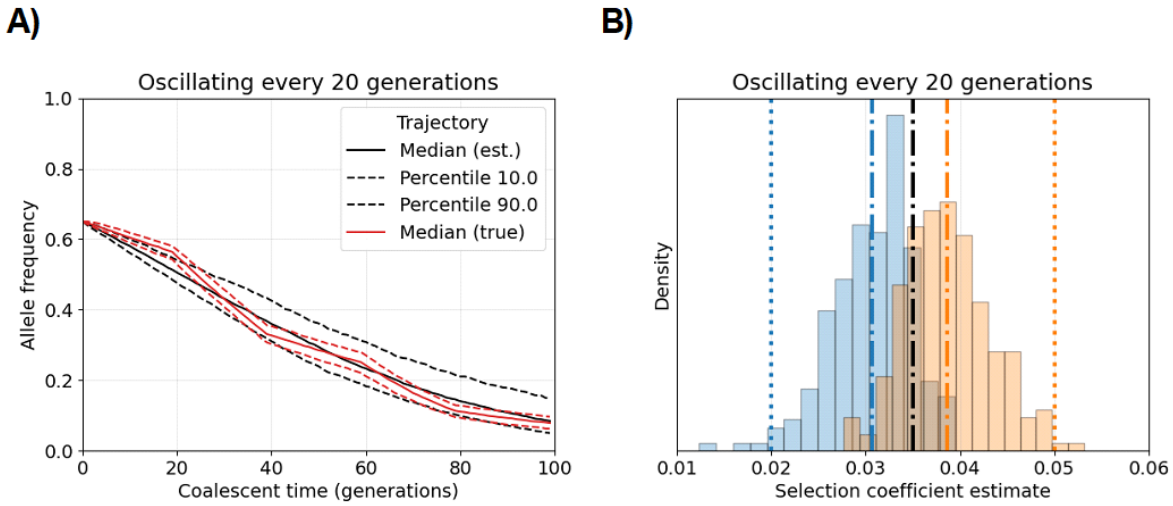

Figure S17: Estimated historical allele frequencies and selection coefficients when selection coefficients oscillate over time. A) The plot shows medians, tenth percentiles, and ninetieth percentiles for allele frequencies $p(t)$ from two hundred parametric bootstraps over identity-by-descent and Wright-Fisher processes for selective sweeps. The selection coefficient changes between 0.02 and 0.05 every twenty generations, starting at $s = 0.02$ in the present generation. B) Histogram of selection coefficient estimates for oscillating selection scenario when $s = 0.02$ (blue and vertical dotted line) or $s = 0.05$ (orange and vertical dotted line) in the present generation. Vertical dotted-dashed lines are the average estimates, and the vertical black line $s = 0.035$ is the midpoint between $s = 0.02$ and $s = 0.05$. The sweeping allele frequency is sixty-five percent in the present generation. The population size is ten thousand diploids. The sample size is one thousand diploids. Estimates are based on true IBD segment lengths greater than 3.0 cM.

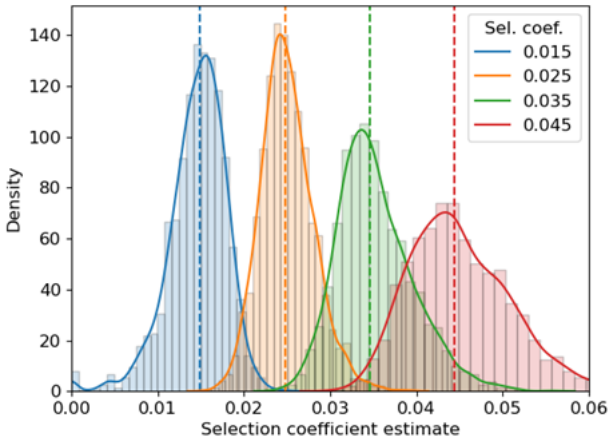

Figure S18: Sampling distribution of the selection coefficient estimator. Histograms show data for selection coefficient estimates of $s = 0.01, 0.02, 0.03, 0.04$ (colors in legend) from 2500 replicate simulations. The population bottleneck is the demographic scenario. The sample size is five thousand diploids. Vertical dashed lines denote the averages of the sampling mean. Kernel density estimates are overlaid on the histograms.

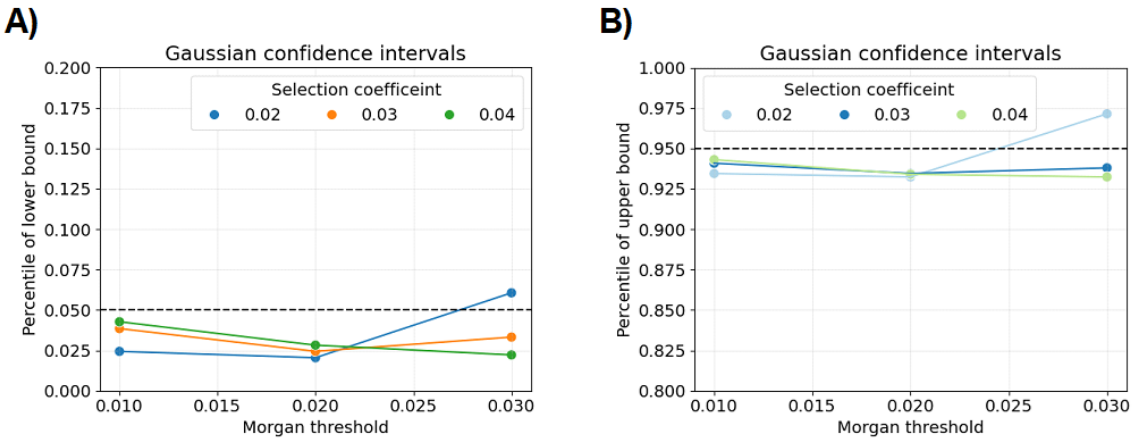

Figure S19: Sampling distribution percentiles of the lower and upper standard normal bounds. Each sampling distribution is determined from 37,500 simulations of true selection coefficients $s = 0.01, 0.02, 0.03$ (color legends). The average percentile of A) lower and B) upper bounds for standard normal confidence intervals are shown as the y-axis variable. The standard deviations are calculated from 250 simulations. Averages are taken over 150 replicates. The significance level is 0.10. Horizontal dashed lines denote 0.05 and 0.95. The population bottleneck is the demographic scenario. The sample size is five thousand diploids. The detection threshold is 3.0 cM.

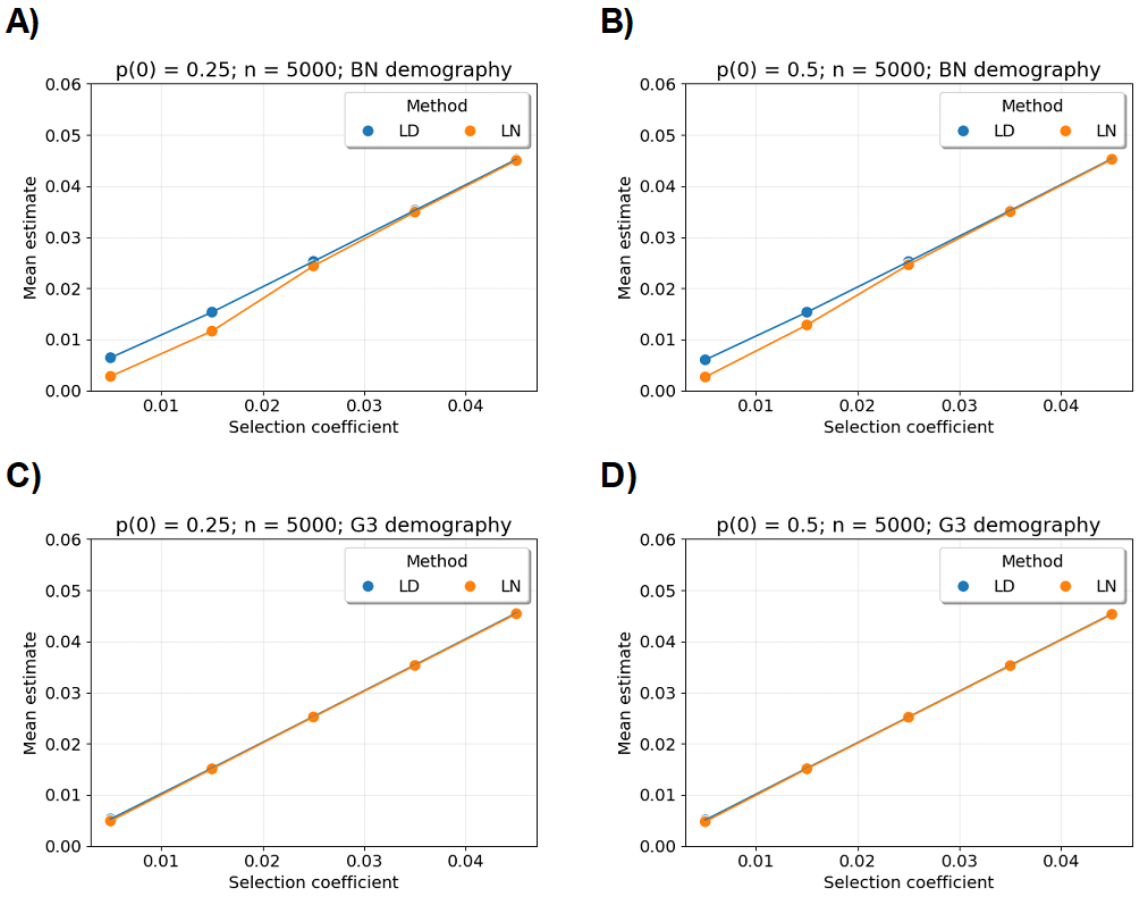

Figure S20: Average selection coefficient estimates using the number of detectable IBD segments versus the IBD length distribution. Line plots show the average overall estimates and the true selection coefficients $s = 0.005, 0.015, 0.025, 0.035$. The estimation methods are denoted in the legend and assigned colors: length distribution (LD) (blue); and the number of detectable IBD segments (LN) (orange). A,B) Population bottleneck (BN) is the demographic model versus C,D) three phases of exponential growth (G3) is the demographic model. A,C) The sweeping allele frequency $p(0)$ is twenty-five percent versus B,D) the sweeping allele frequency $p(0)$ is fifty percent. The sample size n is five thousand diploids. Each data point is based on twenty thousand simulations.

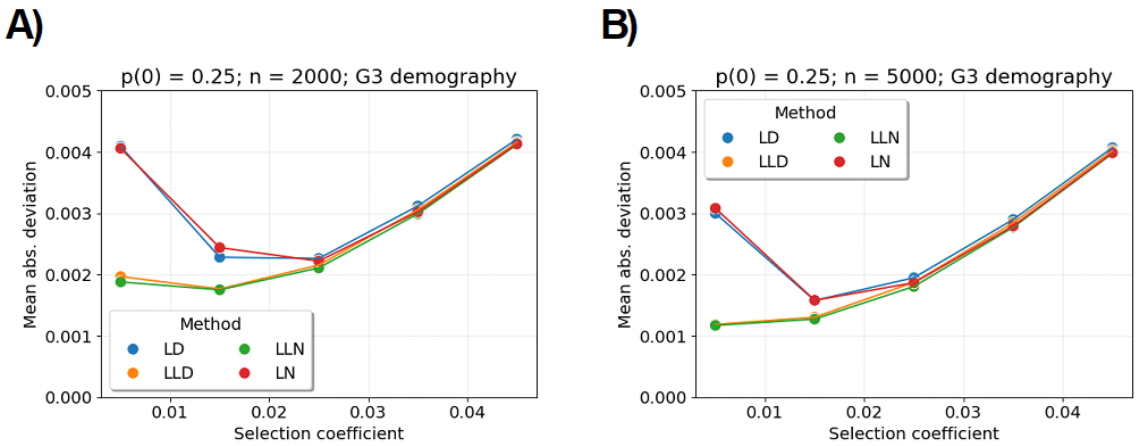

Figure S21: Selection coefficient estimation using the number of detectable IBD segments versus the IBD length distribution and known versus unknown allelic subgroups. Line plots show the mean absolute deviation between estimates and the true selection coefficients $s = 0.005, 0.015, 0.025, 0.035$. The estimation methods are denoted in the legend and assigned colors: length distribution (LD) (blue); length distribution with labeled alleles (LLD) (orange); number of detectable IBD segments with labeled alleles (LLN) (green); and number of detectable IBD segments (LN) (red). A) The sample size is two thousand diploids versus B) the sample size n is five thousand diploids. Three phases of exponential growth (G3) is the demographic model. The sweeping allele frequency $p(0)$ is twenty-five percent. Each data point is based on twenty thousand simulations.

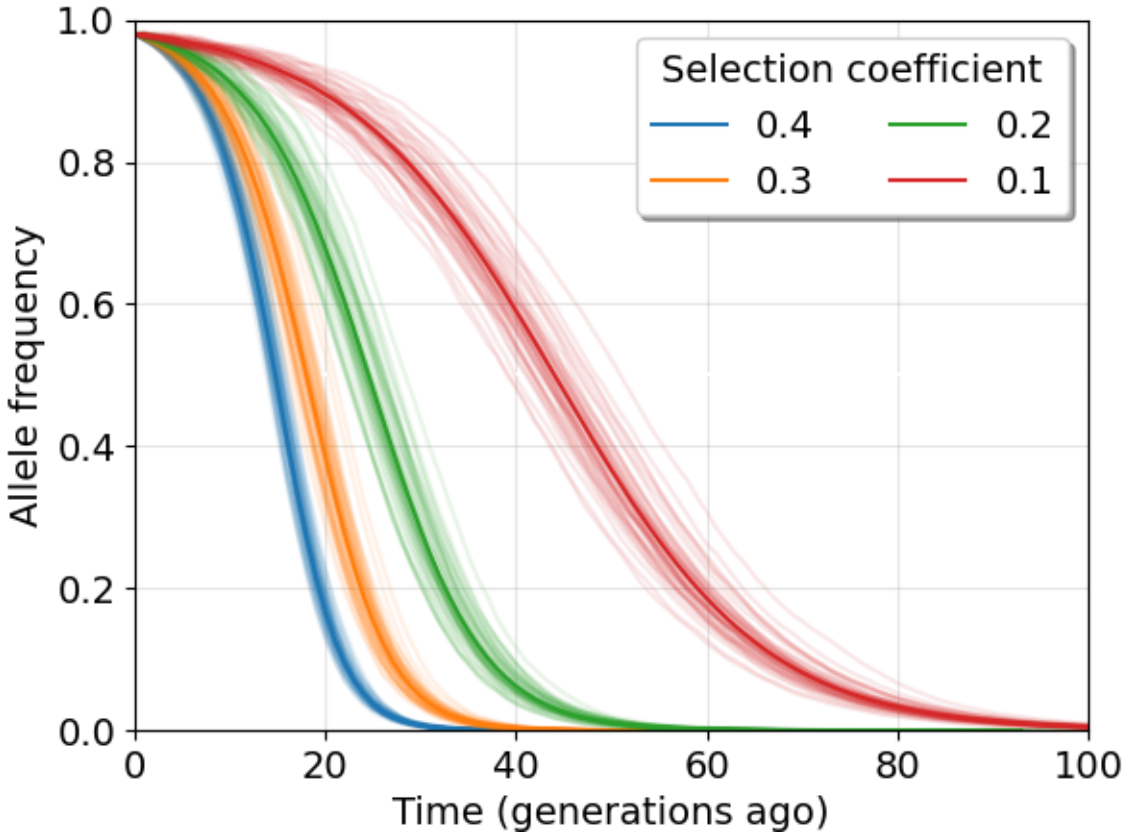

Figure S22: Wright-Fisher simulations with big selection coefficients. Fifty simulations of the Wright-Fisher process for $s \in [0.4, 0.3, 0.2, 0.1]$. Allele frequency ($y$-axis) by time ($x$-axis) is shown for the most recent one hundred generations. The present-day allele frequency is 0.98. The legend defines colors for different selection coefficients. The demography concerns a population decreasing in size exponentially, based on *Plasmodium falciparum* exposed to anti-malaria drugs [65].

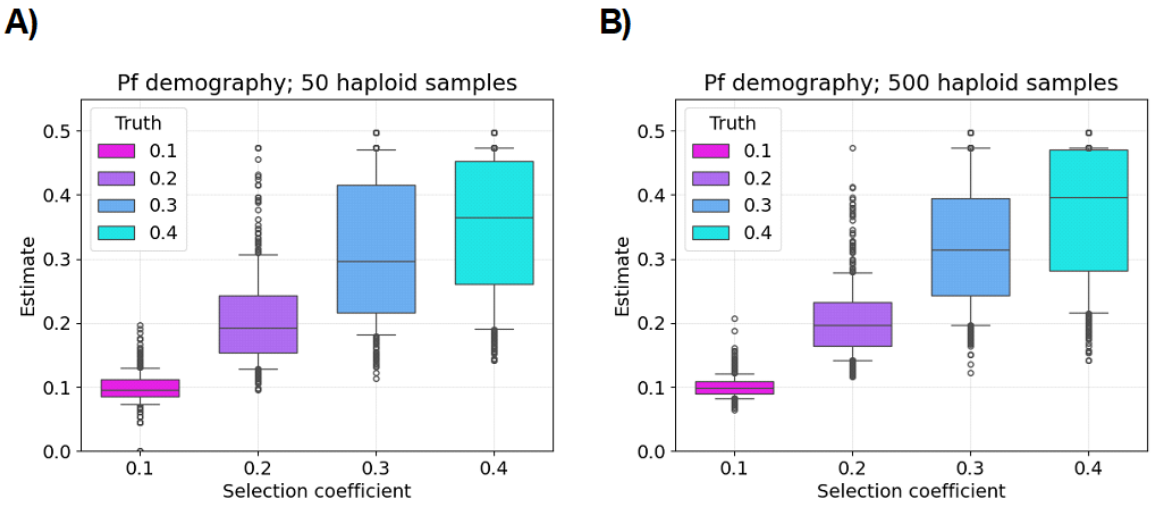

Figure S23: Selection coefficient estimates in *Plasmodium falciparum* model for different sample sizes. Boxplots show the 10th, 25th, 50th, 75th, and 90th percentiles of selection coefficient estimates (y-axis) for $s = 0.10, 0.20, 0.30, 0.40$ (x-axis) with sweeping allele frequency $p(0) = 0.50$ and the example *Plasmodium falciparum* demography. Scatter points denote estimates less than or greater than the 10th and 90th percentiles, respectively. True selection coefficients are assigned different colors (legend). A) There are fifty haploid samples versus B) there are five hundred haploid samples. There are two hundred simulations for each true selection coefficient. The detection threshold is 3.0 cM.

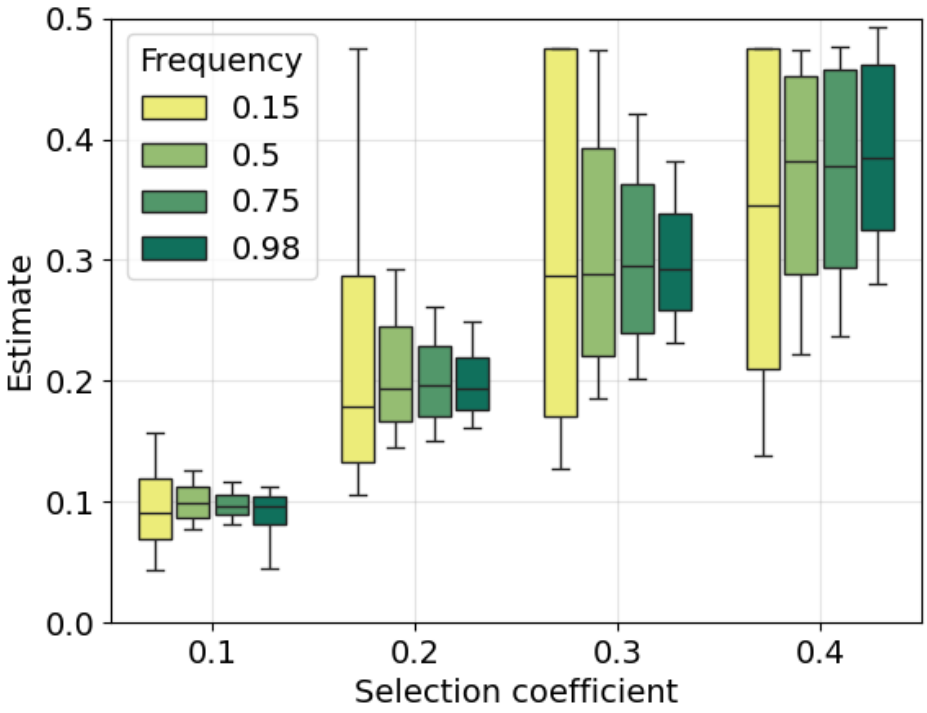

Figure S24: Selection coefficient estimates in *Plasmodium falciparum* model for different sample sizes. Boxplots show the 10th, 25th, 50th, 75th, and 90th percentiles of selection coefficient estimates (y-axis) for $s = 0.10, 0.20, 0.30, 0.40$ (x-axis) with varying sweeping allele frequencies $0.15 \leq p(0) \leq 0.98$ (colors) and the example *Plasmodium falciparum* demography. There are two hundred simulations for each pair of true selection coefficient and sweeping allele frequency. The sample size is fifty haploids. The detection threshold is 3.0 cM.

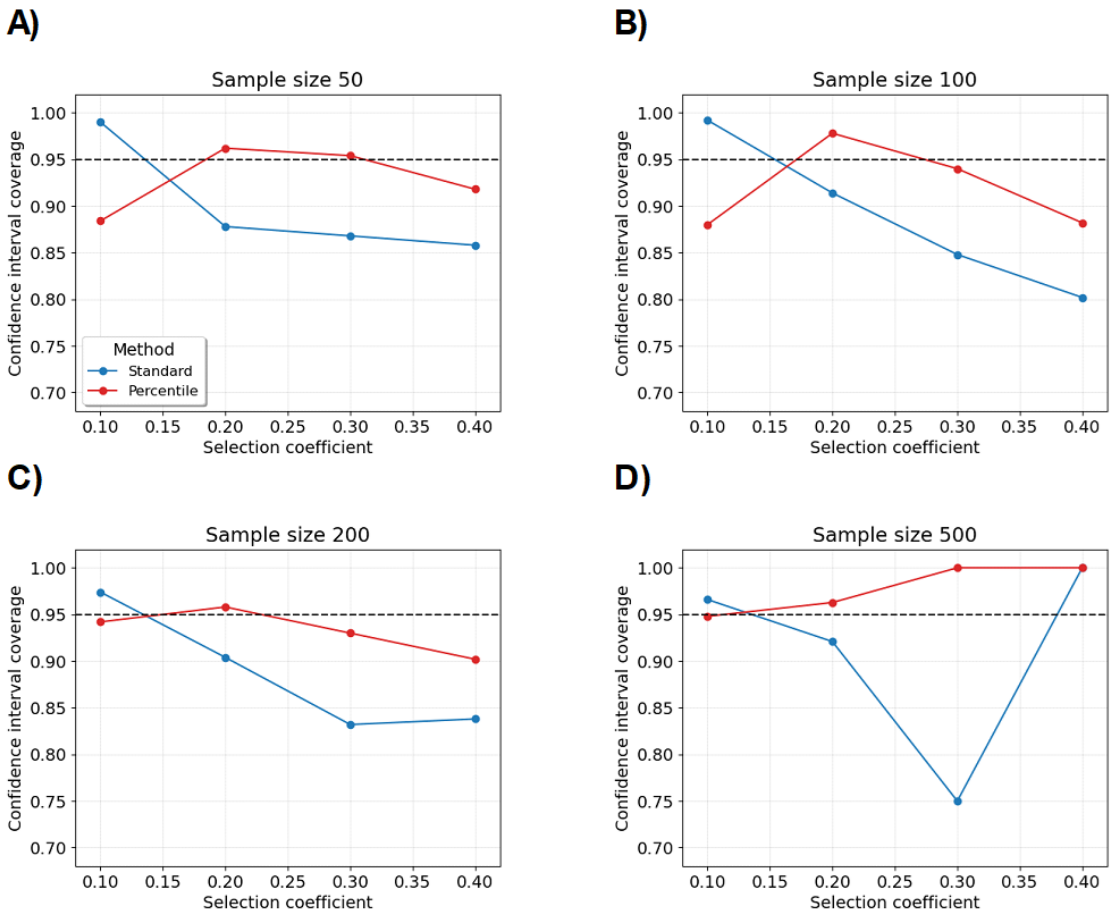

Figure S25: Selection coefficient confidence interval coverages in the Pf model for different sample sizes. Scatter plots show coverage estimates of 95% selection coefficient confidence intervals (y-axis) for s = 0.10, 0.20, 0.30, 0.40 (x-axis) with sweeping allele frequency p(0) = 0.75 and the example Plasmodium falciparum demography. Sample sizes are A) 50, B) 100, C) 200, and D) 500 haploids. The standard normal approach is colored blue and the percentile approach is colored red. Confidence intervals are made from 500 parametric bootstraps. Coverage is calculated from 500 replicate simulations. The detection threshold is 3.0 cM.

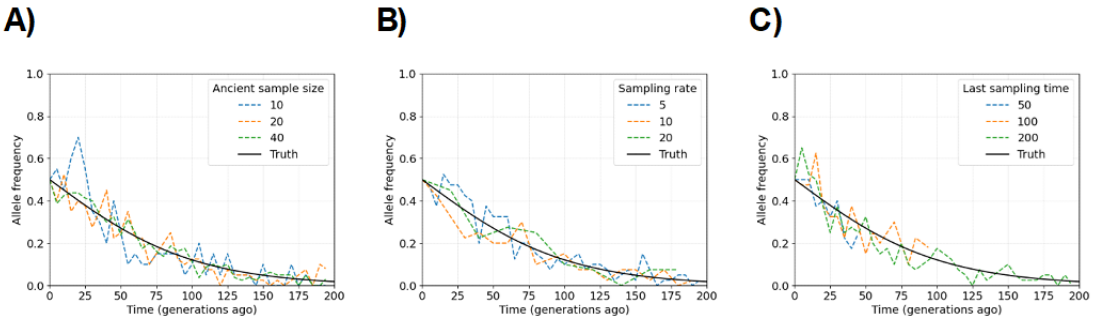

Figure S26: Interpolated Wright-Fisher processes with selection. Line plots show coalescent time (x-axis) by allele frequency (y-axis) for interpolated Wright-Fisher processes with selection. A) Sampling of 10, 20, and 40 diploids is done every five generations until 195 generations ago. B) Sampling of 20 diploids is done every 5, 10, or 20 generations until 195 generations ago. C) Sampling of 20 diploids is done every five generations until 45, 95, or 195 generations ago. The selection coefficient is 0.02. The present-day frequency of the sweeping allele is fifty percent. The sample size in the current generation is one thousand diploids. The population effective size is constant ten thousand diploids. Colors in legends for dashed lines are represented by the parameter perturbations, with labels referring to the parameter under consideration. The black line is the true deterministic curve of the selective sweep model.

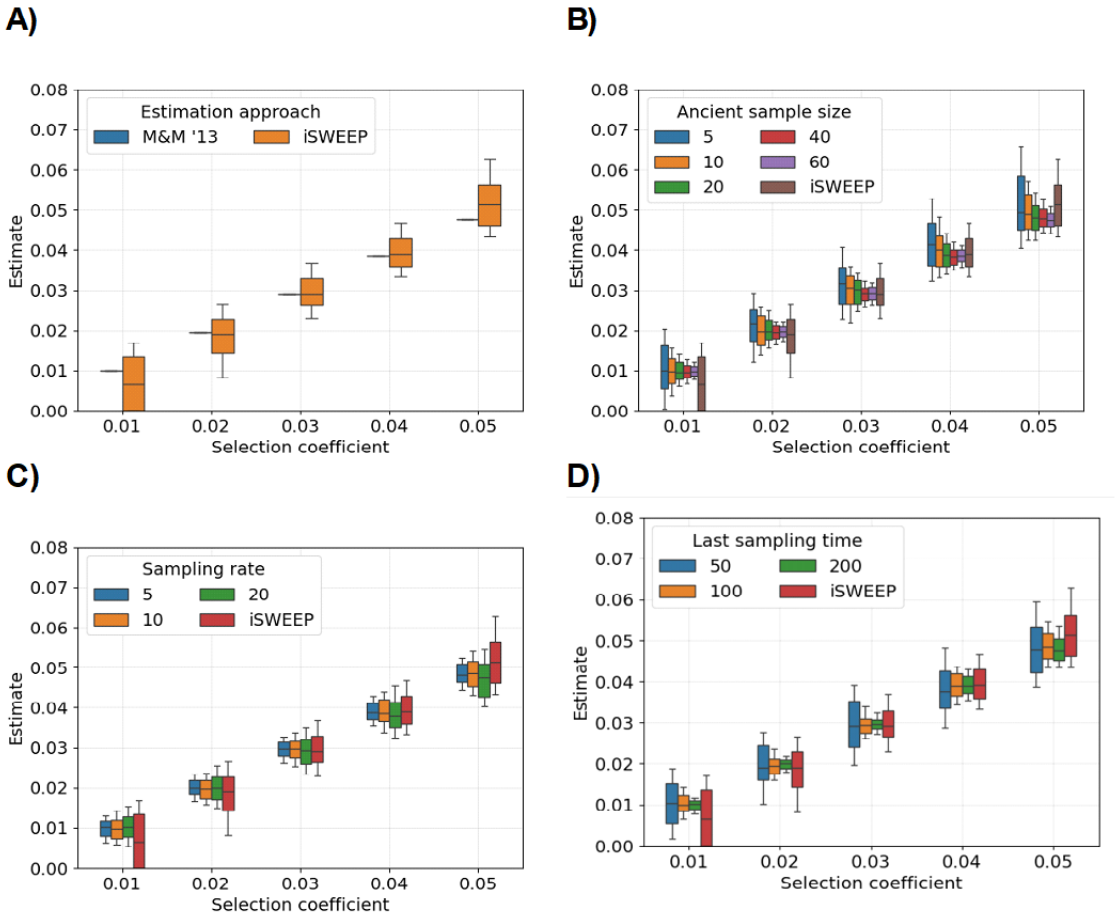

Figure S27: Comparing IBD-based estimates versus the Mathieson and McVean [99] approach in a constant size population. Boxplots represent the medians, interquartile ranges, and tenth and ninetieth percentiles (whiskers) for two hundred estimates of each selection coefficient. Default settings are ancient sample sizes of twenty, the last sampling of ancient data one hundred generations ago, and sampling is done every ten generations. Legends denote in color perturbations from these default settings. A) Mathieson and McVean [99] approach is given the exact and complete historical allele frequencies. B) Ancient sample sizes range from five to sixty at each sampling generation. C) Ancient data is evenly sampled ranging from every five to twenty generations. D) The last time of sampling ancient data ranges from fifty to two hundred generations ago. The IBD-based estimator is called `iSWEEP`.

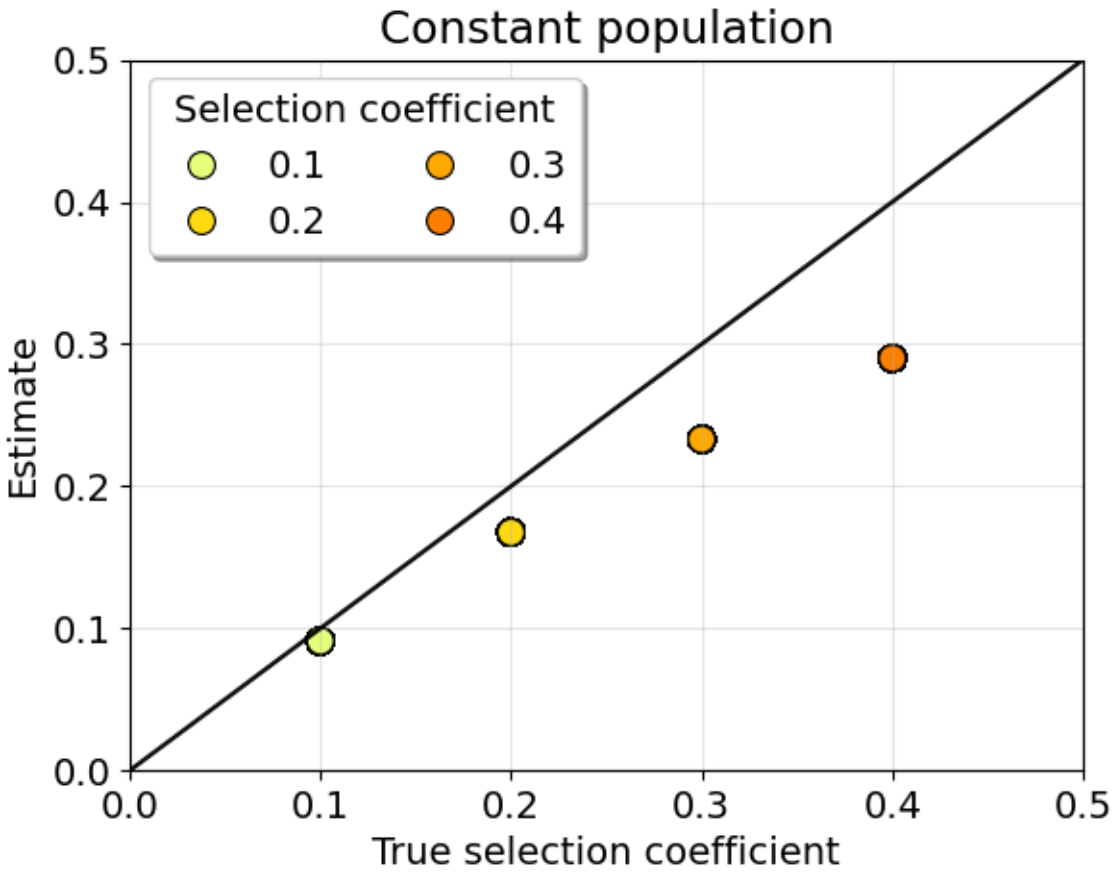

Figure S28: Mathieson and McVean [99] selection coefficient estimator for constant population size and full time series data. The scatter plot shows average selection coefficient estimates from Equation B.3 given a constant population of ten thousand diploids and full time series data. The selection coefficients $s \geq 0.10$ are very large, which violates the first-order Taylor series approximation in Mathieson and McVean [99]. Averages are taken over two hundred simulations.

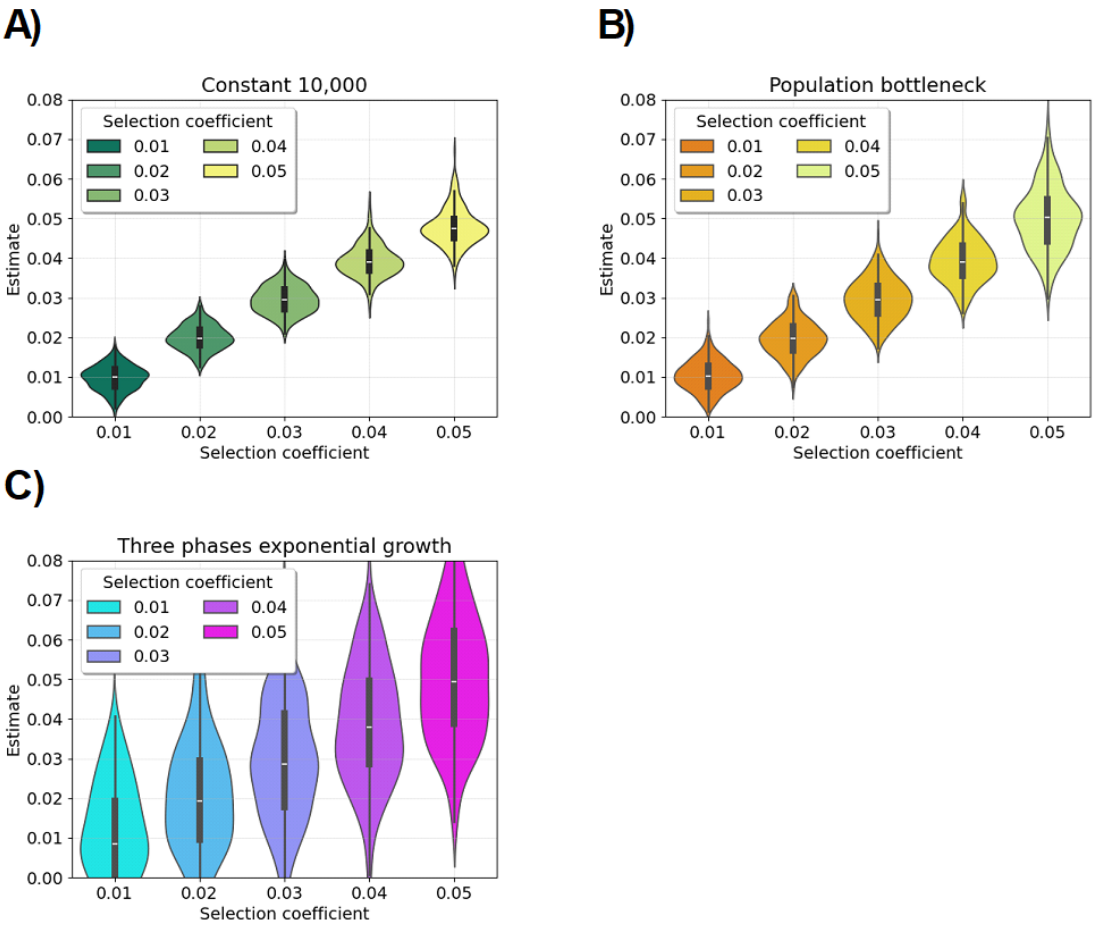

Figure S29: Distributions of modified Mathieson and McVean [99] approach estimates. Kernel density plots and box plots are based on two hundred estimates for each selection coefficient. The population sizes are from A) a constant size population of ten thousand diploids, B) the population bottleneck, or C) the three phases of exponential growth demographic scenario. The Mathieson and McVean [99] approach is an approximate MLE based on historical allele frequencies, and we modify it to handle complex demography and interpolation when allele frequencies are not known in some generations. Default settings are ancient sample sizes of twenty, the last sampling of ancient data one hundred generations ago, and sampling is done every ten generations. Legends denote in color simulations of different selection coefficients. The sweeping allele frequency is fifty percent in the present generation. The sample size is one thousand diploids.

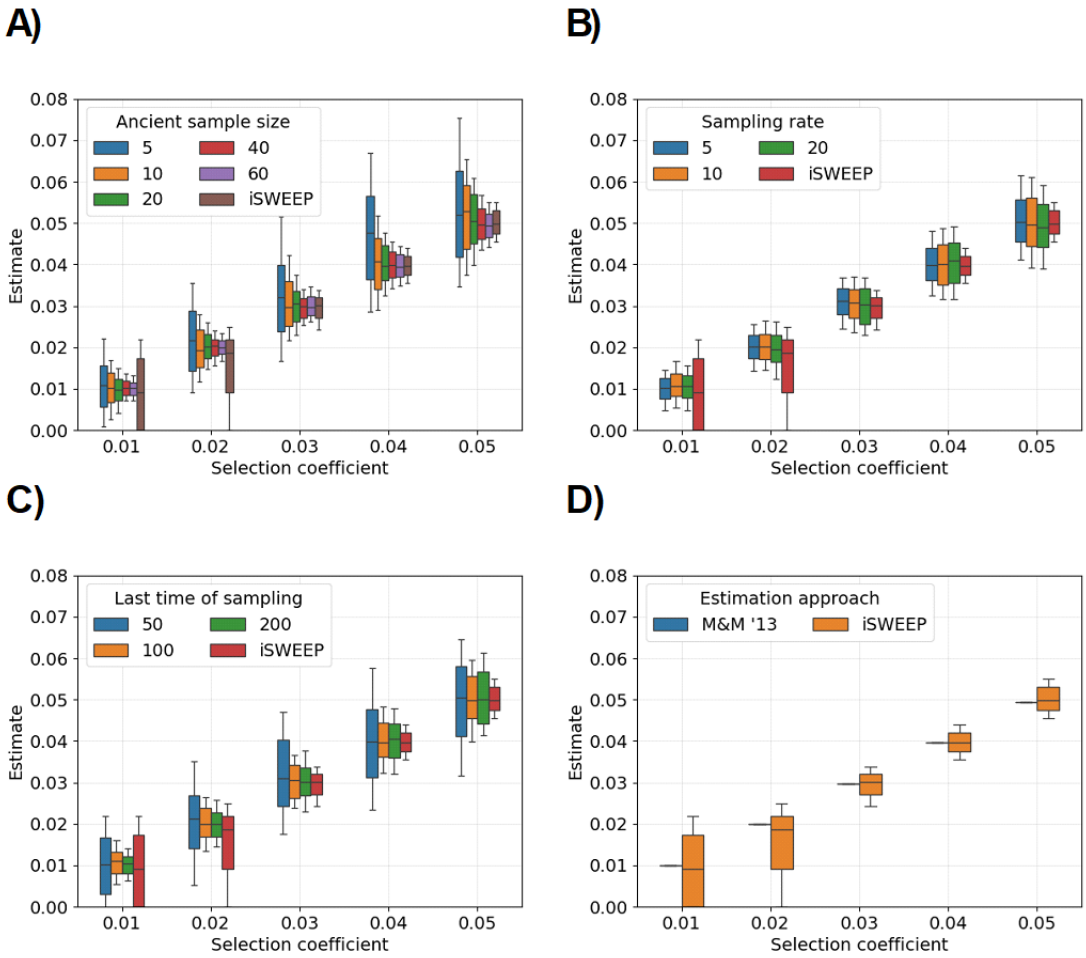

Figure S30: Comparing `iSWEEP` estimates versus modified Mathieson and McVean [99] approach in a population bottleneck demographic scenario. Box plots represent the medians, interquartile ranges, and tenth and ninetieth percentiles (whiskers) for two hundred estimates of each selection coefficient. Default settings are ancient sample sizes of twenty, the last sampling of ancient data one hundred generations ago, and sampling is done every ten generations. Legends denote in color perturbations from these default settings. A) Mathieson and McVean [99] approach is given the exact and complete historical allele frequencies. B) Ancient sample sizes range from five to sixty at each sampling generation. C) Ancient data is evenly sampled ranging from every five to twenty generations. D) The last time of sampling ancient data ranges from fifty to two hundred generations ago. The sweeping allele frequency is fifty percent in the present generation. The sample size in the current generation is one thousand diploids. The IBD-based estimator is called `iSWEEP`.

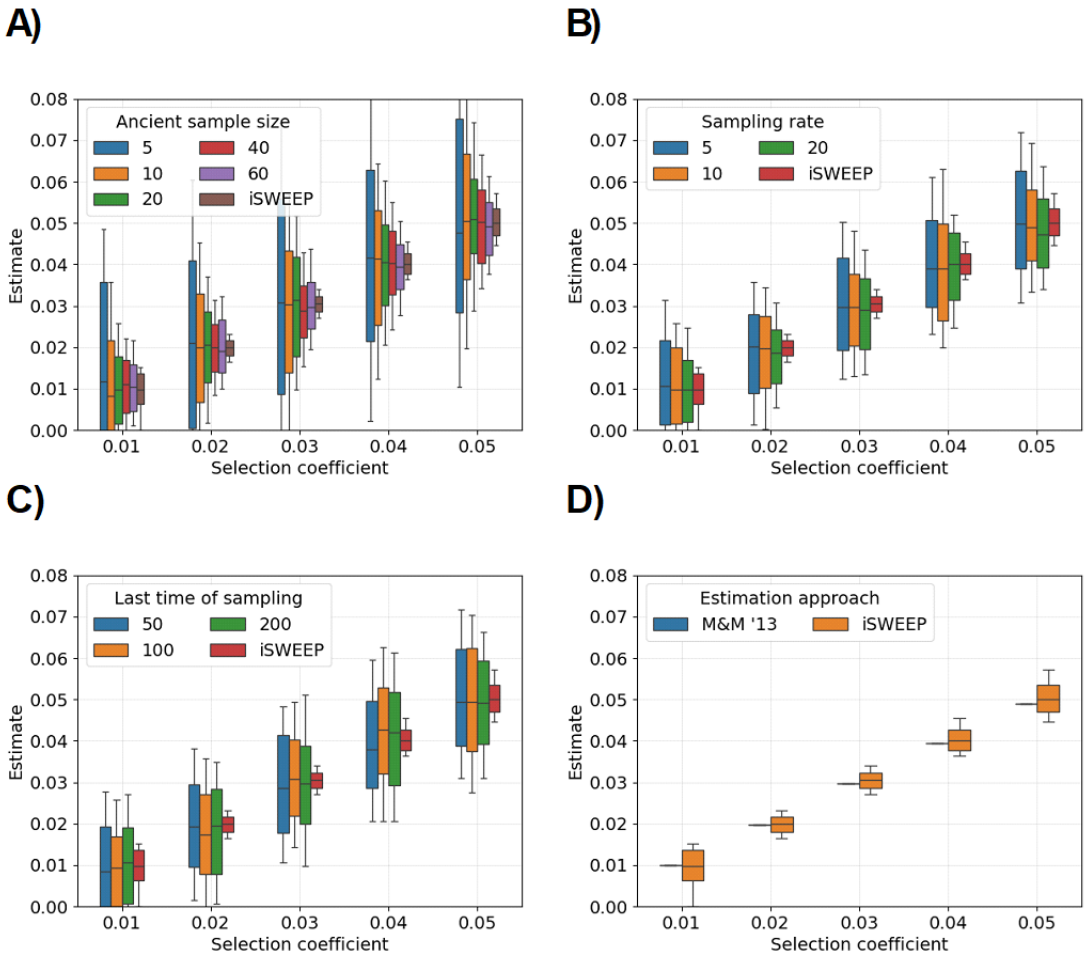

Figure S31: Comparing `iSWEEP` estimates versus modified Mathieson and McVean [99] approach in a three phases of exponential growth demographic scenario. Box plots represent the medians, interquartile ranges, and tenth and ninetieth percentiles (whiskers) for two hundred estimates of each selection coefficient. Default settings are ancient sample sizes of twenty, the last sampling of ancient data one hundred generations ago, and sampling is done every ten generations. Legends denote in color perturbations from these default settings. A) Mathieson and McVean [99] approach is given the exact and complete historical allele frequencies. B) Ancient sample sizes range from five to sixty at each sampling generation. C) Ancient data is evenly sampled ranging from every five to twenty generations. D) The last time of sampling ancient data ranges from fifty to two hundred generations ago. The sweeping allele frequency is fifty percent in the present generation. The sample size in the current generation is one thousand diploids. The IBD-based estimator is called `iSWEEP`.

Methods to study selection in genetic data

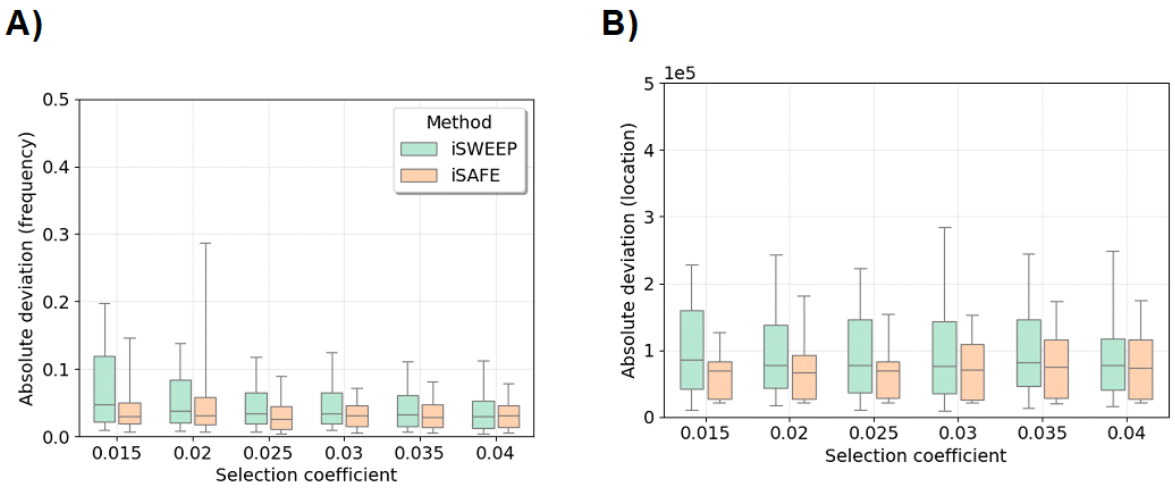

Figure S32: Estimating the frequency and location of the sweeping allele for varying selection coefficients. Comparing our estimator of the sweeping allele frequency and base pair location using `iSWEEP` versus `iSAFE` variant scores in terms of: (A) the absolute deviation between estimates and the true sweeping allele frequency, and (B) the absolute deviation between estimates and the true sweeping allele location. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of estimates for each selection coefficient. There are two hundred simulations of sequence data for each selection coefficient The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

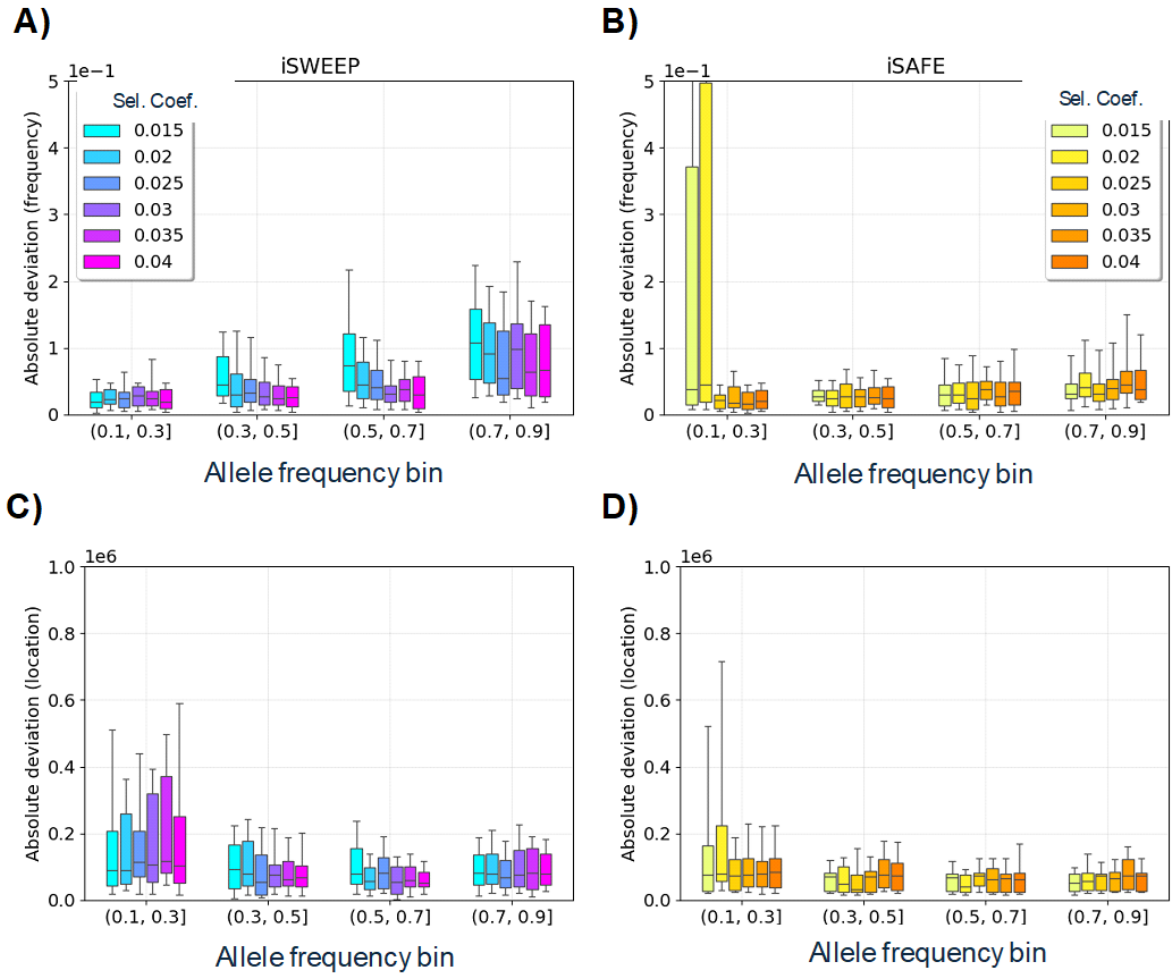

Figure S33: Estimating the frequency and location of the sweeping allele for varying selection coefficients and allele frequencies. Comparing our estimator of the sweeping allele frequency and base pair location using (A,C) `iSWEEP` versus (B,D) `iSAFE` variant scores in terms of: (A,B) the absolute deviation between estimates and the true sweeping allele frequency, and (C,D) the absolute deviation between estimates and the true sweeping allele location. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of estimates for each pair of selection coefficient (colors in legend) and allele frequency bin (x-axis). There are fifty simulations of sequence data for each pair of selection coefficient and allele frequency bin. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

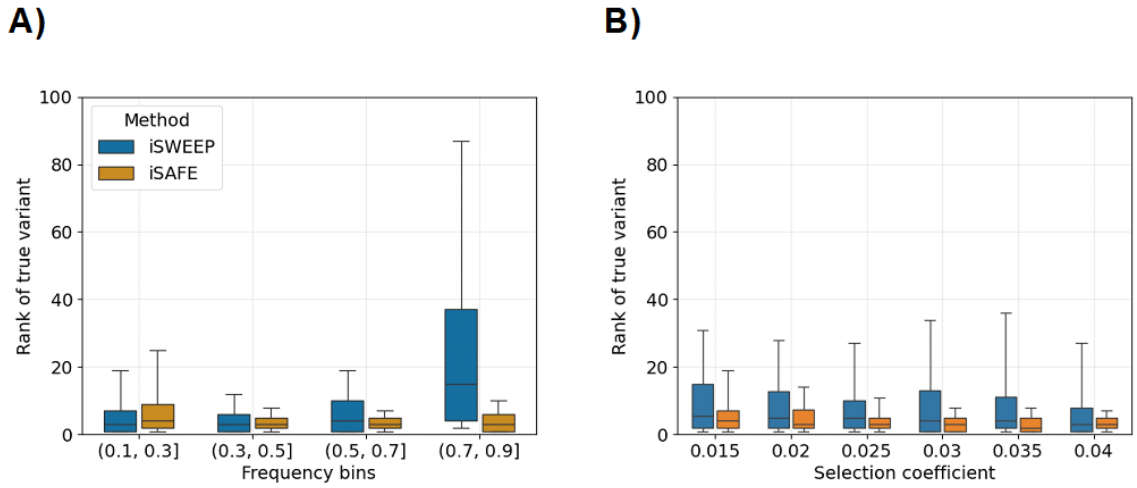

Figure S34: Identifying the sweeping allele in sequence data. Comparing performance between methods `iSWEEP` (blue) and `iSAFE` (orange) in terms of the rank of the true simulated allele, where one is the best possible rank. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of rank estimates for each (A) allele frequency bin and (B) selection coefficient. There are fifty simulations of sequence data for each pair of selection coefficient and allele frequency bin. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

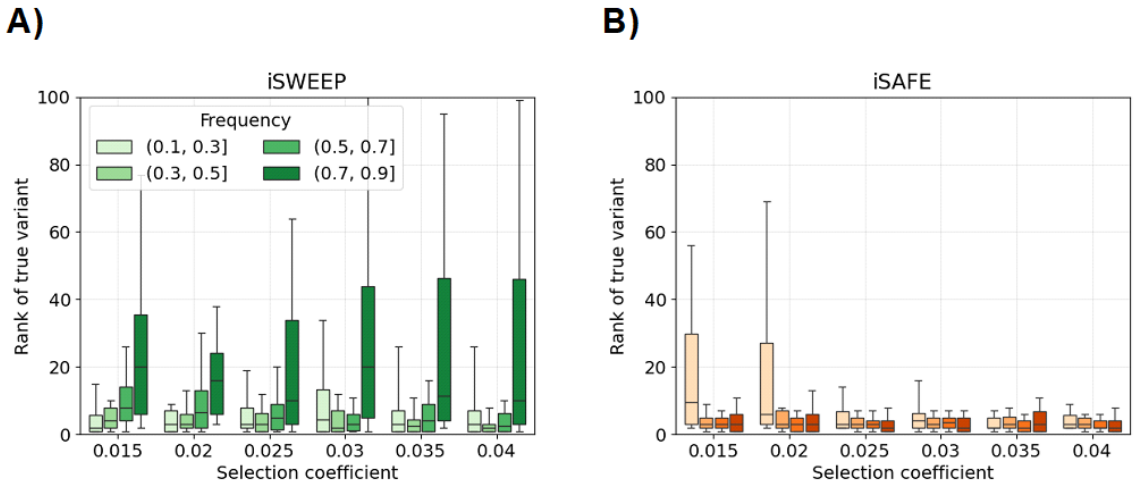

Figure S35: Identifying the sweeping allele in sequence data for varying selection coefficients and allele frequencies. Comparing performance between methods (A) `iSWEEP` and (B) `iSAFE` in terms of the rank of the true simulated allele, where one is the best possible rank. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of rank estimates for each pair of selection coefficient and allele frequency bin. There are fifty simulations of sequence data for each pair of selection coefficient and allele frequency bin. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

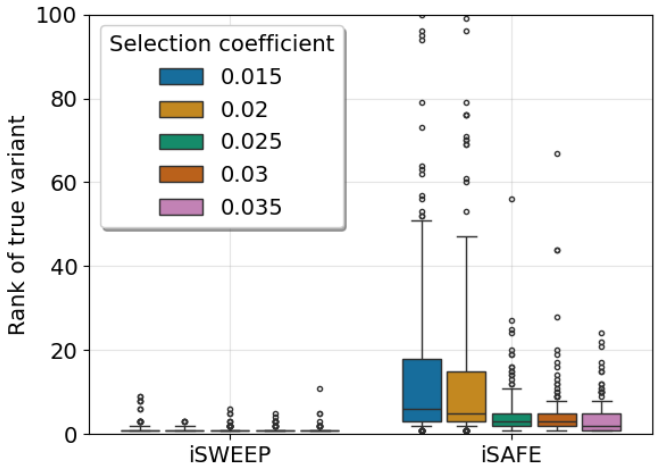

Figure S36: Identifying the sweeping allele in sequence data when `iSWEEP` analysis is centered at the true location. Comparing the ranking methods `iSWEEP` and `iSAFE` when `iSWEEP` is analyzed centered at the true base pair location of the sweeping allele. The best possible rank is one. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of rank estimates for each selection coefficient. There are two hundred simulations of sequence data for each selection coefficient. For `iSWEEP`, the 75th percentile is 1 for all selection coefficients and the 90th percentile is 1 for $s \geq 0.025$. Outliers are data less than the 10th percentile or greater than the 90th percentile. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

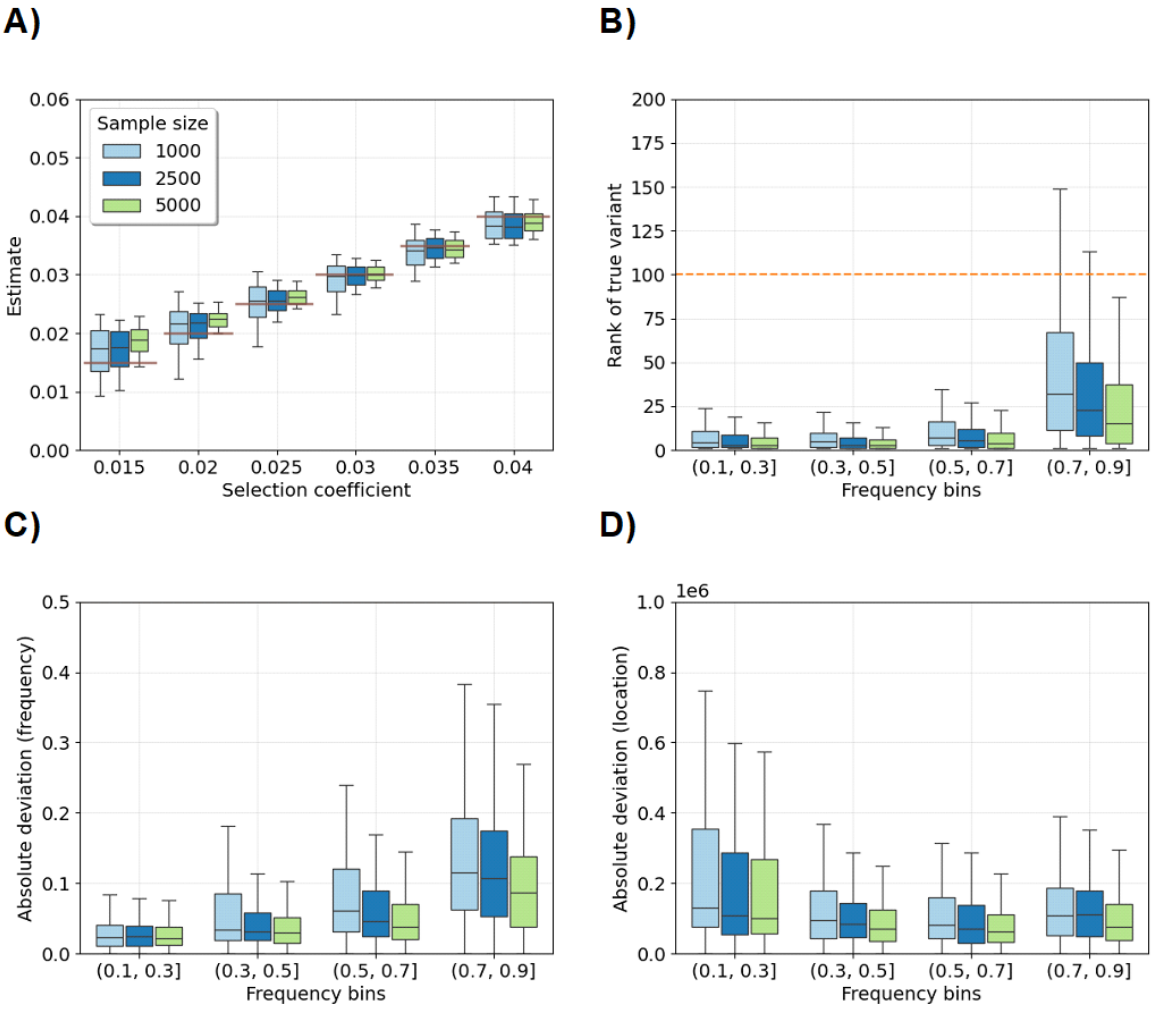

Figure S37: Performance of `iSWEEP` for increasing sample size. Comparing sample sizes 1000 (light blue), 2500 (dark blue), and 5000 (light green) diploids in terms of: (A) estimating selection coefficients, (B) ranking the true simulated allele, (C) estimating the frequency of the sweeping allele, and (D) estimating the base pair location of the sweeping allele. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of estimates from two hundred simulations of each selection coefficient. Horizontal dashed red lines in (A) correspond to the true selection coefficient. The horizontal dashed orange line in (B) is the upper bound in other subplots in Chapter 5. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

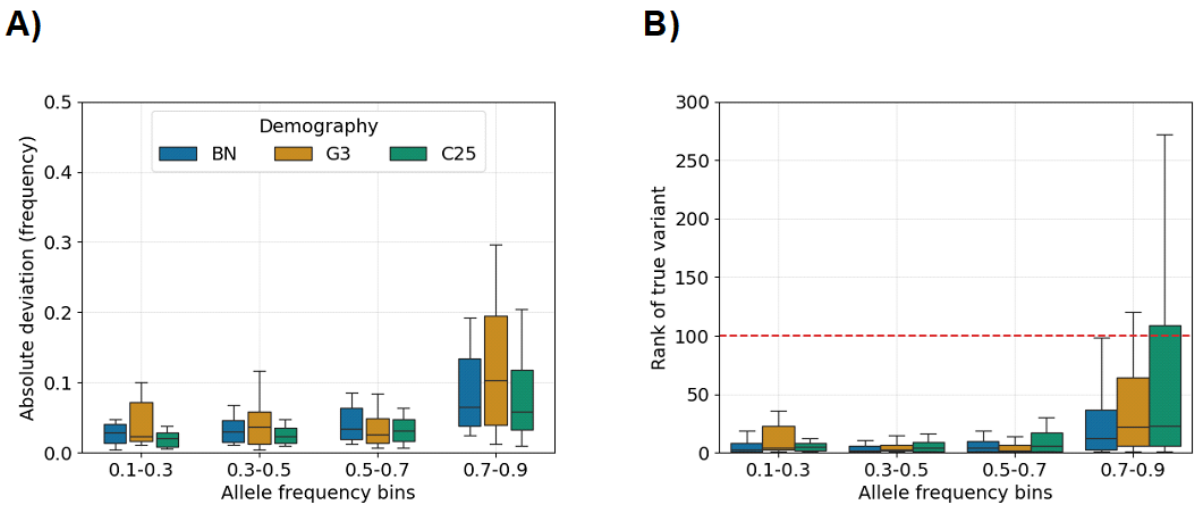

Figure S38: Estimating frequency and identifying the sweeping allele in different demographic scenarios. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of estimates from two hundred simulations of $s = 0.03$ and different demographic models: (A) the absolute deviation between estimates and the true allele frequency; and (B) rankings of the true sweeping allele. Population bottleneck (BN), three phases of exponential growth (G3), and a constant size of 25,000 diploids (C25) are the demographic models (defined in the legend). The horizontal dashed red line in (B) is the upper bound in other subplots in Chapter 5. The sample size is five thousand diploids. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

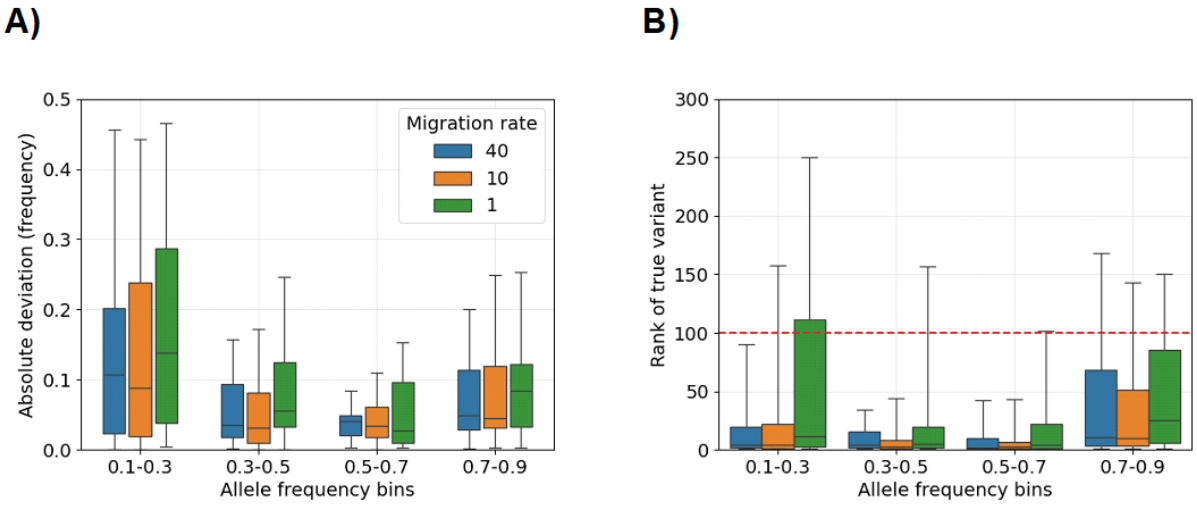

Figure S39: Estimating frequency and identifying the sweeping allele when there is population substructure. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of estimates from two hundred simulations of $s = 0.03$ and different demographic models: (A) the absolute deviation between estimates and the true allele frequency; and (B) rankings of the true sweeping allele. Population bottleneck (BN) is the overall demographic scenario. There are two subpopulations with continuous migration rates of forty, ten, and one percent (defined in legend). The horizontal red line in (B) is the upper bound in other subplots in Chapter 5. The sample size is five thousand diploids. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

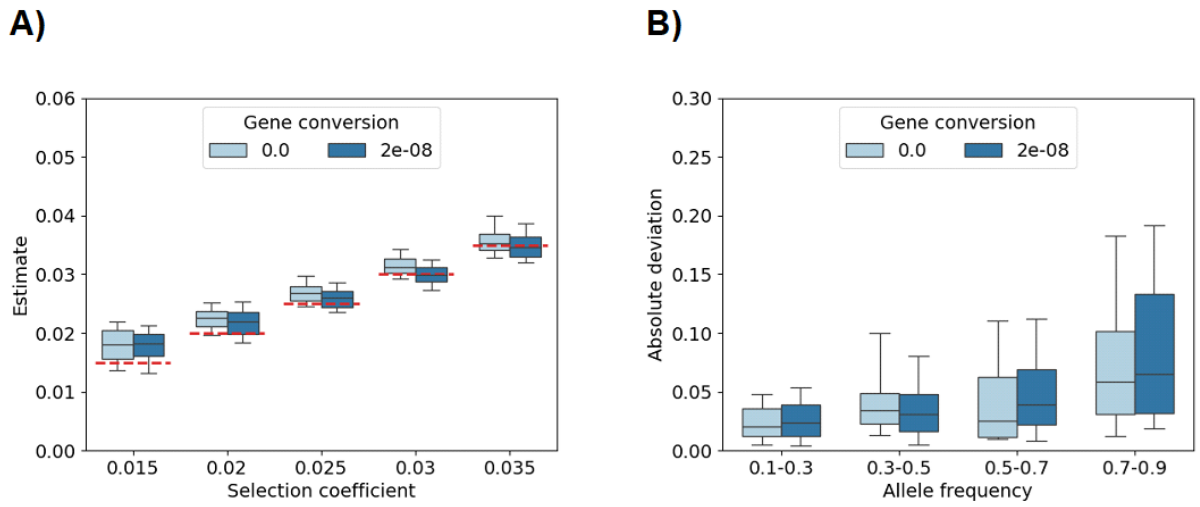

Figure S40: Performance of `iSWEEP` with and without gene conversion. Comparing gene conversion rates of zero (light blue) versus 2e-8 (dark blue): (A) estimating selection coefficients and (B) estimating the frequency of the sweeping allele. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles from sixty simulations of each selection coefficient. Horizontal dashed red lines correspond to the true selection coefficient. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation and recombination rates are 1e-8 and 1e-8.

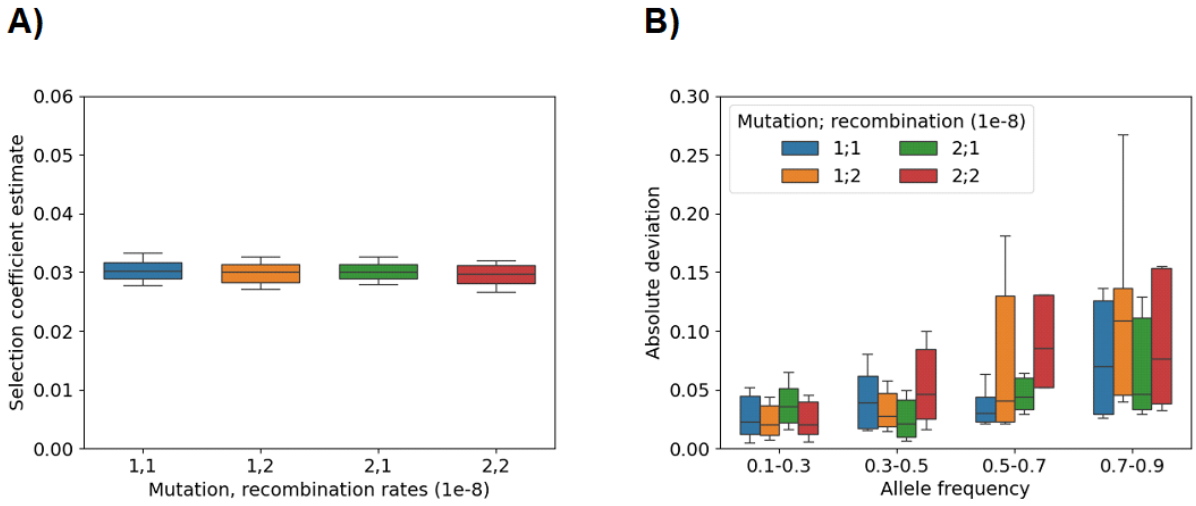

Figure S41: Performance of `iSWEEP` for different mutation and recombination rates. Comparing pair combinations of mutation rate μ and recombination rate ρ: (A) estimating the selection coefficient s=0.03 and (B) estimating the frequency of the sweeping allele. Blue is rates (μ, ρ)= (1e-8,1e-8), orange is rates (1e-8,2e-8), green is rates (2e-8,1e-8), and red is rates (2e-8,2e-8). There are sixty simulations for each setting. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles. The sample size is five thousand diploids. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. The gene conversion rate is 2e-8.

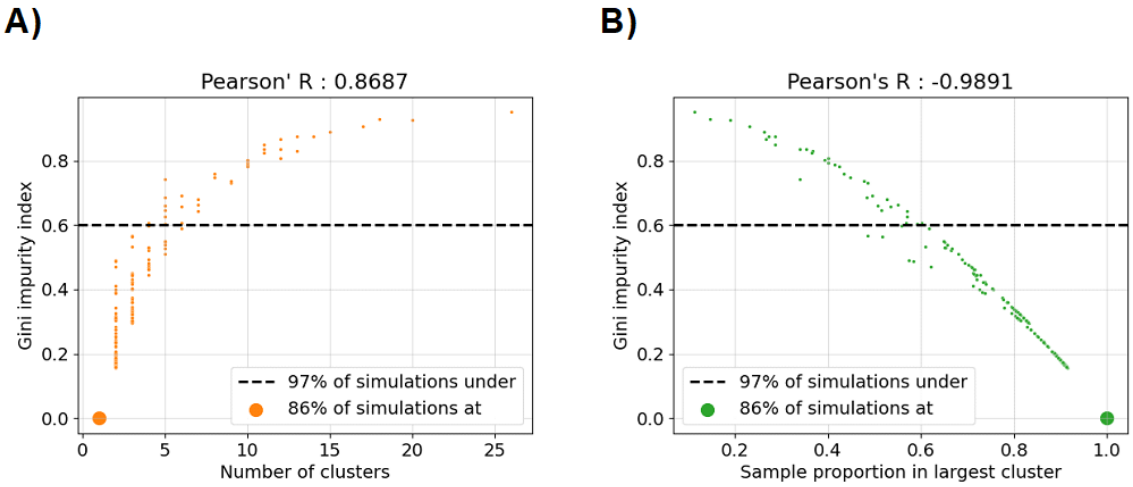

Figure S42: Lack of heterozygosity in excess IBD clusters due to few clusters and one predominant cluster. Gini impurity indices for excess IBD clusters inferred in sequence data (A) increase with the number of clusters and (B) decrease with the proportion of the haplotypes in the outgroup that are in the largest cluster. Results are based on fifty simulations of sequence data for each pair $s \in [0.015, 0.02, 0.025, 0.03, 0.035, 0.04]$ and allele frequency bin in [0.1-0.3, 0.3-0.5, 0.5-0.7, 0.7-0.9]. We determine the 0.6 threshold empirically (horizontal dashed line) as sufficient evidence supporting the recent hard sweep model. Ninety-seven percent of our simulations fall into this category. Eighty-six percent of our simulations involve a single excess IBD cluster. Pearson's correlation in the subplot title is calculated for the Gini impurity index by (A) the number of clusters and (B) the proportion of haplotypes in the outgroup that are in the largest cluster. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

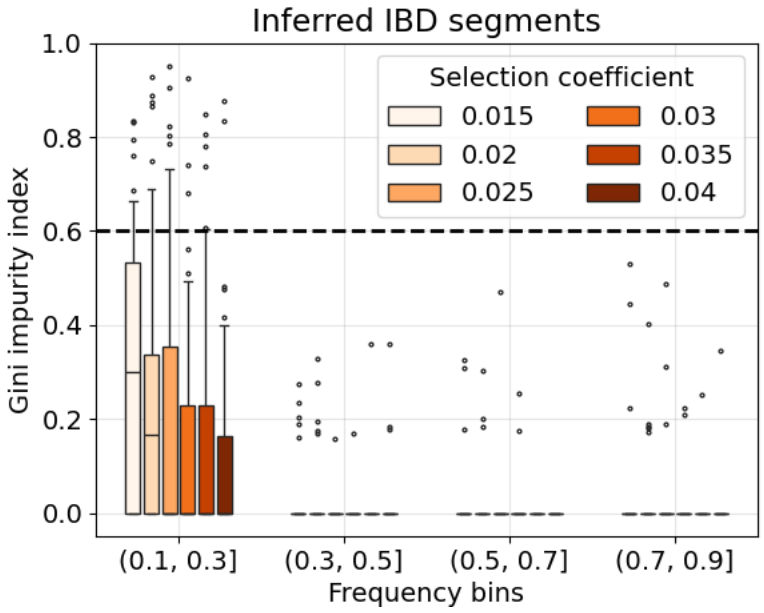

Figure S43: Heterozygosity in excess IBD clusters measured in simulated sequence data. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of Gini impurity indices measured in simulated sequence data. Outliers are data less than the 10th percentile or greater than the 90th percentile. There are fifty simulations of sequence data for each pair of selection coefficient and allele frequency bin. We determine the 0.6 threshold empirically (horizontal dashed line) as sufficient evidence supporting the recent hard sweep model. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

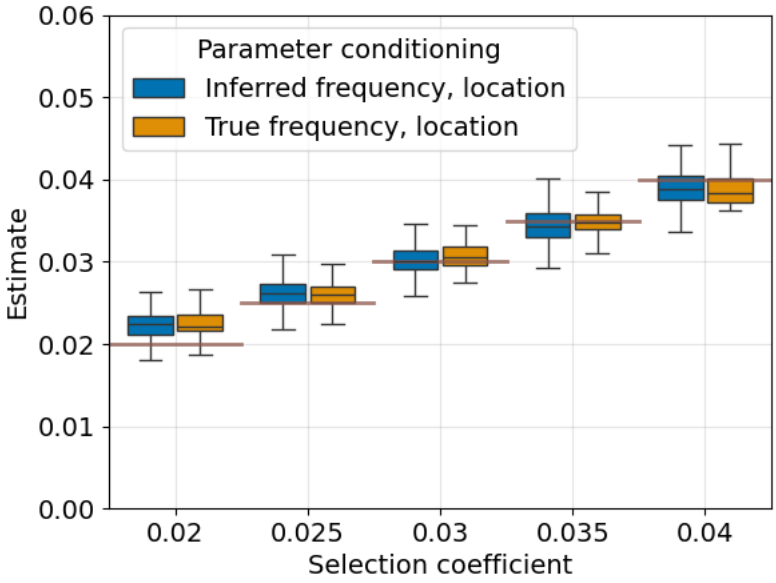

Figure S44: Estimating selection coefficients when the frequency and location of the sweeping allele are unknown. Selection coefficient estimates from `iSWEEP` are based on two hundred simulations of sequence data for each selection coefficient. Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of estimates based on inference conditional on (blue) inferred frequency and base pair location or (orange) true frequency and base pair location. Horizontal light brown lines correspond to the true selection coefficient. The IBD segment detection threshold is 3.0 cM. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

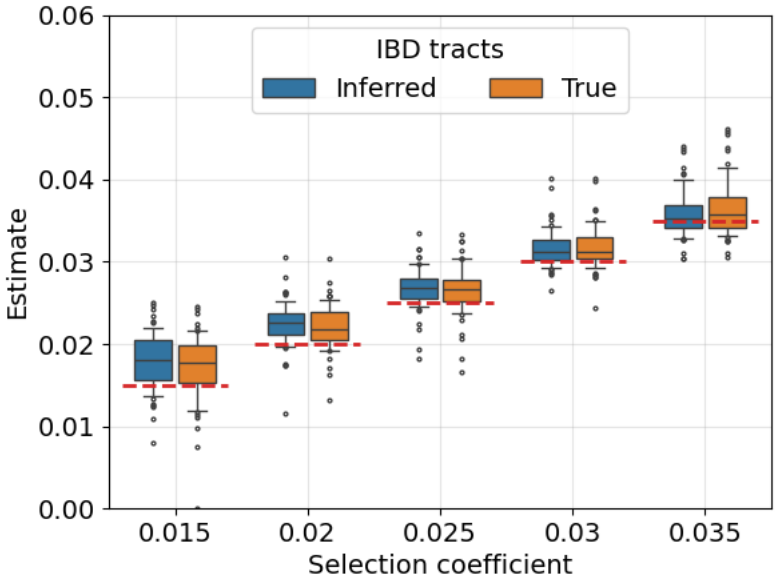

Figure S45: Estimating selection coefficients using true versus inferred IBD segments. Selection coefficient estimates from `iSWEEP` based on simulations of zero gene conversion and either inferred IBD segments from `hap-ibd` and `ibd-ends` (blue) or true IBD segments from `tskibd` (orange). Box plots show 10th, 25th, 50th, 75th, and 90th percentiles of estimates based on sixty simulations for each selection coefficient. Points less than the 10th percentile or greater than the 90th percentile are outliers. Horizontal dashed red lines correspond to the true selection coefficient. The IBD segment detection threshold is 3.0 cM inferred at the true base pair location of the sweeping allele. Estimation is conditional on the true frequency sweeping allele. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation and recombination rates are 1e-8 and 1e-8.

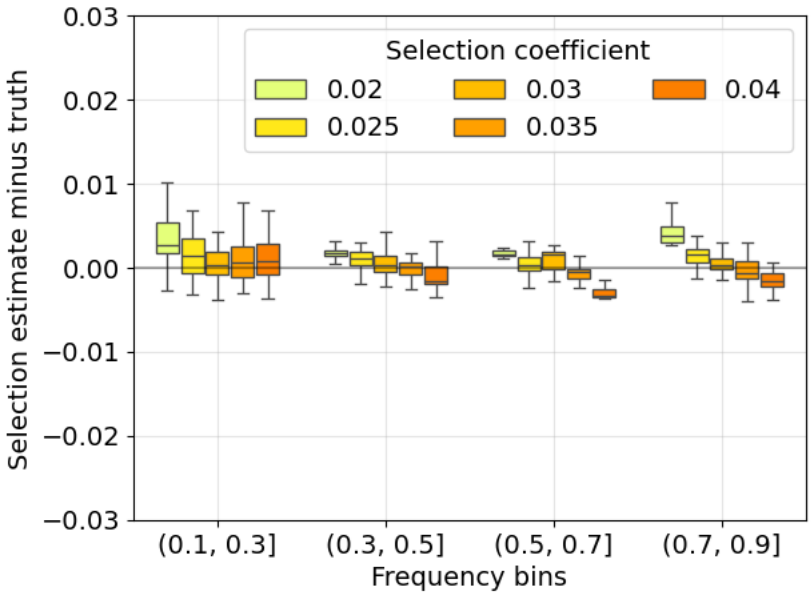

Figure S46: Estimating selection coefficients using inferred IBD segments and varying sweeping allele frequencies. Selection coefficient estimates from `iSWEEP` are aggregated by allele frequency bins and varying selection coefficients. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of estimates minus the truth based on fifty simulations of sequence data for each selection coefficient $s \in [0.01, 0.02, 0.03, 0.04]$ and allele frequency bin in [0.1-0.3, 0.3-0.5, 0.5-0.7, and 0.7-0.9]. The IBD segment detection threshold is 3.0 cM. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation and recombination rates are 1e-8 and 1e-8.

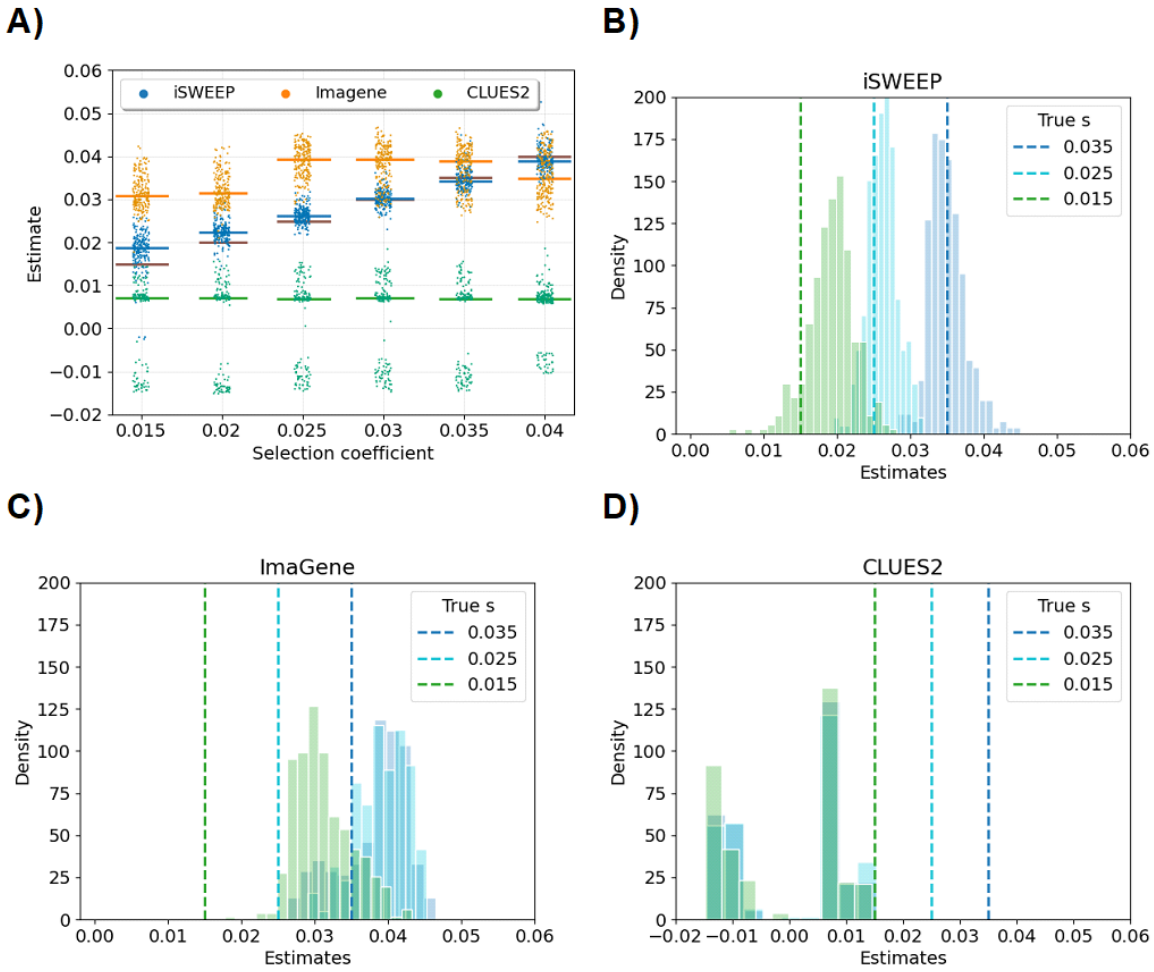

Figure S47: Variability and empirical distribution of selection coefficient estimates for different methods. (A) The strip plot shows the variability in the point estimates for `ImaGene`, `iSWEEP`, and `CLUES2`, where the horizontal lines represent the median estimates (colored) and true selection coefficients (brown). Histogram plots show the empirical distribution of two hundred estimates for selection coefficients $s = 0.015,\ 0.025,\ 0.035$. Each method is shown in a different panel: (B) `iSWEEP`, (C) `ImaGene`, and (D) `CLUES2`. Implementation details are in the main text and Figure 2. Vertical dashed lines are the true selection coefficients. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

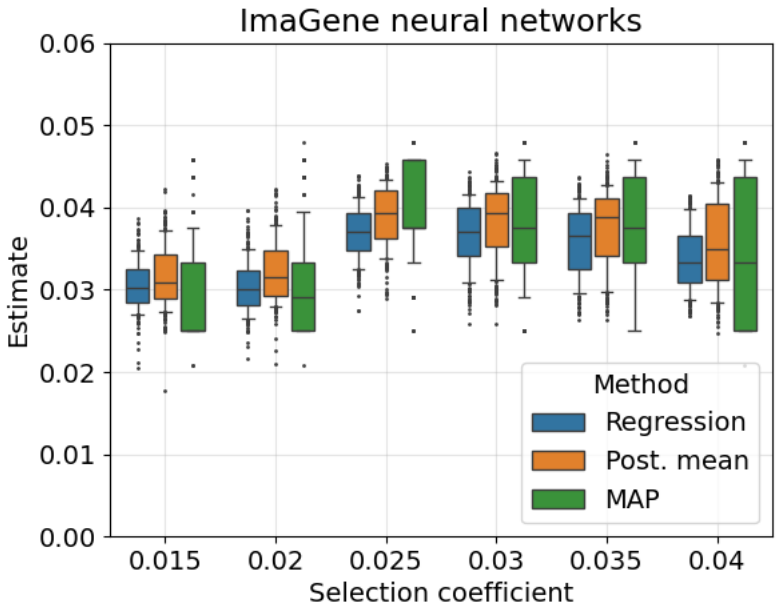

Figure S48: Selection estimates for different `ImaGene` neural network models. Selection coefficient estimates are based on fitted `ImaGene` models using (blue) optimizer=rmsprop, loss=mse, and metrics=mae or (orange and green) optimizer=adam, loss=categorical_crossentropy, and metrics=accuracy. The Posterior mean (orange) and Maximum *a posteriori* (MAP) (green) approaches come from the model with categorical labels. The Regression model (blue) treats the fine-grained categorical labels as continuous responses. The main text describes the model training in detail.

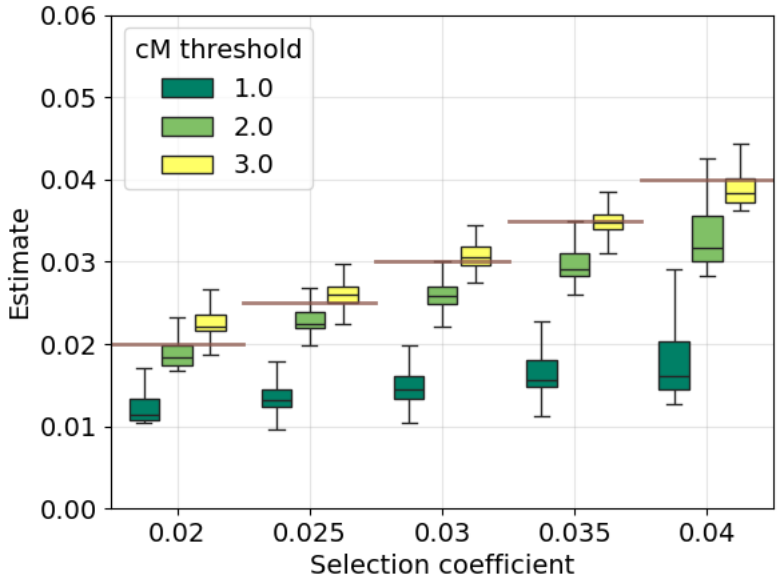

Figure S49: Estimating selection coefficients when inferring IBD segments with different detection thresholds. Selection coefficient estimates from `iSWEEP` are based on two hundred simulations of sequence data for each selection coefficient. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of estimates based on (green) 1.0 cM, (lime) 2.0 cM, and (yellow) 3.0 cM IBD segment detection thresholds. Horizontal light brown lines correspond to the true selection coefficient. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

Scanning for excess identity-by-descent rates

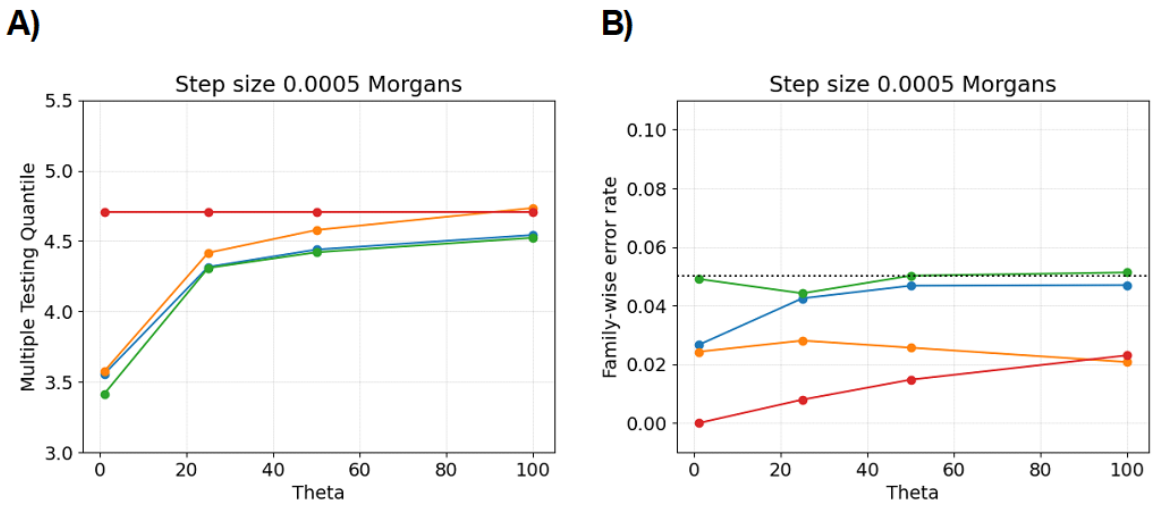

Figure S50: Multiple testing in simulations of Ornstein-Uhlenbeck processes and different hypothesis testing methods. Line plots show A) multiple testing quantiles or B) family-wise error rates (y-axis) of Ornstein-Uhlenbeck processes with different θ (x-axis). Colors represent different hypothesis testing methods: (blue) Siegmund and Yakir discrete approximation, (orange) Siegmund and Yakir continuous approximation, (green) simulation-based approach, and (red) the Bonferroni correction. The simulation-based approach is based on ten thousand simulations. The step size is hypothesis testing every 0.0005 Morgans. The amount of data for each simulation is equivalent to twenty chromosomes of uniform length one Morgan.

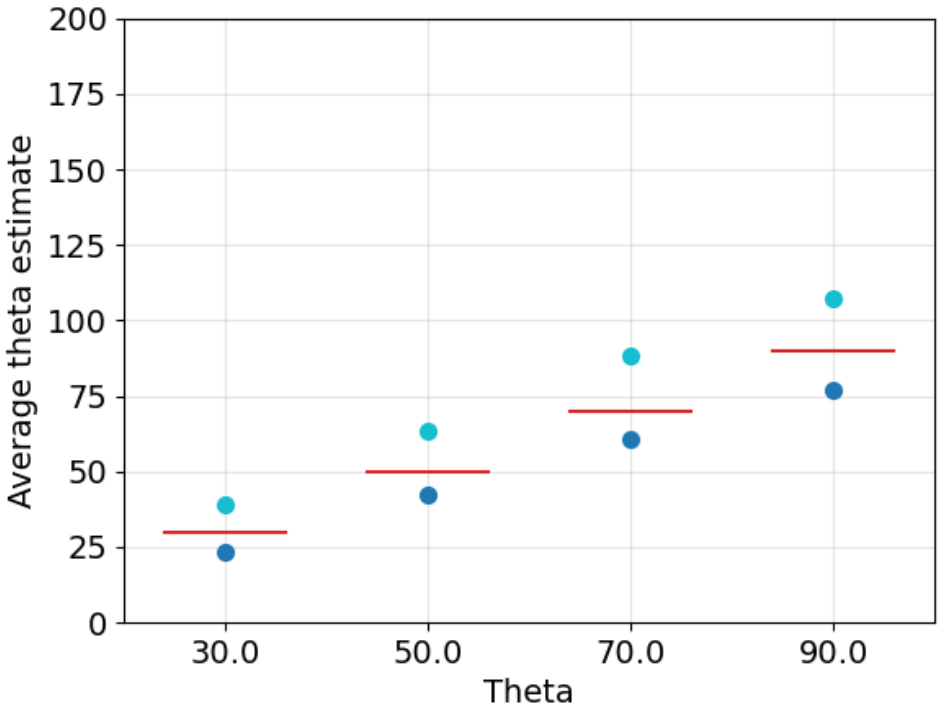

Figure S51: Estimating the exponential decay parameter θ from simulated Ornstein-Uhlenbeck processes. The 10th and 90th percentile of $\hat{\theta}$ estimates (y-axis) of true θ (x-axis) are based on covariance decays up to 4.0 cM and step size 0.02 cM. Percentiles are taken over five hundred simulations for each θ. True θ parameters are also denoted with horizontal red lines. The amount of data for each simulation is equivalent to ten chromosomes of uniform length 100 cM.

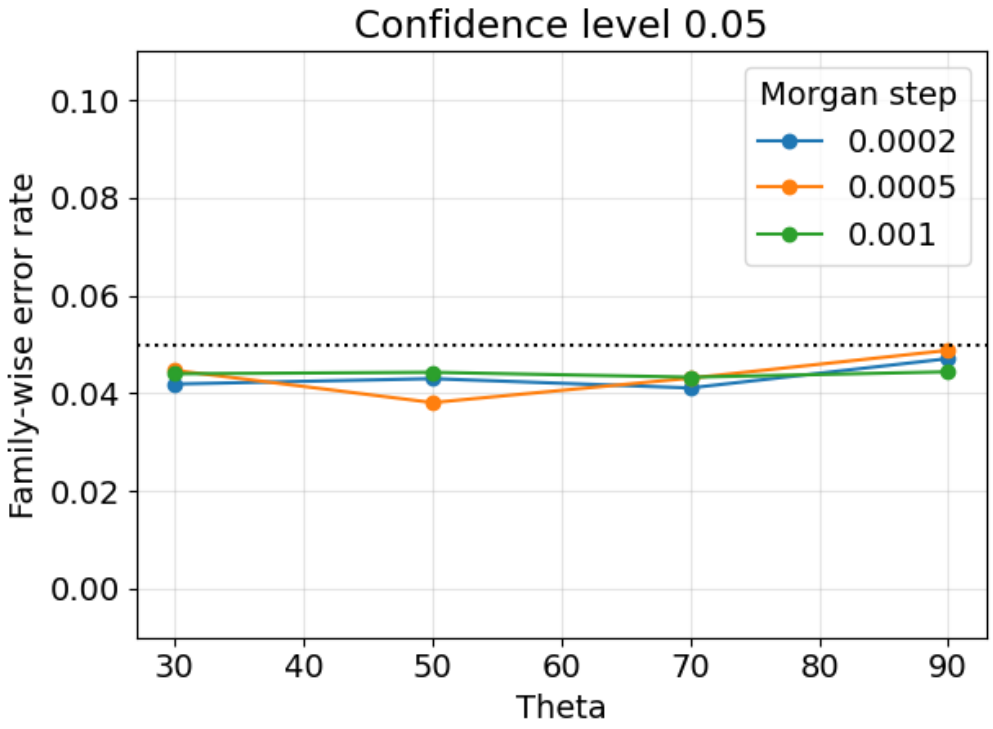

Figure S52: Family-wise error rates for null model simulations of Ornstein-Uhlenbeck processes when exponential decay parameter θ is estimated. Line plots show the family-wise error rates (y-axis) by the true theta parameter (x-axis). Hypothesis tests of family-wise significance level 0.05 are based on the Siegmund and Yakir discrete approximation with estimated θ. Estimates $\hat{\theta}$ are based on covariance decays up to 4.0 cM for different true thetas (x-axis) and different step sizes (legend). Ten thousand Ornstein-Uhlenbeck processes are simulated for each pair of θ and step size (colors in legend). The amount of data for each simulation is equivalent to ten chromosomes of uniform length 100 cM.

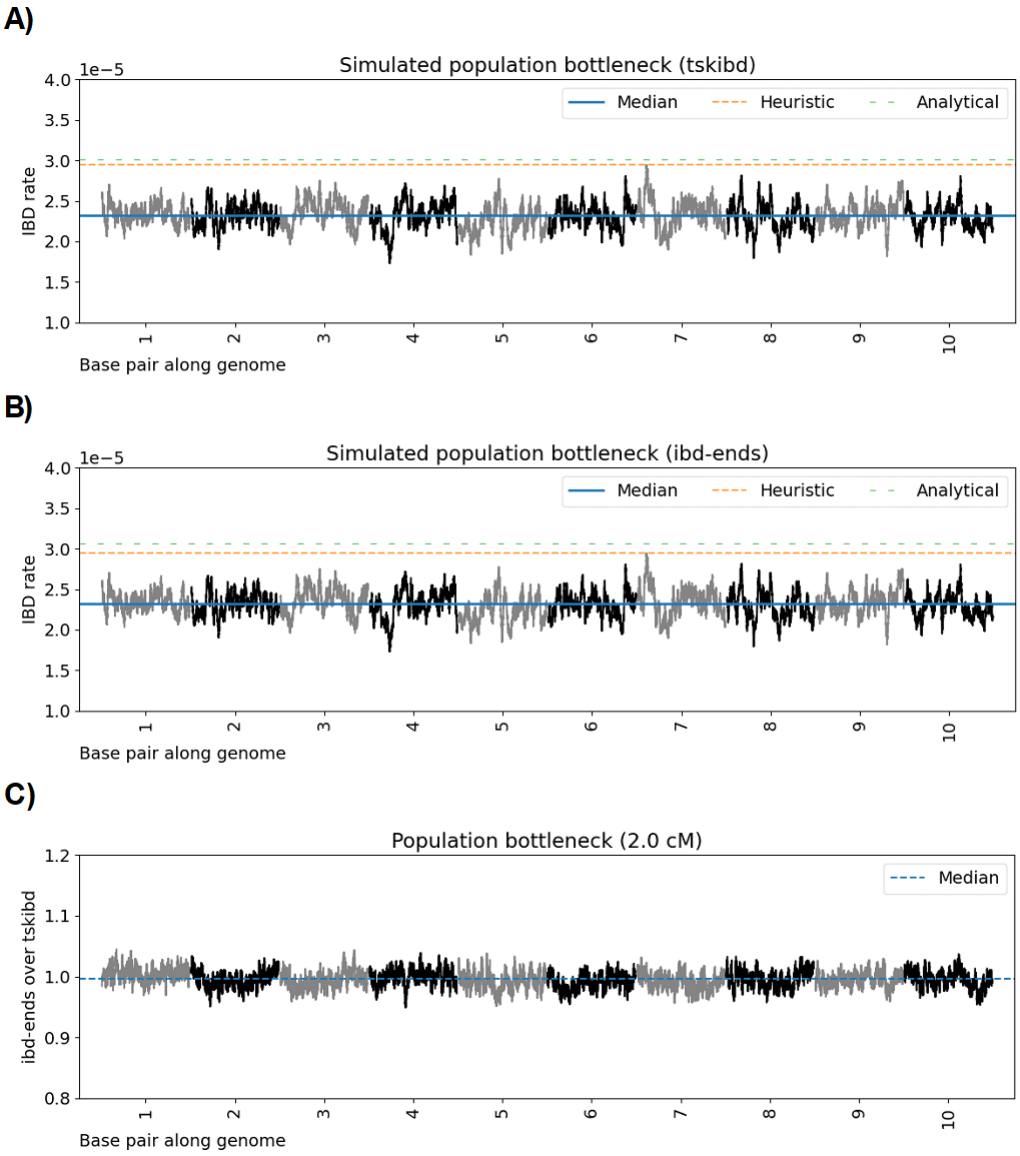

Figure S53: Genome-wide IBD rate scan in simulated population bottleneck data. Line plots show IBD rates $\geq$ 2.0 cM (y-axis) for cM positions along ten simulated chromosomes. Scans are based on A) `tskibd` true IBD segments [65] or B) `ibd-ends` inferred IBD segments [20]. In C), we divide the IBD rates in B) from those in A). Each chromosome is 100 cM. The IBD rate is calculated every 0.02 cM. Data is based on twenty-five hundred diploid samples from the simulated population bottleneck demographic scenario. Horizontal dashed lines show (blue) the genome-wide median IBD rate, (orange) the heuristic threshold of four standard deviations above the median, and (green) the Siegmund and Yakir discrete approximation (S&Y). The family-wise significance level is 0.05.

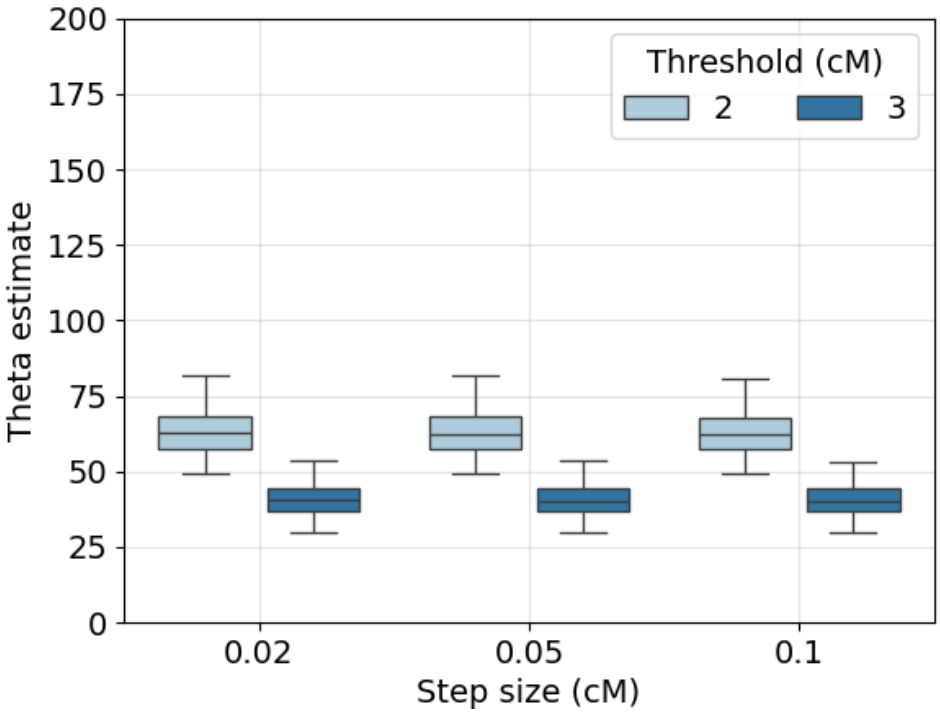

Figure S54: Estimating the exponential decay parameter θ from simulated IBD rate processes with different cM length thresholds. Box plots show the 1st, 25th, 50th, 75th, and 99th, percentiles of estimates $\hat{\theta}$ using the IBD rate processes with simulated true IBD segments (dark blue) $\geq$ 2.0 cM and (light blue) $\geq$ 3.0 cM from `tskibd`. Estimates $\hat{\theta}$ are based on covariance decays up to 4.0 cM for different step sizes (x-axis). There are fifteen hundred simulations for each step size. The demographic model is the population bottleneck. There are twenty-five hundred diploid samples. The amount of data for each simulation is equivalent to ten chromosomes of uniform length 100 cM.

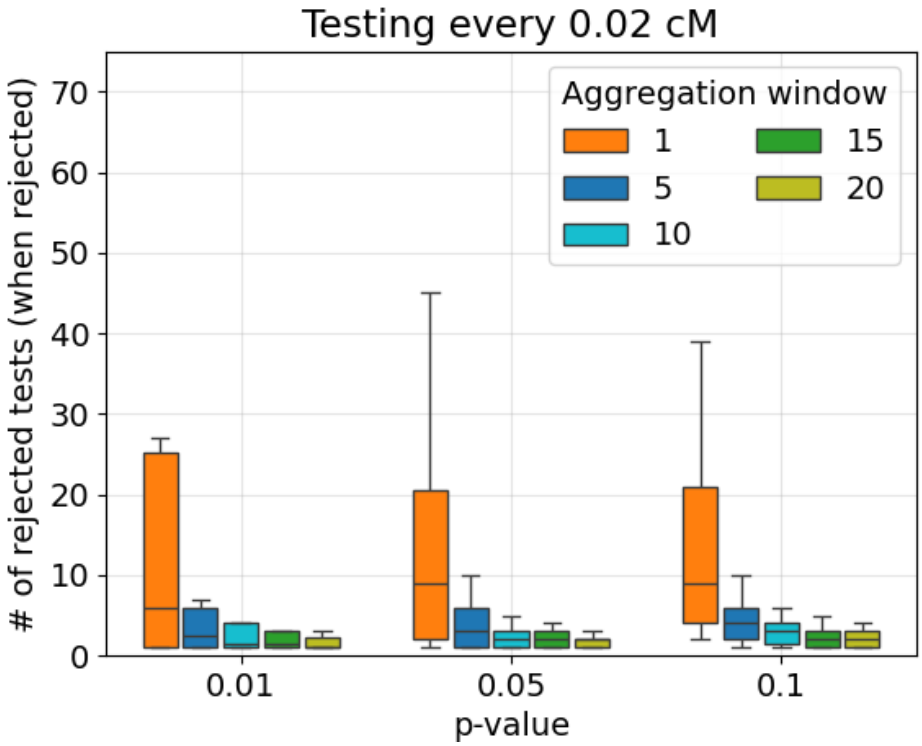

Figure S55: The number of rejected hypothesis tests aggregated by cM position. Box plots show the 10th, 25th, 50th, 75th, and 90th percentiles of the number of rejected tests (y-axis) in non-overlapping windows aggregated over 1, 5, 10, 15, and 20 marginal test statistics (colors in legend). Data where the null model is not rejected are excluded. Using true IBD segments $\geq$ 2.0 cM from simulated IBD rate processes. The hypothesis testing method is the Siegmund and Yakir discrete approximation using true IBD segments $\geq$ 2.0 cM from simulated IBD rate processes. Estimates $\hat{\theta}$ are based on covariance decays up to 4.0 cM. There are five hundred simulations for each family-wise significance level (x-axis). The step size is 0.02 cM. The demographic model is the population bottleneck. There are twenty-five hundred diploid samples. The amount of data for each simulation is equivalent to ten chromosomes of uniform length 100 cM.

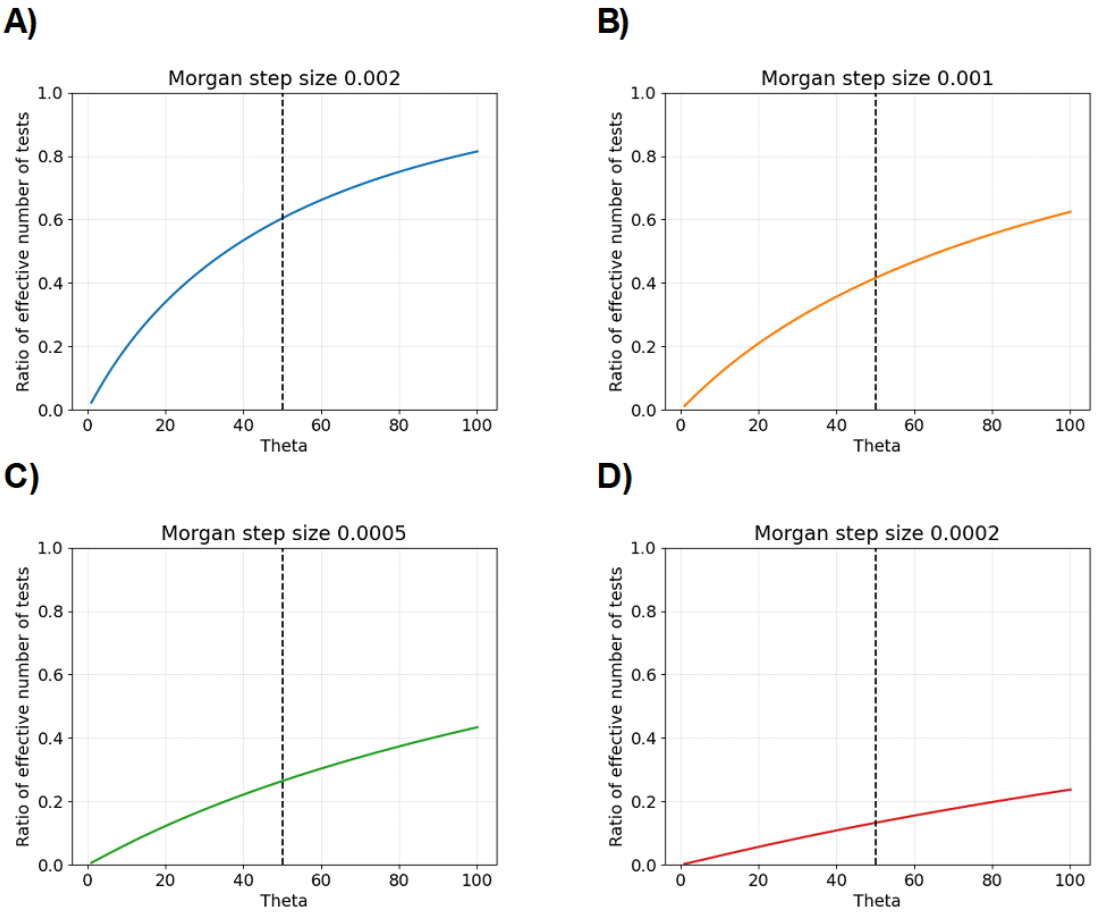

Figure S56: Effective number of tests from Siegmund and Yakir interpretation. Line plots show the interpreted effective number of tests (Appendix A.3) divided by the actual number of tests (y-axis). The x-axis shows the exponential decay parameter θ in the Ornstein-Uhlenbeck process. Vertical black dashed lines denote Theta equal to fifty. The Morgan step sizes for each hypothesis test are A) 0.002, B) 0.001, C) 0.0005, and D) 0.0002.

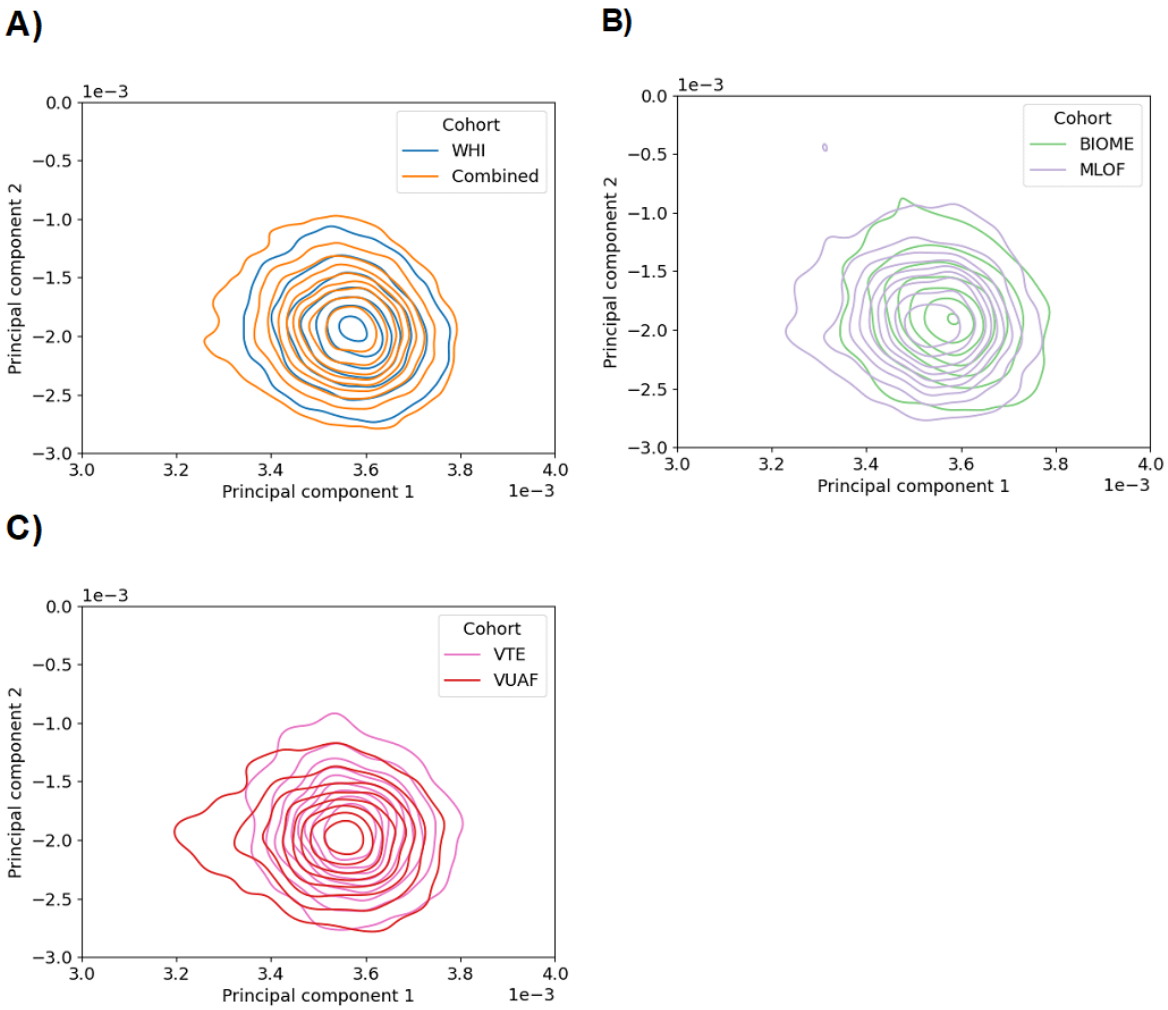

Figure S57: Kernel density plots of principal components with negligible geographic differentiation between European ancestry cohorts. Each plot compares two groups: (A) the EUR1 group and WHI; (B) BioMe and MLOF; and (C) VTE and VUAF.

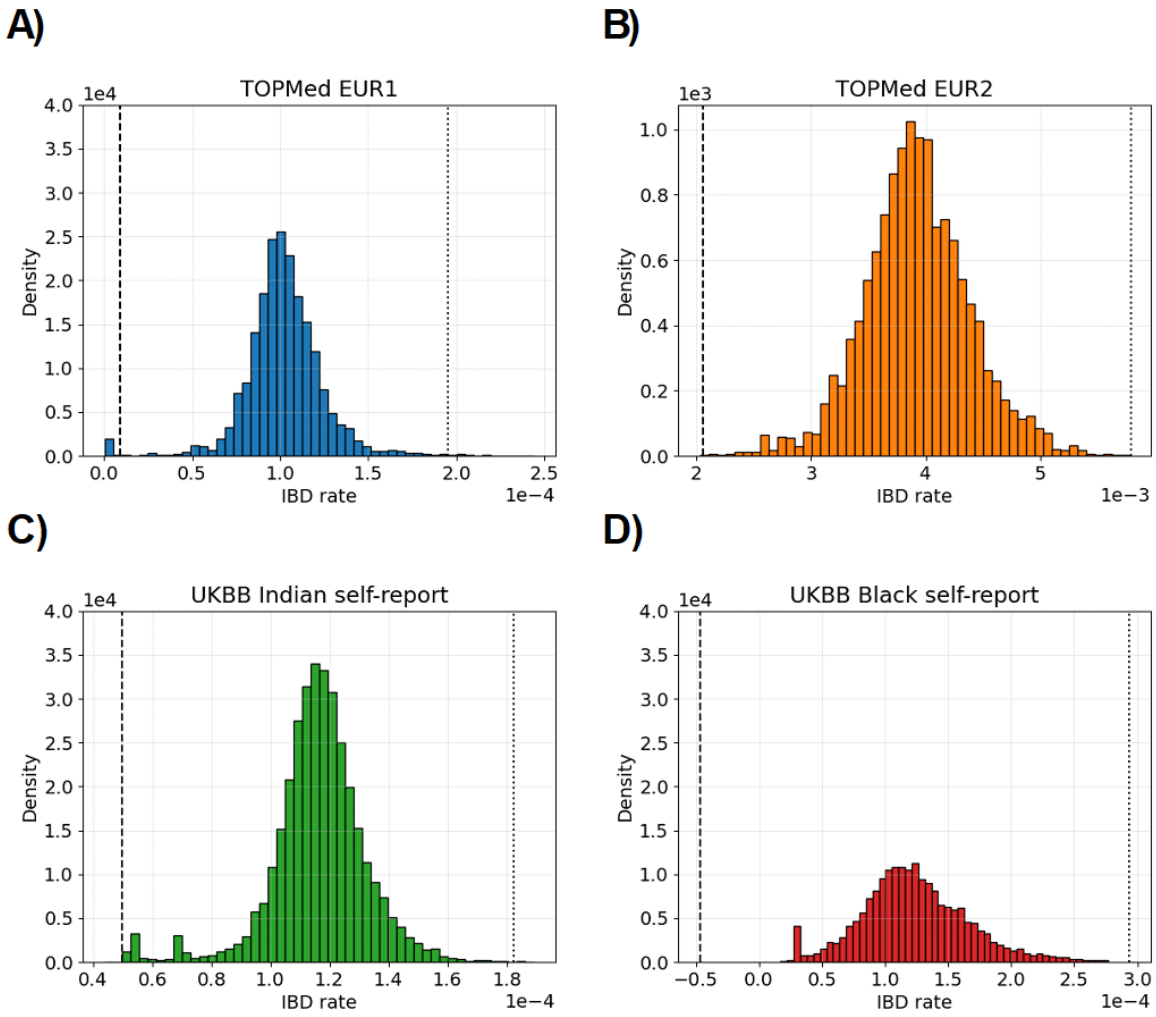

Figure S58: Histograms of IBD rates around a locus in human populations. The inferred IBD rates $\geq$ 2.0 cM (x-axis) are shown for A) TOPMed EUR1, B) TOPMed EUR2, C) UKBB self-report Indian, and D) UKBB self-report Black groups. The y-axis scale is the same in A), C), and D) and different in B) for visibility. There are fifty bins in all plots. The vertical dashed and dotted lines represent the mean $\pm$ four standard deviations after removing outliers (described in the main text).

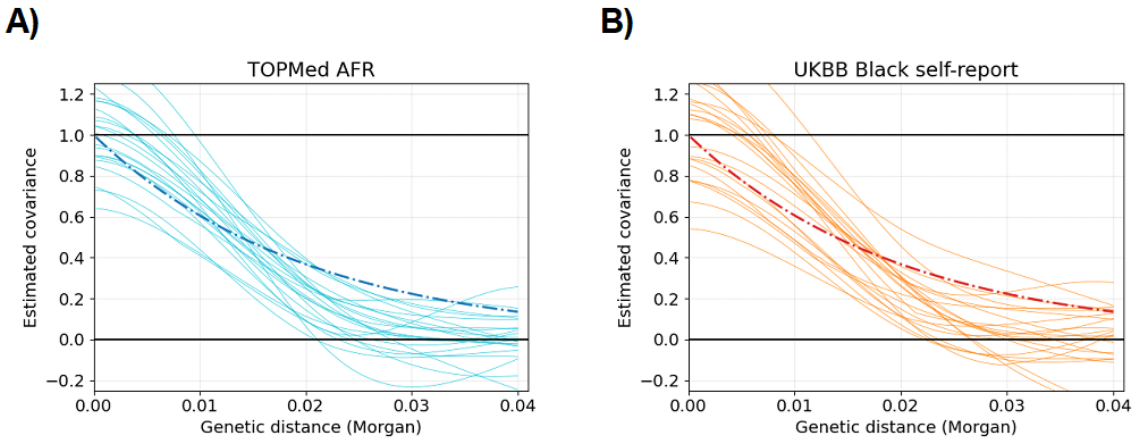

Figure S59: Estimating exponential decay parameter θ in African ancestry data. Each faint colored line shows estimated covariances (y-axis) for different cM distances (x-axis) and a specific chromosome. Dark-colored dashed lines show the predicted covariances from estimates $\hat{\theta}$ in fitted Ornstein-Uhlenbeck processes. There are twenty-two chromosomes in human autosomal data. Centimorgan distances in the TOPMed data A) come from the deCODE 2019 GRCh38 genetic map. Centimorgan distances in the UKBB data B) from the Bherer GRCh37 genetic map. Inferred IBD segments from `hap-ibd` and `ibd-ends` (real data) $\geq$ 2.0 cM. Step sizes are 0.02 cM.

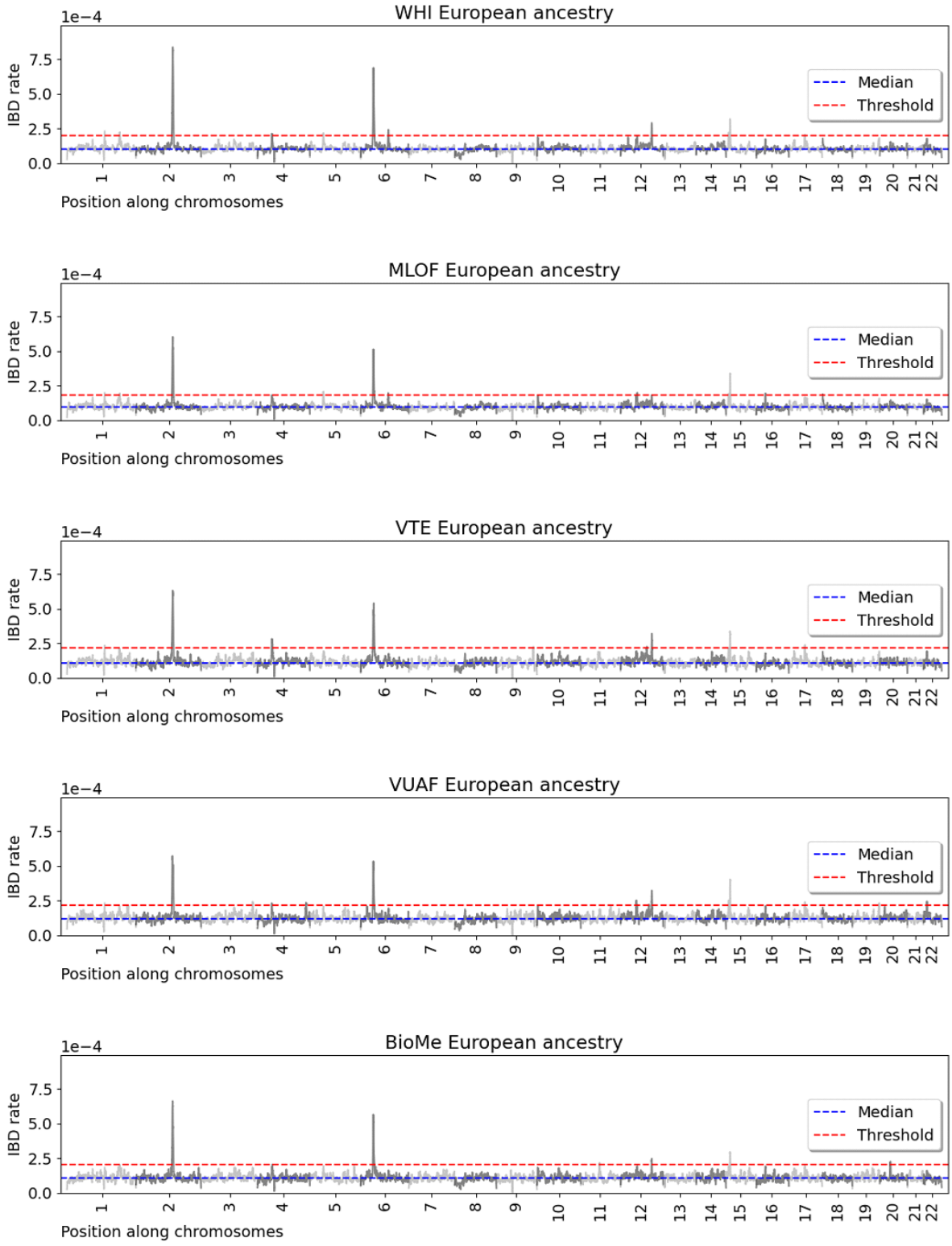

Figure S60: Cohort study selection scans of TOPMed European Americans

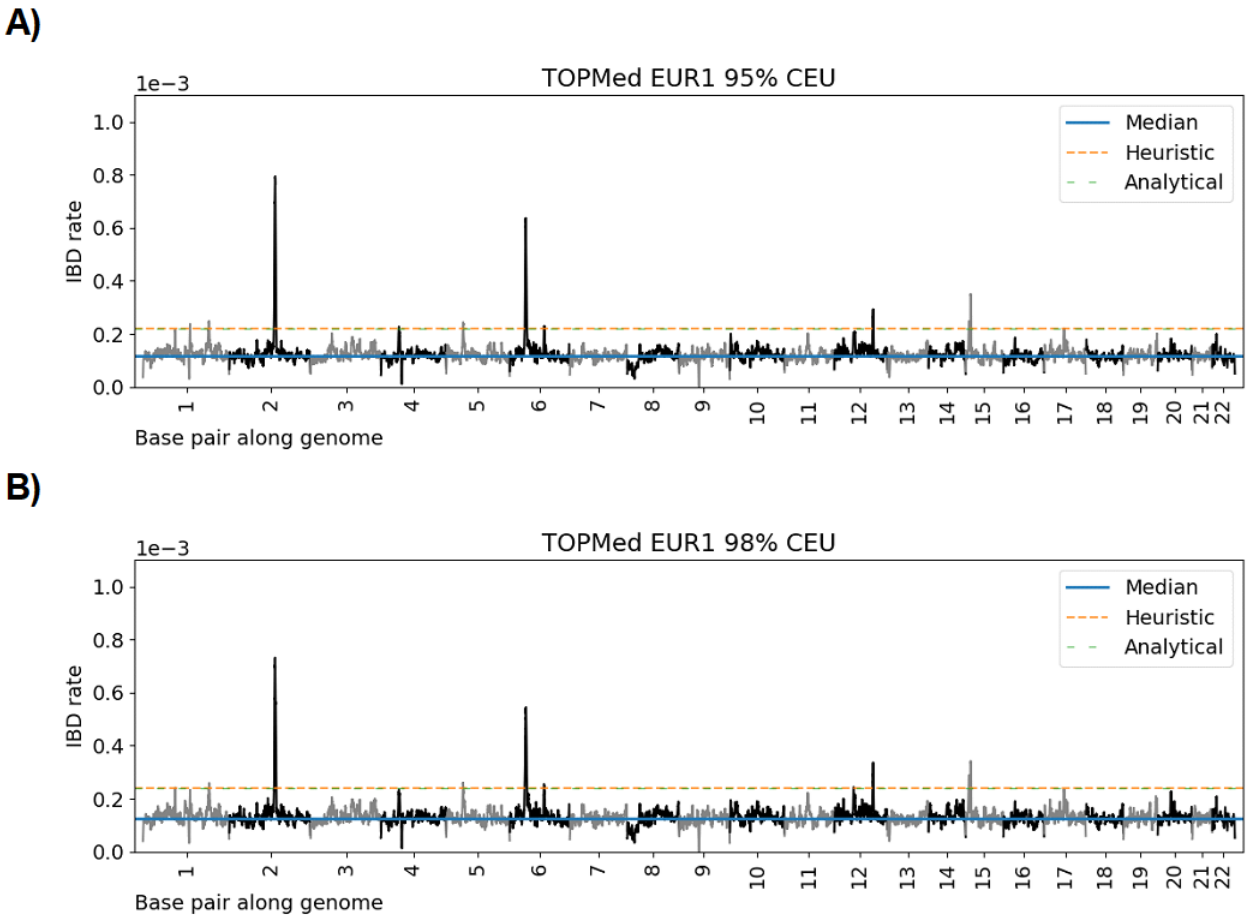

Figure S61: Selection scans of TOPMed samples with different CEU ancestry proportions. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on the first inferred European ancestry samples in the TOPMed project. Subsets of individuals with A) 95% and B) 98% CEU ancestry are analyzed. The main text describes the second versus the first inferred European groups. Horizontal dashed lines show (blue) the genome-wide median IBD rate, (orange) the heuristic threshold of four standard deviations above the median, and (green) the Siegmund and Yakir discrete approximation (S&Y). The S&Y method is calculated assuming hypothesis testing every 0.02 cM.

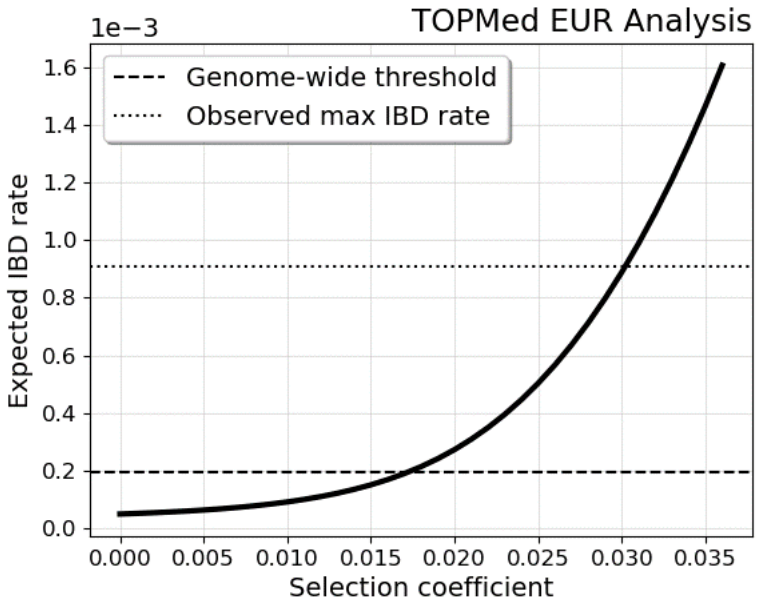

Figure S62: Expected IBD rates in the TOPMed EUR analysis for different selection coefficients. The expected IBD rate (y-axis) is calculated for different selection coefficients (x-axis). This calculation is conditional on the recent effective sizes that are inferred for the EUR1 group. The expected IBD rate is an average of the expected IBD rates conditional on allele frequencies between 0.1 to 0.9 (discretized every 0.1). The horizontal dashed line is the genome-wide scan threshold we use in our analysis. The horizontal dotted line is the observed maximum IBD rate (near the LCT gene).

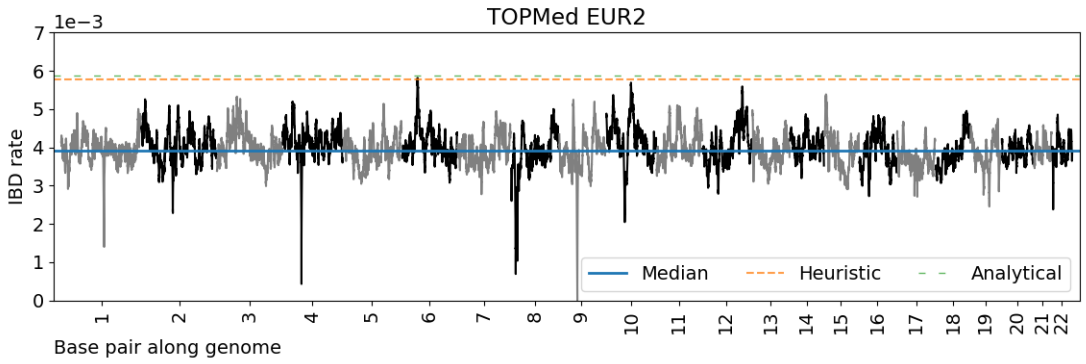

Figure S63: Genome-wide IBD rate scans in second European ancestry data. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on the second inferred European ancestry samples in the TOPMed project. The main text describes the second versus the first inferred European groups. Each subplot has a different y-axis scale. Horizontal dashed lines show (blue) the genome-wide median IBD rate,(orange) the heuristic threshold of four standard deviations above the median, and (green) the Siegmund and Yakir discrete approximation (S&Y). The S&Y method is calculated assuming hypothesis testing every 0.02 cM.

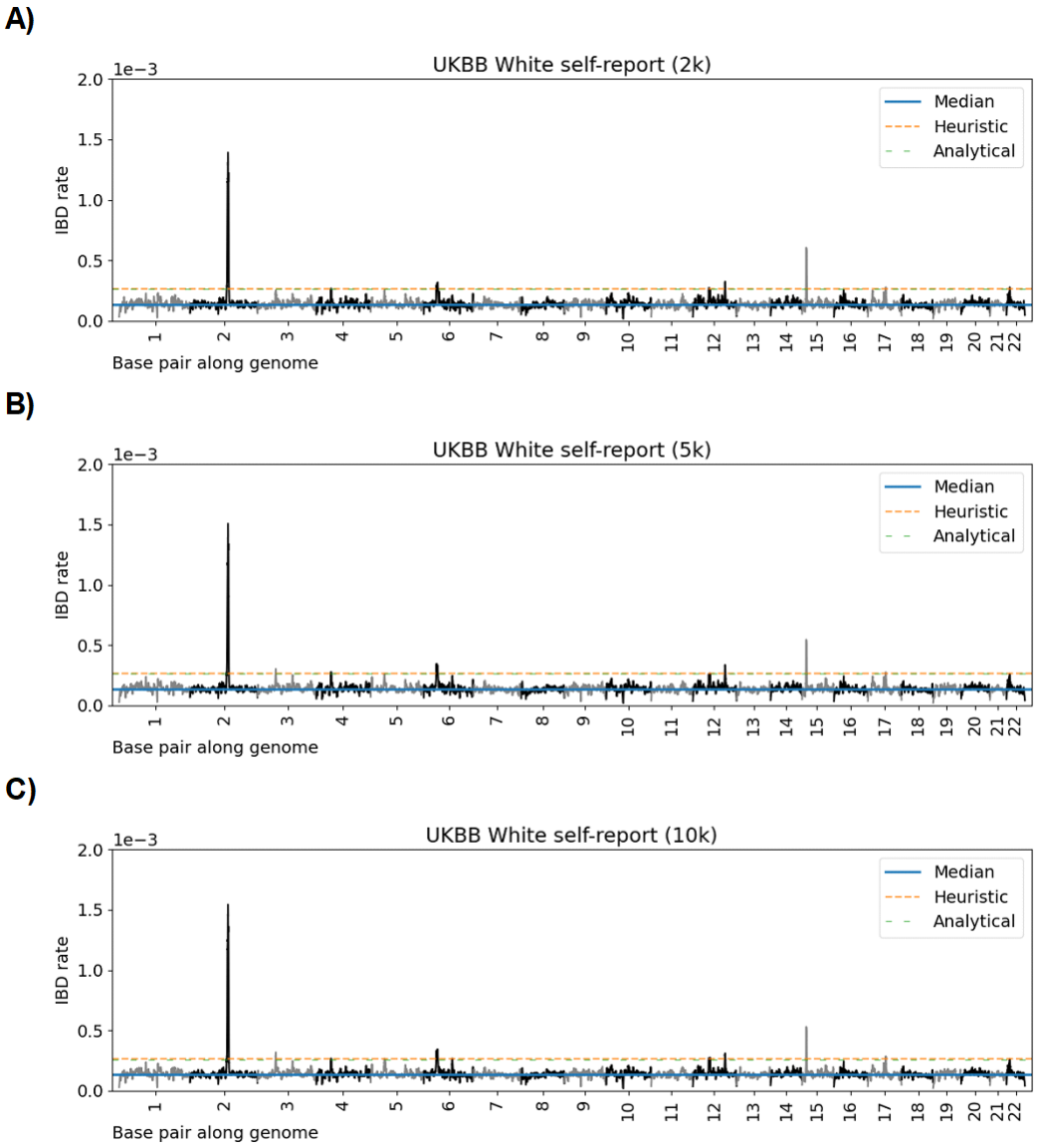

Figure S64: Genome-wide IBD rate scans in UKBB White British subsets of different size. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on UKBB White British subsets of A) 2k, B) 5k, and C) 10k samples. Horizontal dashed lines show (blue) the genome-wide median IBD rate, (orange) the heuristic threshold of four standard deviations above the median IBD rate, and (green) the Siegmund and Yakir discrete approximation (S&Y) calculated assuming hypothesis testing every 0.02 cM.

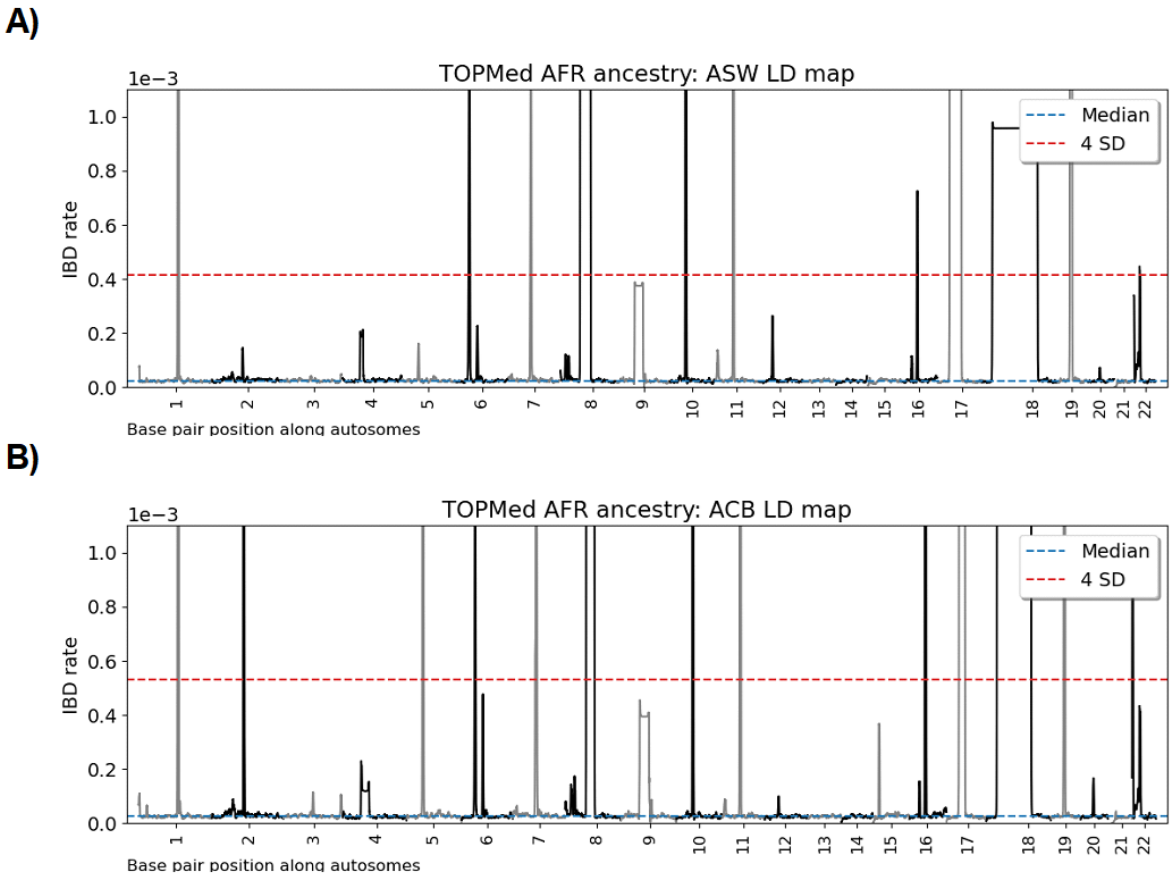

Figure S65: Genome-wide IBD rate scans in African ancestry data using LD-based genetic maps. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on inferred African ancestry samples in the TOPMed project using the Spence and Song [141] LD-based genetic maps with reference cohorts A) ASW and B) ACB from the HapMap project [10]. The human reference genome is GRCh38. The y-axis scale is the same as is used in the analysis with pedigree-based genetic maps. Horizontal dashed lines show (blue) the genome-wide median IBD rate and (red) the heuristic threshold of four standard deviations above the median.

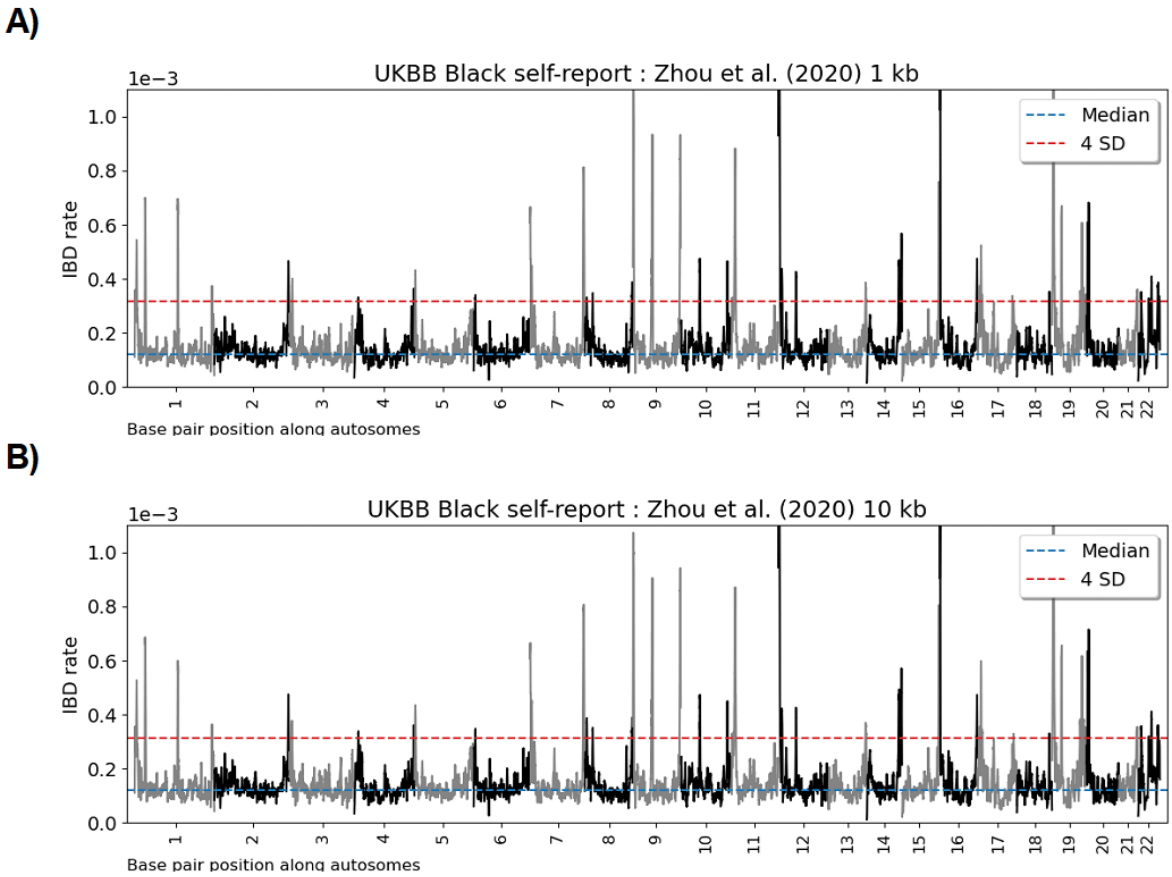

Figure S66: Genome-wide IBD rate scans in African ancestry data using IBD-based genetic maps. Line plots show IBD rates (y-axis) for base pair positions along twenty-two human autosomes. The IBD rate is calculated every 20 kb. Data is based on UKBB Black self-report samples using Zhou et al. [162] IBD-based genetic maps with resolutions A) 1 kb and B) 10 kb. The human reference genome is GRCh37. The y-axis scale is the same as is used in the analysis with pedigree-based genetic maps. Horizontal dashed lines show (blue) the genome-wide median IBD rate and (red) the heuristic threshold of four standard deviations above the median.

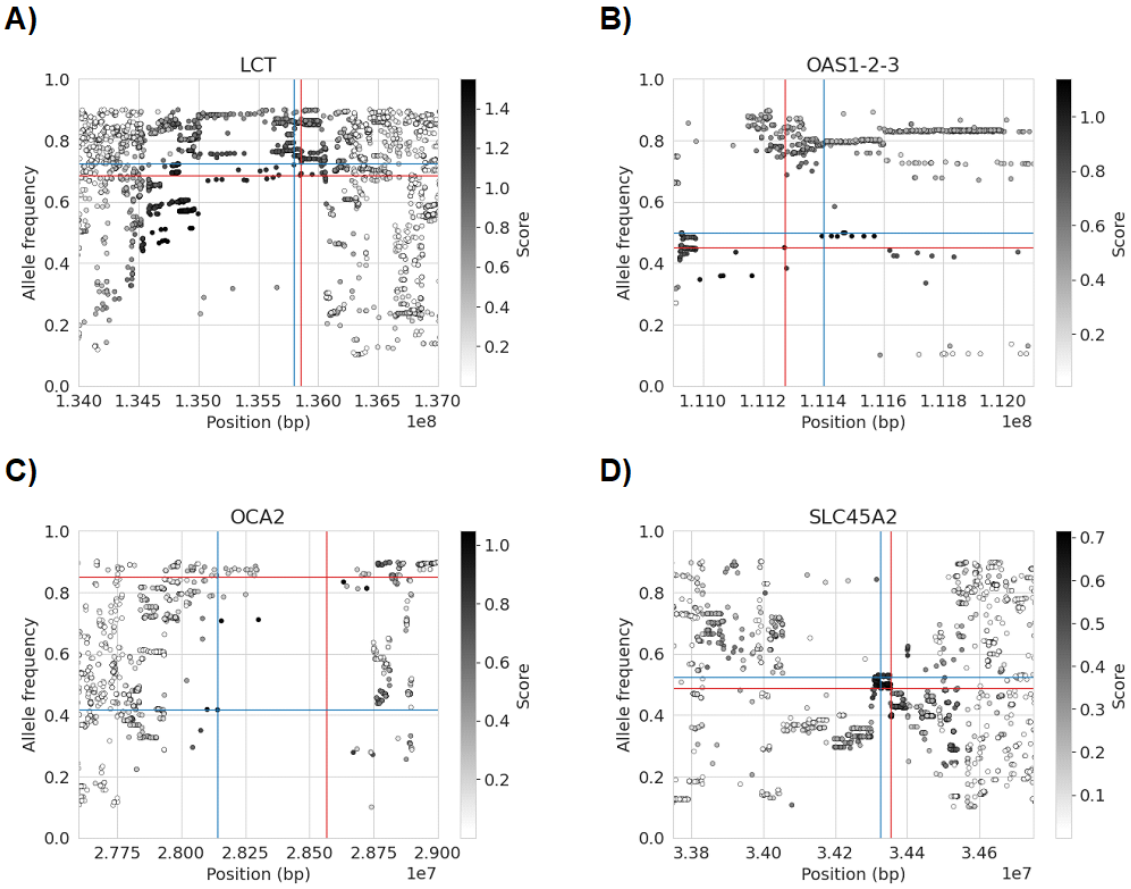

Figure S67: Spectrum of common variants at selected loci. Position in basepairs (x-axis) by allele frequency (y-axis) for common variants nearby loci passing the selection scan: (A) LCT; (B) OAS1-2-3; (C) OCA2; and (D) SLC45A2. Color bars show the z_j scores, where darker grays correspond to greater scores. Blue lines denote the haplotype-based allele frequency and base pair location estimates. Red lines denote the frequency and base pair location of the SNP with the greatest z_j score. Results are shown for the EUR1 group (A, B, D) and the WHI cohort (C).

Appendix E

SUPPLEMENTARY TABLES

Identity-by-descent in large samples

Type	Structure	Avg	Var	Min	Max	S.W.t.
G3	Edges	3,085.66	12,827.06	2,554	3,716	0.29
	Largest	15.60	9.22	9	50	1.00
	Tree2	757.96	635.16	644	880	0.08
	Tree3	99.71	94.29	54	149	0.42
	Complete	251.55	230.65	185	3,118	0.14
BN	Edges	587.73	694.55	469	716	0.12
	Largest	4.32	0.39	3	11	1.00
	Tree2	473.24	393.50	377	574	0.09
	Tree3	9.66	9.57	0	27	1.00
	Complete	35.53	34.76	11	488	1.00

Table S1: Summary statistics of IBD graphs for the three phases of exponential growth (G3) and the population bottleneck (BN) demographic scenarios. Network structures of interest are the number of edges (Edges), the degree of the largest components (Largest), the number of trees of order 2 and 3 (Tree-2 and Tree-3), and the number of complete components of degree 3 or more (Complete). Summary statistics are aggregated over at least six hundred thousand simulations. Shapiro-Wilk tests at a confidence level of 0.05 are performed with 1000 replicates for at least 600 simulations. The proportions of rejected null hypotheses are reported as S.W.t. The sample size is five thousand individuals. The Morgan length threshold is 0.03.

Type	Structure	Avg	Var	Min	Max	S.W.t.
$s = 0.01$	Edges	3,407.38	21,526.32	2,916	4,143	0.36
	Largest	**24.33**	50.80	11	89	0.97
	Tree2	737.77	626.12	636	842	0.05
	Tree3	95.81	92.23	57	138	0.07
	Complete	242.41	215.76	187	305	0.05
$s = 0.02$	Edges	4,693.51	140,436.48	3,579	8,212	0.95
	Largest	**73.97**	1,219.95	22	346	0.97
	Tree2	697.19	588.38	596	791	0.10
	Tree3	86.65	83.70	53	126	0.09
	Complete	220.37	199.88	161	281	0.10
$s = 0.03$	Edges	8,242.12	2,283,864.57	4,998	37,933	0.97
	Largest	**230.39**	12,224.19	39	819	0.97
	Tree2	659.10	565.21	562	759	0.07
	Tree3	78.43	74.69	46	119	0.11
	Complete	199.95	181.88	145	254	0.06
$s = 0.04$	Edges	16,486.56	24,295,227.62	7,747	72,775	0.97
	Largest	**484.92**	38,683.32	89	1,229	0.97
	Tree2	630.68	529.35	546	731	0.02
	Tree3	72.95	70.26	41	108	0.11
	Complete	185.76	167.85	135	241	0.07

Table S2: Summary statistics of IBD graphs for different selection coefficients. There is directional selection with different selection coefficients $s \in [0.01, 0.02, 0.03, 0.4]$. The same description of IBD graph features as in Table 3.1. Shapiro-Wilk tests at a confidence level of 0.05 are performed with 250 replicates for 150 simulations. The proportions of rejected null hypotheses are reported as S.W.t. The demographic scenario is three phases of exponential growth (G3). The sample size is five thousand individuals. The Morgan length threshold is 0.03.

Type	Structure	Avg	Var	Min	Max	S.W.t.
$s = 0.01$	Edges	612.05	753.44	504.00	736.00	0.06
	Largest	**4.71**	0.75	3.00	14.00	0.97
	Tree2	481.32	400.48	397.00	566.00	0.06
	Tree3	11.33	11.24	1.00	25.00	0.90
	Complete	39.25	37.75	15.00	66.00	0.19
$s = 0.02$	Edges	722.33	1,349.58	582.00	967.00	0.38
	Largest	**9.79**	20.27	4.00	56.00	0.97
	Tree2	497.56	407.99	416.00	581.00	0.03
	Tree3	16.38	16.05	3.00	34.00	0.72
	Complete	50.79	48.02	24.00	81.00	0.15
$s = 0.03$	Edges	1,090.00	16,537.54	808.00	2,360.00	0.97
	Largest	**40.15**	456.43	8.00	172.00	0.97
	Tree2	501.78	424.81	409.00	592.00	0.06
	Tree3	20.80	20.37	4.00	43.00	0.47
	Complete	61.55	58.15	33.00	93.00	0.14
$s = 0.04$	Edges	2,177.58	284,697.22	1,219.00	7,591.00	0.97
	Largest	**122.45**	2,833.45	18.00	354.00	0.97
	Tree2	492.44	425.42	412.00	578.00	0.01
	Tree3	22.28	21.94	6.00	44.00	0.46
	Complete	66.05	63.26	36.00	99.00	0.19

Table S3: Summary statistics of IBD graphs for different selection coefficients and the population bottleneck demographic scenario. There is directional selection with different selection coefficients $s \in [0.01, 0.02, 0.03, 0.4]$. The same description of IBD graph features as in Table 3.1. Shapiro-Wilk tests at a confidence level of 0.05 are performed with 250 replicates for 150 simulations. The proportions of rejected null hypotheses are reported as S.W.t. The sample size is five thousand individuals. The Morgan length threshold is 0.03.

Selection coefficient estimation

	Parameter		s			
Scenario	**True**	**Assumed**	**True**	**Estimate** $\bar{\hat{s}}$	**Width**	**Coverage**
$p(0)$	0.40	0.50	0.0300	0.0305	0.0074	95.8%
$p(0)$	0.60	0.50	0.0300	0.0301	0.0063	94.6%
$p(0)$	0.20	0.25	0.0300	0.0316	0.0100	93.6%
$p(0)$	0.30	0.25	0.0300	0.0289	0.0080	90.3%
$N_e(t)$	BN	0.9x	0.0300	0.0282	0.0070	75.9%
$N_e(t)$	BN	1.1x	0.0300	0.0314	0.0064	92.1%
SV	0.05	0.00	0.0300	0.0300	0.0093	94.6%
SV	0.02	0.00	0.0300	0.0300	0.0073	94.9%
SV	0.01	0.00	0.0300	0.0300	0.0064	94.9%

Table S4: Selection coefficient estimation based on true IBD segments and model misspecification. Each row represents the results of 10,000 replicates. Summary statistics are the mean estimate, the mean confidence interval width, and the percentage of confidence intervals containing the true selection coefficient (coverage). Estimation is conditional on true IBD segments longer than 3.0 cM. Ninety-five percent confidence intervals are based on fifty bootstraps. Default settings are $s = 0.03$, $p\,(0) = 0.50$, population bottleneck demographic history (BN), the multiplicative genetic model, and an ongoing sweep from a new beneficial mutation. The sample size is five thousand diploids. The Assumed column is the parameter value used in the conditional estimation of s. Abbreviations 0.9x and 1.1x indicate that the $N_e(t)$ is uniformly scaled by that value. SV concerns the allele frequency of standing neutral variation.

Methods to study selection in genetic data

cM segment type	Software	Algorithm settings	Purpose
≥ 2.0 and ≥ 3.0	`hap-ibd`	min-seed=0.5 min-extend=0.2 min-output=1.0 min-maf=0.4	Candidate segments for `ibd-ends`
≥ 2.0 and ≥ 3.0	`ibd-ends`	min-maf=0.001 err=2e-4 Median endpoints	Selection scan and estimation
≥ 1.0	`hap-ibd`	min-seed=0.5 min-extend=0.2 min-output=1.0 min-maf=0.1	IBD outgroups

Table S5: Algorithm settings for detecting IBD segments in sequence data. Unless otherwise specified in this table, we use the default settings in `hap-ibd` and `ibd-ends`. We use `hap-ibd` version 1.0 from May 10, 2023 and `ibd-ends` version 1.1 from May 20, 2022. The IBD segments ≥ 2.0 and ≥ 3.0 cM are a filtered subset of the `ibd-ends` output. The error rate in `ibd-ends` should be modified if in a preliminary data analysis the estimated error rate differs from the default by more than a factor of three.

Name	Purpose	Approach	Input data	Assumptions	Time
`iSAFE`	Ranking SNPs	EHH-based	Phased variants	Derived allele known	34.07 min
`iSWEEP`	Ranking SNPs	IBD-based	Phased variants; $\geq$1.0 cM IBD	Inferred adaptive allele	5.83 min
`iSWEEP`	Estimating $p(0)$	IBD-based	Phased variants; $\geq$1.0 cM IBD	Inferred adaptive allele	11.75 sec
`CLUES2`	Estimating s	Modeling coalescent	Inferred ARGs	Causal allele known	1-2 days
`ImaGene`	Estimating s	Deep learning	Simulated training data	N/A	$\geq$ 1 week training
`iSWEEP`	Estimating s	IBD-based	Phased variants; $\geq$3.0 cM IBD	Estimates for $p(0)$ & locus	1.67 sec

Table S6: Algorithm settings for simulation study on sequence data. Average CPU time for `iSAFE` and `iSWEEP` is based on four s=0.03 simulations of 5000 samples. (Wall clock time for `iSWEEP` is faster if running `hap-ibd` in parallel across cores.) `CLUES2` and `ImaGene` are analyzed using 500 of 5000 samples due to runtime limitations. `iSWEEP` time to estimate s is the total time divided by 100 bootstraps. All methods except `ImaGene` require a genetic map; `ImaGene` requires a genome-wide recombination rate.

True s	**Estimate** $\bar{\hat{s}}$	**MAD**	**Width**	**Coverage**
0.040	0.0391	0.0022	0.0103	99.0%
0.035	0.0350	0.0013	0.0087	98.0%
0.030	0.0306	0.0013	0.0078	100.0%
0.025	0.0259	0.0017	0.0075	96.5%
0.020	0.0227	0.0029	0.0100	76.2%
0.015	0.0193	0.0048	0.0119	43.5%

Table S7: Selection coefficient estimates based on IBD segments inferred from sequence data given the known sweeping allele. For each row, the average of estimates $\hat{s}$, mean absolute deviation (MAD) between estimates and the truth, average confidence interval width, and confidence interval coverage are aggregated over two hundred simulations of sequence data. Estimation is conditional on IBD segments longer than 3.0 cM and the known frequency and location of the sweeping allele. Ninety-five percent confidence intervals are based on one hundred bootstraps. The sample size is five thousand diploids. Population bottleneck (BN) is the demographic scenario. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

Selection coefficient	`iSWEEP` **interval width**	`ImaGene` **interval width**	**Ratio of** `ImaGene` **to** `iSWEEP` **width**
0.04	0.0106	0.0217	2.05
0.035	0.0090	0.0195	2.14
0.03	0.0078	0.0191	2.42
0.025	0.0077	0.0193	2.50
0.02	0.0104	0.0233	2.22
0.015	0.0142	0.0230	1.62

Table S8: Comparing uncertainty quantification in selection coefficient estimation between `iSWEEP` and `ImaGene`. For each row, the average confidence interval width is aggregated over two hundred simulations of sequence data. Ninety-five percent confidence intervals for `iSWEEP` are based on one hundred bootstraps. Ninety-five percent high posterior density intervals for `ImaGene` are based on ten thousand bootstraps. Width is the difference between the right and left endpoints of the interval. The simulation study settings are described in the main text.

$N_e(t)$	**True** s	**Estimate** $\bar{\hat{s}}$	**MAD**	**Width**	**Coverage**
BN	0.030	0.0302	0.0015	0.0078	95.0%
C25	0.030	0.0300	0.0025	0.0137	96.0%
G3	0.030	0.0283	0.0023	0.0076	76.5%
$m = 0.4$	0.030	0.0295	0.0020	0.0074	83.5%

Table S9: Selection coefficient estimation based on IBD segments inferred from sequence data and different demographic scenarios. For each row, the average of estimates $\hat{s}$, mean absolute deviation (MAD) between estimates and the truth, average confidence interval width, and confidence interval coverage are aggregated over two hundred simulations of sequence data. Estimation is conditional on IBD segments longer than 3.0 cM and inferred frequency and location of the sweeping allele. Ninety-five percent confidence intervals are based on one hundred bootstraps. The sample size is five thousand diploids. Demographic scenarios in the first column are abbreviated as bottleneck (BN), three phases of exponential growth (G3), constant size N_e=25,000 (C25), and continuous mixing with rate $m = 0.40$ between two subpopulations. Mutation, recombination, and gene conversion rates are 1e-8, 1e-8, and 2e-8.

Scanning for excess identity-by-descent rates

Nominal level	**Adjusted** Analytical	Simulation	Bonferroni	**FWER** Analytical	Simulation	Bonferroni
0.01	1.58e-6	2.01e-6	2.08e-7	0.008	0.012	0.002
0.05	9.23e-6	1.06e-5	1.04e-6	0.030	0.034	0.006
0.10	2.03e-5	2.29e-5	2.08e-6	0.066	0.078	0.012

Table S10: Significance levels and family-wise error rates after multiple testing corrections with IBD segments $\geq$ 3.0 cM. Significance levels are adjusted for multiple testing based on scans over 10 chromosomes of size 100 cM and tests every 0.02 cM (50,000 total tests). The analytical approximation of Siegmund and Yakir [137] and the simulation method are based on a fitted Ornstein-Uhlenbeck process. Each simulation has a different threshold as a result of estimating θ. Family-wise error rate (FWER) is the percentage of five hundred genome-wide scans that have at least one statistically significant result. The sample size is twenty-five hundred diploids. The demographic scenario is the population bottleneck. The IBD segment detection threshold is 3.0 cM.

Chr	Max IBD rate	Position (cM)	Position (Mb)
2	6.62E-04	140.27	136.98
6	2.40E-04	51.46	33.92
15	2.09E-04	15.88	31.48
16	2.05E-04	35.78	17.92

Table S11: Regions highlighted in UKBB Indian self-report selection scan. Loci where the identity-by-descent (IBD) rate exceeds the multiple testing adjusted threshold of 1.81e-4 at the 0.05 family-wise significance level. Physical and genetic positions for the location of maximum IBD rate are shown in megabases (Mb) and centiMorgans (cM). Locations are aligned to build GRCh37. The pedigree-based recombination map from Bhérer et al. [15] is used when inferring IBD segments.

Chr	Max IBD rate	Position (cM)	Position (Mb)
11	9.66E-04	9.98	5.22
16	6.46E-04	35.40	17.76
22	4.54E-04	10.69	21.38
7	4.24E-04	88.60	80.36
6	4.12E-04	52.19	34.42
17	3.82E-04	9.87	3.72
5	3.76E-04	22.61	9.58
11	3.76E-04	66.40	61.02
8	3.63E-04	56.07	37.18
10	3.62E-04	89.44	79.38
1	3.57E-04	245.78	240.70

Table S12: Regions highlighted in UKBB Black self-report selection scan. Loci where the identity-by-descent (IBD) rate exceeds the multiple testing adjusted threshold of 3.54e-4 at the 0.05 family-wise significance level. Physical and genetic positions for the location of maximum IBD rate are shown in megabases (Mb) and centiMorgans (cM). Locations are aligned to build GRCh37. The pedigree-based recombination map from Bhérer et al. [15] is used when inferring IBD segments.

Map	Type	cM at 17 Mb	cM at 18 Mb	cM difference
deCODE [70]	Pedigree	34.45	36.59	2.14
Bhérer et al. [15]	Pedigree	34.00	35.80	1.80
Spence and Song [141] ASW	LD	20.47	21.48	1.01
Zhou et al. [162] JHS	IBD	38.91	39.45	0.54

Table S13: African ancestry specific genetic maps versus fine scale pedigree-based genetic maps from European ancestry at XYLT1 gene. The third and fourth columns give the conversion to cM for 17 Mb and 18 Mb from different recombination maps. The last column is the difference in cM genetic positions. The XYLT1 gene is within 17 Mb to 18 Mb in both GRCh37 and GRCh38 references. The Bhérer et al. [15] and Zhou et al. [162] genetic maps are with respect to the GRCh37 reference. The deCODE [70] and Spence and Song [141] genetic maps are with respect to the GRCh38 reference. ASW stands for African-Americans in the Southwest, a cohort sample from the HapMap project. JHS stands for Jackson Heart Study, a predominantly African-American cohort in the TOPMed project. LD stands for linkage disequilibrium. IBD stands for identity-by-descent. The Spence and Song [141] genetic map for the ACB cohort is approximately the same as the Spence and Song [141] genetic map for the ASW cohort. The Zhou et al. [162] genetic map is for increments of 1 kb. cM positions and differences are approximately the same with the Zhou et al. [162] genetic map for 10 kb.

Gene	Cohort	Location	$\hat{p}(0)$	$\hat{s}$	Gini	Freq. $\hat{A}$
LCT	EUR	2:135,851,076	0.6864	0.0325	0	0.65
	BioMe	2:135,851,076	0.6769	0.0291	0	0.58
	MLOF	2:135,851,076	0.6599	0.0294	0	0.57
	VTE	2:135,851,076	0.7067	0.0319	0	0.61
	VUAF	2:135,080,336	0.6834	0.0300	0	0.47
	WHI	2:135,851,076	0.6863	0.0317	0	0.64
OAS	EUR	12:111,270,654	0.4506	0.0182	0.71	0.19
	BioMe	12:111,270,654	0.4607	0.0115	0	0.24
	MLOF	.	.	.	.	.
	VTE	12:111,598,263	0.1138	.	0.66	0.05
	VUAF	12:111,270,654	0.4360	0.0196	0.61	0.10
	WHI	12:111,446,804	0.4475	0.0175	0.61	0.19
OCA2	EUR	.	.	.	.	.
	BioMe	.	.	.	.	.
	MLOF	15:28,632,448	0.8294	0.0174	0.30	0.64
	VTE	.	.	.	.	.
	VUAF	15:28,100,878	0.4221	0.0199	0	0.40
	WHI	15:28,141,480	0.4170	0.0164	0	0.69
TRPM1	EUR	15:31,437,832	0.1223	0.0273	0.96	0.26
	BioMe	15:31,437,832	0.1223	0.0264	0.73	0.05
	MLOF	15:31,437,832	0.1249	0.0289	0.82	0.10
	VTE	15:31,437,832	0.1163	0.0297	0.74	0.05
	VUAF	15:31,437,832	0.1322	0.0297	0.76	0.08
	WHI	15:31,437,832	0.1209	0.0277	0.91	0.12

Table S14: Cohort-specific estimation for TOPMed EUR cohorts. Gini impurity index, the proportion of haplotypes in the IBD outgroup $\hat{A}$, and allele frequency, location, and selection coefficient estimates are reported for selected loci. Either the Gini impurity index is < 0.60 or there is consensus across studies of the best-ranked variant. Boldface marks SNPs that are best-ranked among multiple cohorts. Dot . marks if the locus does not pass the scan.

Appendix F

SUPPLEMENTARY MATERIALS

S1 Funding

The research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number HG005701. S.D.T. acknowledges funding support from the National Defense Science and Engineering Graduate Fellowship and National Institute of Health T32 GM081062 Pre-doctoral Training Grant in Statistical Genetics. Finally, we acknowledge the Departments of Statistics and Biostatistics at the University of Washington for their support and maintenance of computing resources.

S2 Software Resources

The following list includes software developed in this dissertation:

- https://github.com/sdtemple/isweep: a Python package and an analysis workflow to implement methods in Chapter 5
- https://github.com/sdtemple/flare-pipeline: an analysis workflow for phasing, local ancestry inference, and identity-by-descent segment detection

The following are key software resources developed by other authors:

- https://github.com/avaughn271/CLUES2: estimating selection coefficients using ancestral recombination graphs
- https://github.com/browning-lab/hap-ibd: detecting identity-by-descent segments in marker data

- https://github.com/browning-lab/ibd-ends: detecting identity-by-descent segments in marker data
- https://faculty.washington.edu/browning/Beagle/Beagle.html: haplotype phasing
- https://github.com/Ying001/IBDkin: estimating kinship using identity-by-descent segments
- https://faculty.washington.edu/browning/ibdne.html: recent demographic inference using identity-by-descent segments
- https://github.com/mfumagalli/ImaGene: estimating selection coefficients using deep learning
- https://github.com/alek0991/iSAFE: ranking candidate alleles for selection
- https://myersgroup.github.io/relate/: inferring the ancestral recombination graph
- https://tskit.dev/: utilities for tree sequences
- https://github.com/bguo068/tskibd: deriving identity-by-descent segments from tree sequences
- https://messerlab.org/SLiM/: simulating genetic data in complex scenarios
- https://github.com/snakemake/snakemake: programming language to develop bioinformatics pipelines

S3 Data Acknowledgements

S3.1 Open source data

We thank those individuals who contributed genetic data to the HapMap, 1000 Genomes, and Human Genetic Diversity Panel projects. We also thank the researchers involved in

making this data publicly available.

S3.2 UK Biobank

This research has used the UK Biobank Resource under Application Number 19934.

S3.3 Trans-omics for Precision Medicine

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung, and Blood Institute (NHLBI). Core support including centralized genomic-read mapping and genotype calling along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I).

We gratefully acknowledge the studies and participants who provided biological samples and data for the TOPMed project. Funding for the Barbados Asthma Genetics Study (BAGS) was provided by National Institutes of Health (NIH) R01HL104608, R01HL087699, and HL104608 S1. The Mount Sinai BioMe Biobank (BioMe) has been supported by The Andrea and Charles Bronfman Philanthropies and in part by funds from the NHLBI and the National Human Genome Research Institute (NHGRI) (U01HG00638001; U01HG007417; X01HL134588); genome sequencing was funded by contract HHSN268201600037I. The Cleveland Clinic Atrial Fibrillation study (CCAF) was supported by NIH grants R01 HL 090620 and R01 HL 111314, the NIH National Center for Research Resources for Case Western Reserve University and Cleveland Clinic Clinical and Translational Science Award UL1-RR024989, the Cleveland Clinic Department of Cardiovascular Medicine philanthropy research funds, and the Tomsich Atrial Fibrillation Research Fund; genome sequencing was supported by R01HL092577. The Framingham Heart Study (FHS) was supported by contracts

NO1-HC-25195, HHSN2682015000001I and 75N92019D00031 from the NHLBI and grant supplement R01 HL092577-06S1; genome sequencing was funded by HHSN268201600034I and U54HG003067. The Hypertension Genetic Epidemiology Network Study (HyperGen) is part of the NHLBI Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472, U01 HL054473, U01 HL054495, and U01 HL054509; genome sequencing was funded by R01HL055673. The Jackson Heart Study is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from NHLBI and the National Institute for Minority Health and Health Disparities (NIMHD); genome sequencing was funded by HHSN268201100037C. The My Life, Our Future samples (MLOF) and data are made possible through the partnership of Bloodworks Northwest, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation, and Bioverativ; genome sequencing was funded by HHSN268201600033I and HHSN268201500016C. The Venous Thromboembolism project (VTE) was funded in part by grants from the NIH, NHLBI (HL66216 and HL83141) and the NHGRI (HG04735). The Vanderbilt Genetic Basis of Atrial Fibrillation study (VUAF) was supported by grants from the American Heart Association (EIA 0940116N), and grants from the National Institutes of Health (HL092217, U19 HL65962, and UL1 RR024975), and by CTSA award (UL1TR000445) from the National Center for Advancing Translational Sciences; genome sequencing was funded by R01HL092577. The Women's Health Initiative program (WHI) is funded by NHLBI through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005; genome sequencing was funded by HHSN268201500014C.