Highlights

Identity-by-descent segments in large samples

Seth D. Temple, Elizabeth A. Thompson

- We show the asymptotic normality of the identity-by-descent rate, a mean of correlated binary random variables that arises in population genetics studies.
- We describe an efficient algorithm capable of simulating long identity-by-descent segments around a locus in large sample sizes.
- In enormous simulation studies, we use this algorithm to characterize the distributional properties of the identity-by-descent rate.
- In finite samples, we reject the null hypothesis of normality more often than the nominal significance level, indicating that genome-wide scans based on identity-by-descent rates may be anti-conservative.

Identity-by-descent segments in large samples

Seth D. Temple[a], Elizabeth A. Thompson[a]

[a]*Department of Statistics, Seattle, 98195, Washington, United States of America*

Abstract

If two haplotypes share the same alleles for an extended gene tract, these haplotypes are likely to derive identical-by-descent from a recent common ancestor. 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 at a locus is normally distributed for large sample size and large scaled population size. We use efficient and exact simulations to study the distributional behavior of the detectable identity-by-descent rate in finite samples. One consequence of non-normality in finite samples is that genome-wide scans based on identity-by-descent rates may be subject to anti-conservative Type 1 error control.

Keywords: identity-by-descent, coalescent, covariance, asymptotic normality

1. Introduction

Two individuals share a haplotype segment identical-by-descent (IBD) if they inherit it from the same common ancestor. Here we study the length of IBD segments overlapping a single focal location. Ignoring gene conversion, IBD segments are randomly cut by crossover recombination in each future generation. The length of an IBD segment is thus shorter with higher probability the more removed its

common ancestor is from the present-day. This measurement can serve as a proxy for the time of common ancestry at a locus [30].

Using modern methods, sufficiently long IBD segments can be detected with high accuracy from genetic data [14, 27, 36]. Detectable segments can provide rich information about the recent genetic history of a population sample. For instance, detected IBD segments have been used to test for rare variant associations when a disease allele is untyped or a genome-wide association study is underpowered [7, 15, 23]. They have also been used to estimate relatedness [14, 27, 36], haplotype phase [3, 21], mutation rates [26, 31, 32], recombination rates [34], gene conversion rates [6, 26], demographic changes [4, 8, 25], and positive selection [29].

Some IBD-based methods that have been proposed to estimate population genetics parameters solve composite likelihoods [25, 29, 31]. Composite likelihood approaches are commonly used in genetics when it is analytical intractable or computationally expensive to address dependencies in the data [20]. For independent, identically distributed data, maximizing likelihood estimators are asymptotically consistent, efficient, and normal under regularity conditions [9]. Such properties are appealing in statistical inference. To what extent these properties extend to composite likelihood estimators is generally unknown [20]. Studying these estimators can be especially challenging if the maximum does not have a closed form [25, 31].

Quantifying uncertainty is another challenge in IBD-based studies. Bootstrap resampling is commonly employed in IBD-based estimation procedures [4, 6, 8, 25, 31]. For significance level α, these existing works do not demonstrate that their $(1-\alpha)\%$ bootstrap confidence intervals contain a true parameter in $(1-\alpha)\%$

of simulations. Moreover, nonparametric bootstrapping tends to give confidence intervals that are not wide enough to satisfy coverage [22]. In a simple example, Temple et al. [29] show that simulating IBD segments independently can understate the variance in the number of detectable IBD segments at a locus. The parametric bootstrap approach proposed in Temple et al. [29] gives adequate coverage in selection coefficient estimation for numerous simulation studies. The technique implicitly assumes that the number of detectable IBD segments at a locus, and certain functions[1] thereof, are normally distributed in large samples.

Here we derive sufficient conditions under which the proportion of detectable IBD segments at a locus[2] 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 dominates the covariance between detectable IBD segments. Our conditions involve a minimum length of detectable IBD segments and the population size 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. This condition we appeal to as the driving force for the efficiency of *in silico* studies on normality. The overall contribution of this work is to support IBD-based statistical inference with rigorous theory and extensive simulation studies.

The outline of the paper is as follows. In Section 2, we formally define our probability model for IBD segments. In Section 3, we present and prove our

[1]Functions that satisfy the Delta method conditions [9]: the first derivative at the expected number of detectable IBD segment at a locus exists and is non-zero.

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

main result for the asymptotic normality of the detectable IBD rate. In Section 4, we discuss and build upon an algorithm to efficiently simulate detectable IBD segments for huge sample sizes. In Section 5, we use simulation to investigate the statistical properties of IBD-based estimators and IBD graphs at a locus. Many calculations of covariance terms are left to the Appendix.

2. Shared haplotypes overlapping a focal location

The IBD segment model integrates over two waiting time distributions: the time until a common ancestor and the genetic length until a crossover. The end-points of an extant haplotype segment can be modeled conditional on the coalescent time, where the surviving length is the minimum of independent crossover points over generations. First, we review two ancestral tree processes: the discrete-time Wright-Fisher model and the n-coalescent. Second, we relate the coalescent time to IBD segment lengths.

2.1. The time until a common ancestor

Let n be the sample size and $k \leq n$ be the size of a subsample. Define $N(t)$ to be the population size t generations ago. 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

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

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

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

when k is much smaller than $\min_t N(t)$ [17]. The geometric model assumes that multiple coalescent events in a single generation is 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 A.3.1.

The n-coalescent comes from the continuous time limit of Equations 1 and 2 for constant population size N [3]. Specifically, T_k converges weakly to Exponential($\binom{k}{2}$) for $k \ll N$, $N \to \infty$ and time scaled in units of N generations [18, 19]. In our proofs, we assume this weak convergence for all $k \leq n$. Henceforth, we consider the positive real-valued T_k in units of N generations.

[3] Varying population sizes are implemented through changes to the coalescent process [17].

2.2. 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 [16] derives that the genetic distance until a 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[4].

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 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)[5].

Figure 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 $W_a := L_a + R_a \,|\, t \sim \text{Gamma}(2, t)$.

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

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

The segment length overlapping a focal location is necessarily longer than the segment length in one direction[6]. Because recombination 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[7].

2.3. The presence of detectable IBD segments

Relative to a focal point, we focus on the detection of long IBD segments in a sample. 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 $\text{Cov}_2, \text{Cov}_3$, 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. [11]. The detectable IBD rate is

$$\bar{\mathbf{X}}_{\binom{n}{2}} := \binom{n}{2}^{-1} \sum_{(a,b)} X_{a,b}. \tag{3}$$

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 $\mathbf{Z}_{-a,b} := \sum_{(c,d)} Z_{c,d} - Z_{a,b}$.

[6]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.

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

The sum of variances of all IBD segment indicators is

$$\Omega_{\binom{n}{2}} := \sum_{(a,b)} \mathrm{Var}(X_{a,b}) = \binom{n}{2} \times \mathbb{E}_2[X_{a,b}] \times (1 - \mathbb{E}_2[X_{a,b}]) \tag{4}$$

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

$$\bar{\mathbf{Z}}_{\binom{n}{2}} := \Omega_{\binom{n}{2}}^{-1/2} \times (\bar{\mathbf{X}}_{\binom{n}{2}} - \mathbb{E}_2[X_{a,b}]). \tag{5}$$

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. Supplementary Figure S1 provides a conceptual example calculating $\bar{\mathbf{Y}}$ for four sample haplotypes.

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[8]. 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 dwarfed by the variance of the random variables themselves is the basis of the general central

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

limit theorem for dependent data that is given in Chandrasekhar and Jackson [10] and Chandrasekhar et al. [11]. The following central limit theorems concern large sample size n and large population size N scaled by Morgan length threshold w [9].

Theorem 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. [11]. Without loss of generality, we derive integrals over a tree with two sample 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.

$$\mathbb{E}_2[X_{a,b}] = \int \exp(-2Nt_2w)\exp(-t_2)\,dt_2 = (2Nw+1)^{-1} = O((Nw)^{-1}). \tag{6}$$

It is easy to show that $\mathbb{E}_2[X_{a,b}] \to 0$ uniformly for large population size. The second condition implies that $\Omega_{\binom{n}{2}} \to \infty$. The assumption in Chandrasekhar et al. [11] 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 [10]. Lastly, given $n = o(Nw)$, we show that

$$\sum_{(a,b)\neq(c,d)} \mathrm{Cov}(\, X_{a,b}, X_{c,d}\,) = o(\Omega_{\binom{n}{2}}). \tag{7}$$

[9] 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 [14, 28, 35].

In Appendix A.1, 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.

$$\Omega_{\binom{n}{2}} \sim n^2 \cdot O((Nw)^{-1}) = o((Nw)^2) \cdot O((Nw)^{-1}) = o(Nw); \tag{8}$$

$$\sum_{a,b,c} \mathrm{Cov}(X_{a,b}, X_{a,c}) \sim n^3 \cdot O((Nw)^{-2}) = o((Nw)^3) \cdot O((Nw)^{-2}) = o(Nw); \tag{9}$$

$$\sum_{a,b,c,d} \mathrm{Cov}(X_{a,b}, X_{c,d}) \sim n^4 \cdot O((Nw)^{-3}) = o((Nw)^4) \cdot O((Nw)^{-3}) = o(Nw). \tag{10}$$

Therefore, the covariance across IBD segment indicators is controlled by the covariance within IBD segment indicators (Equation 7). □

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 location have strictly non-negative covariance, which we show in Appendix A.1.

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.1. 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.1), concluding that this condition might hold in large samples.

The asymptotic normality of $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ follows from same arguments as those of the proof in Theorem 3.1. We show in Appendix A.1 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 3.2. *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 3.1. 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 at 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 at 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. *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 3.1, except we use N_1 and N_2 to upper and lower bound covariance terms.

$$\Omega_{\binom{n}{2}} \sim n^2 \cdot O((N_2 w)^{-1}) = o(N_2 w); \tag{11}$$

$$\sum_{a,b,c} \mathrm{Cov}(X_{a,b}, X_{a,c}) \sim n^3 \cdot O((N_2 w)^{-2}) = o(N_2 w); \tag{12}$$

$$\sum_{a,b,c,d} \mathrm{Cov}(X_{a,b}, X_{c,d}) \sim n^4 \cdot O((N_2 w)^{-3}) = o(N_2 w). \tag{13}$$

□

Theorem 3.1 is a special case of Theorem 3.3 when $N_1 = N_2$. The conditions in Theorem 3.3 are unlikely to hold in real data examples, and they are more difficult to interpret. Note that the proof of Theorem 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 at a focal location the most. As is the case in Theorem 3.2, we can extend Theorem 3.3 to address IBD segments overlapping a focal location.

4. Identity-by-descent segment lengths drawn from the coalescent tree

Simulating IBD segments overlapping a focal point is one way to evaluate the distributional behavior of the detectable IBD rate, and functions thereof, in finite samples. A naive algorithm is to simulate a coalescent tree for a sample from an arbitrary population, draw recombination endpoints to the left and right of a focal point at each coalescent event, and then calculate IBD segment lengths. Making pairwise comparisons of recombination endpoints is quadratic in sample size and therefore prohibitive in large samples. Meanwhile, testing for normality requires enormous simulations to generate empirical distributions of the IBD rate around a locus.

Temple et al. [29] propose two modifications to the naive algorithm which they claim improves runtime in their simulations. "Pruning" is when a sample haplotype is not considered for IBD segment calculation at coalescent events if its haplotype segment length is less than a threshold. We elaborate in Appendix A.3.2 on the rare probability of long haplotype segments in large populations. "Merging" is when two sample haplotypes are combined for IBD segment calculation at coalescent events if they share the exact same left and right recombination

endpoints. After enough time, many recombination endpoints are descendant from the same common ancestor. We derive results in Appendix A.3.3 concerning the probability of merging.

In Algorithm 1, we formally state the method to simulate detectable IBD segments at a single locus. We also offer two more modifications that improve the runtime for the coalescent process: the discrete-time Wright-Fisher model can be approximated when there is likely to be more than one coalescent event in a generation (Appendix A.3.1), and the discrete-time Wright-Fisher process can be exchanged for the basic coalescent once number of non-coalesced haploids is much smaller than the effective population sizes. The latter implementation is similar to the hybrid simulation approach in Bhaskar et al. [2].

4.1. Simulation study setup

Our main empirical studies involve sample sizes increasing from one to ten thousand individuals and constant population sizes increasing from ten thousand to ten million individuals. The sample and population individuals are "diploids", which we implement as a haploid model multiplied by a factor of 2. We also explore the effects of demography and selection on the IBD segment lengths at a locus with respect to Algorithm 1's runtime.

The demographic scenarios are described in Temple et al. [29], which they refer to as examples of three phases of exponential growth and a population bottleneck. Supplementary Figure S2 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. The population bottleneck scenario involves an ancestral population of

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
   (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 $t \pm 1$
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 $\tau \pm 1$
   (**Option:** Use the Kingman coalescent if $k \ll N(t)$ for all remaining t.)

ten thousand individuals that grew exponentially at a fixed rate but experienced an instantaneous reduction in size at twenty generations from the present-day.

The genetic model for selection is described in Temple et al. [29]. 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.

4.2. Empirical runtime in simulating IBD segment lengths

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 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.

Supplementary 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 population bottleneck scenario, and $s = 0.04$.

Overall, we benchmark that Algorithm 1 can be run hundreds of times within a day on one core processing unit of an Intel 2.2 GHz compute node. Despite performance savings, we observe that Algorithm 1 maintains quadratic behavior in sample size (Figure 2 and Supplementary 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 [2].

4.3. Coalescent time conditional on IBD 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 marginal notion of the conditional coalescent time of sample haplotypes a and b is to solve for t such that $\mathbb{E}[W_{a,b}] = 2t^{-1} = w$. 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.

Supplementary 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 millions diploids. The distributions

appear similar to the Gamma family distributions (which is not surprising[10]).

Supplementary 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 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.

Supplementary Figure S6 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.

5. Empirical studies

We would be limited in our ability to study the distributional behavior of the standardized detectable IBD rate $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ overlapping a focal location if not for Al-

[10]For the first event time of sample haplotypes a and b, the conditional coalescent time $NT_n \,|\, W_{a,b} \geq w$ 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.

gorithm 1. To investigate normality, we require massive simulations to form tens of thousands of empirical distributions. The speed of Algorithm 1 is more important to this study on distributions than it is to the coverage studies of bootstrap intervals in Temple et al. [29].

5.1. The identity-by-descent rate in finite samples

5.1.1. Hypothesis testing for normality

Recall that $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ is shown to be asymptotically normal under regularity conditions on sample size and scaled population size (Theorem 3.2). Using the Shapiro-Wilk test, we investigate if empirical distributions of $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ resemble normal distributions as sample size, population size, and the Morgan length threshold 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 normal. We report the proportion of times we reject the null hypothesis at the 0.05 confidence level.

Figure 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. Supplementary Figure S7 shows that this trend does not depend on the confidence level. These observations align with the condition $n = o(Nw)$ in Theorem 3.1 and Theorem 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 3.1 and Theorem 3.2 (there are too few observed IBD segments).

Supplementary Figure S8 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 has no effect on the marginal correlations of IBD segments (Lemmas A.3, A.4, and A.5).

Supplementary Figure S9 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. 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 though the proportions of rejected tests around 0.3 and 0.1 are not close to the nominal significance level 0.05. Additionally, these proportions are less than their corresponding proportions in the population of twenty-five thousand individuals (Figure 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 (Supplementary Figure S2). While the conditions in Theorem 3.3 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.

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. Nevertheless, the trends of increasing sample size and scaled population size suggest the fidelity of our central limit theorems.

5.1.2. Genome-wide scans based on identity-by-descent rates

Excess IBD rate overlapping a focal location may indicate genomic regions under strong and recent positive selection [5, 29]. Let $\bar{\mathbf{Y}}_{1:M}$ be the detectable IBD rate overlapping M focal locations along a genome. The genome-wide sample mean and standard deviations are

$$\hat{\mu} := M^{-1} \sum_{m=1}^{M} \bar{\mathbf{Y}}_m; \tag{14}$$

$$\hat{\sigma} := \sqrt{(M-1)^{-1} \sum_{m=1}^{M} (\bar{\mathbf{Y}}_m - \bar{\bar{\mathbf{Y}}})^2}. \tag{15}$$

Browning and Browning [5] and Temple et al. [29] report regions where the detectable IBD rate overlapping a focal location exceeds $\hat{\mu}+4\hat{\sigma}$ as putative loci under strong and recent selection.

When Theorem 3.3 holds, we can interpret each such marginal hypothesis test as a one-sample one-sided $t-$test. Under this null hypothesis, we can compute p-values for each $\bar{\mathbf{Y}}_m$. We evaluate the control of Type 1 error in finite samples from finite populations. Using the same simulated data as in Section 5.1.1, Figure 4 shows the average upper bound $\hat{\mu} + 4 \times \hat{\sigma}$ divided by the 99.99683 percentile over two million simulations, where the Gaussian cumulative distribution function $\Phi(4) = 0.9999683$. Supplementary Figure S10 shows the same as in Figure 4 ex-

cept 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 closer to the percentile threshold as population size and Morgan length threshold increase, which is a result consistent with Section 5.1.1 and our central limit theorems. Supplementary Figure S9 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. Supplementary Figure S11 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[11]. 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

[11]Ignoring spatial correlation of tests

be if Type 1 error control 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.

5.2. Identity-by-descent graphs at a locus

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. 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 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 5.1.1.

5.2.1. 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 [12]. 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 [25]. 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 [13]. 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 [13].

In a sample size of five thousand from a population of one hundred thousand individuals, Figure 5 shows that some empirical distributions of graph features resemble normal distributions. Table 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[12]. 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 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.

[12]The expected number of edges should be the same, if not for the approximation [25].

5.2.2. The impact of complex demograpy

Supplementary Figure S12 shows that the apparent normality of some graph features extends to the three phases of exponential growth and population bottleneck demographic scenarios. Supplementary Table S1 reports that proportions of rejected hypothesis tests for numbers of trees of order 2 are close to 0.05 for both demographic scenarios. We also cannot 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 at a locus can be similar for large constant size populations and demographic scenarios with large recent population sizes.

5.2.3. The impact of strong positive selection

Strong directional selection increases the IBD rate [29], 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 selections coefficients between 0.01 and 0.04 and the three phases of exponential growth and population bottleneck scenarios. Supplementary 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 growth of a largest detectable IBD cluster that

includes haplotypes with a beneficial allele. This idea is a major motivation for the suite of methods developed in Temple et al. [29].

6. Discussion

In this article, we leverage ideas from coalescent theory and haplotype sharing to develop statistical theory and motivate methodology in IBD-based inference. Most notably, we prove a central limit theorem for the detectable IBD rate at 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 not large IBD clusters about a locus.

The conceptual framework for these 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.

Deriving asymptotic results can be difficult in genetics due to the dependencies among even distantly related individuals. Our proof strategy is to control the sources of covariance that arise from two waiting time distributions. The scaled population size is the term determining weak marginal correlations between IBD segment lengths. More meioses separate sample haplotypes the larger the population size is. Recombination serves to reduce the expected length of ancestral haplotypes with each meiosis. The techniques we use might be generally useful to

other studies involving these processes.

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 can lead to anti-conservative control of Type 1 error, we observe no trend between this anti-conservative behavior and sample size.

Our empirical studies amount to hundreds of millions of simulations. A heuristic yet efficient and exact algorithm to simulate long IBD segments has been proposed in prior work [29] and improved upon in this work. We provide probabilistic arguments that give evidence to the approximately linear, not quadratic, runtime of the algorithm. We benchmark that simulating a joint distribution of detectable IBD segment lengths at a locus can take as little as a couple seconds for sample sizes of order 10^4. The method for a single locus is measured in seconds to minutes whereas existing methods for entire chromosomes can take hours [1, 24]. 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.

Together our theoretical results and simulation studies support ongoing methodological developments based on IBD segments. For instance, the confidence intervals in Temple et al. [29] should give proper coverage of the selection coefficient if the detectable IBD rate is approximately normally distributed in scenarios of positive selection and the Delta method conditions hold for the function that maps the detectable IBD rate to the selection coefficient [9]. Additionally, existing genome-wide scans for excess IBD rate lack formal or exact hypothesis testing frameworks. The IBD-based selection scans in Browning and Browning [5] and Temple et al. [29] use a heuristic threshold of four standard deviations to determine significance. Browning and Thompson [7] use an approximate Markov model for the IBD rate at a spatial position to associate rare diseases with the difference in IBD rate between case-case pairs and case-control pairs, adjusting for multiple testing either heuristically or by computationally expensive permutations of their test statistic. 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 can be assumed to be Gaussian, and their correlation is a consequence of the same IBD segments overlapping spatially adjacent locations on a chromosome. Future work could extend the ideas presented here to derive a genome-wide significance threshold for IBD-based selection scans and rare variant association studies.

Acknowledgements

Computational resources were supported by the National Human Genome Research Institute of the National Institutes of Health under award number HG005701 and the Department of Statistics at the University of Washington.

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.

We thank Professor Sharon Browning for helpful discussions.

Software Resources

We use the Python package https://github.com/sdtemple/isweep for all simulation studies. This software is freely available under the open source CC0 1.0 Universal License.

Author Contributions

S.D.T. proposed the study, planned the study, wrote the software, conducted the simulation studies, and wrote the manuscript. Both authors worked on the theoretical results and contributed to editing the manuscript.

Declaration of Interests

The authors declare no competing interests.

References

[1] F. Baumdicker, G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, B. Eldon, E. C. Ellerman, J. G. Galloway, A. L. Gladstein, G. Gorjanc, B. Guo, B. Jeffery, W. W. Kretzschumar, K. Lohse, M. Matschiner, D. Nelson, N. S. Pope, C. D. Quinto-Cortés, M. F. Rodrigues, K. Saunack, T. Sellinger, K. Thornton, H. van Kemenade, A. W. Wohns, Y. Wong, S. Gravel, A. D. Kern, J. Koskela, P. L. Ralph, and J. Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220 (3), Mar. 2022.

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

[3] B. L. Browning, X. Tian, Y. Zhou, and S. R. Browning. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.*, 108(10):1880–1890, Oct. 2021.

[4] S. R. Browning and B. 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, Sept. 2015.

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

[6] S. R. Browning and B. L. Browning. Biobank-scale inference of multi-individual identity by descent and gene conversion. Nov. 2023.

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

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

[9] G. Casella and R. L. Berger. Statistical inference. *Biometrics*, 49(1):320, Mar. 1993.

[10] A. G. Chandrasekhar and M. O. Jackson. A network formation model based on subgraphs. Nov. 2016.

[11] A. G. Chandrasekhar, M. O. Jackson, T. H. McCormick, and V. Thiyageswaran. General covariance-based conditions for central limit theorems with dependent triangular arrays. Aug. 2023.

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

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

[14] W. A. Freyman, K. F. Mcmanus, S. S. Shringarpure, E. M. Jewett, K. Bryc, and A. Auton. Fast and robust identity-by-descent inference with the templated positional Burrows–Wheeler transform. *Mol. Biol. Evol.*, 38(5):2131–2151, May 2021.

[15] A. Gusev, E. E. Kenny, J. K. Lowe, J. Salit, R. Saxena, S. Kathiresan, D. M. Altshuler, J. M. Friedman, J. L. Breslow, and I. Pe'er. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am. J. Hum. Genet.*, 88(6):706–717, June 2011.

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

[17] J. Hein, M. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory.* Oxford University Press, USA, Dec. 2004.

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

[19] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, Sept. 1982.

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

[21] P.-R. Loh, P. F. Palamara, and A. L. Price. Fast and accurate long-range phasing in a UK biobank cohort. *Nat. Genet.*, 48(7):811–816, July 2016.

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

[23] J. Nait Saada, G. Kalantzis, D. Shyr, F. Cooper, M. Robinson, A. Gusev, and P. F. 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, Nov. 2020.

[24] P. F. Palamara. ARGON: fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics*, 32(19):3032–3034, Oct. 2016.

[25] P. F. Palamara, T. Lencz, A. Darvasi, and I. Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.*, 91(5):809–822, Nov. 2012.

[26] P. F. Palamara, L. C. Francioli, P. R. Wilton, G. Genovese, A. Gusev, H. K. Finucane, S. Sankararaman, Genome of the Netherlands Consortium, S. R. Sunyaev, P. I. W. de Bakker, J. Wakeley, I. Pe'er, and A. L. Price. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am. J. Hum. Genet.*, 97(6):775–789, Dec. 2015.

[27] M. D. Ramstetter, T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala, J. Blangero, J. G. Mezey, and A. L. Williams. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, 207(1):75–82, 2017.

[28] R. Shemirani, G. M. Belbin, C. L. Avery, E. E. Kenny, C. R. Gignoux, and J. L. Ambite. Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nat. Commun.*, 12(1):3546, June 2021.

[29] S. D. Temple, R. D. Waples, and S. R. Browning. Modeling recent positive selection in americans of european ancestry. Nov. 2023.

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

[31] X. Tian, B. L. Browning, and S. R. Browning. Estimating the genome-wide mutation rate with three-way identity by descent. *Am. J. Hum. Genet.*, 105 (5):883–893, Nov. 2019.

[32] X. Tian, R. Cai, and S. R. Browning. Estimating the genome-wide mutation rate from thousands of unrelated individuals. *Am. J. Hum. Genet.*, 109(12): 2178–2184, Dec. 2022.

[33] J. Wakeley and T. Takahashi. Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.*, 20(2):208–213, Feb. 2003.

[34] Y. Zhou, B. L. Browning, and S. R. Browning. Population-specific recombination maps from segments of identity by descent. *Am. J. Hum. Genet.*, 107 (1):137–148, July 2020.

[35] Y. Zhou, S. R. Browning, and B. 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, Apr. 2020.

[36] Y. Zhou, S. R. Browning, and B. L. Browning. IBDkin: fast estimation of kinship coefficients from identity by descent segments. *Bioinformatics*, 36 (16):4519–4520, Aug. 2020.

Figures

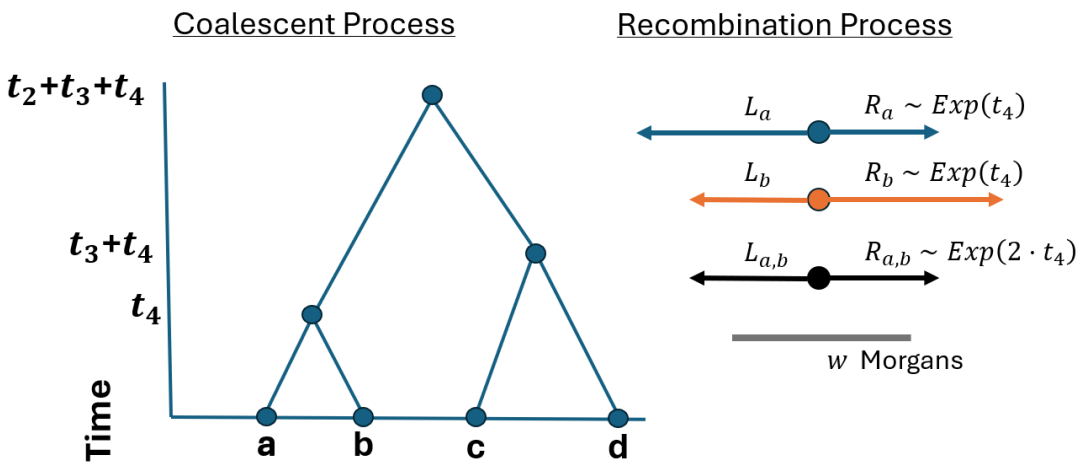

Figure 1: Conceptual framework for IBD segment lengths. (Left) Sample haplotypes a, b, c, d trace their lineages back to common ancestors at times $t_4, t_4 + t_3, t_4 + t_3 + t_2$. (Right) Relative to a focal point, the haplotype segments lengths R_a, R_b, L_a, L_b are independent, identically distributed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

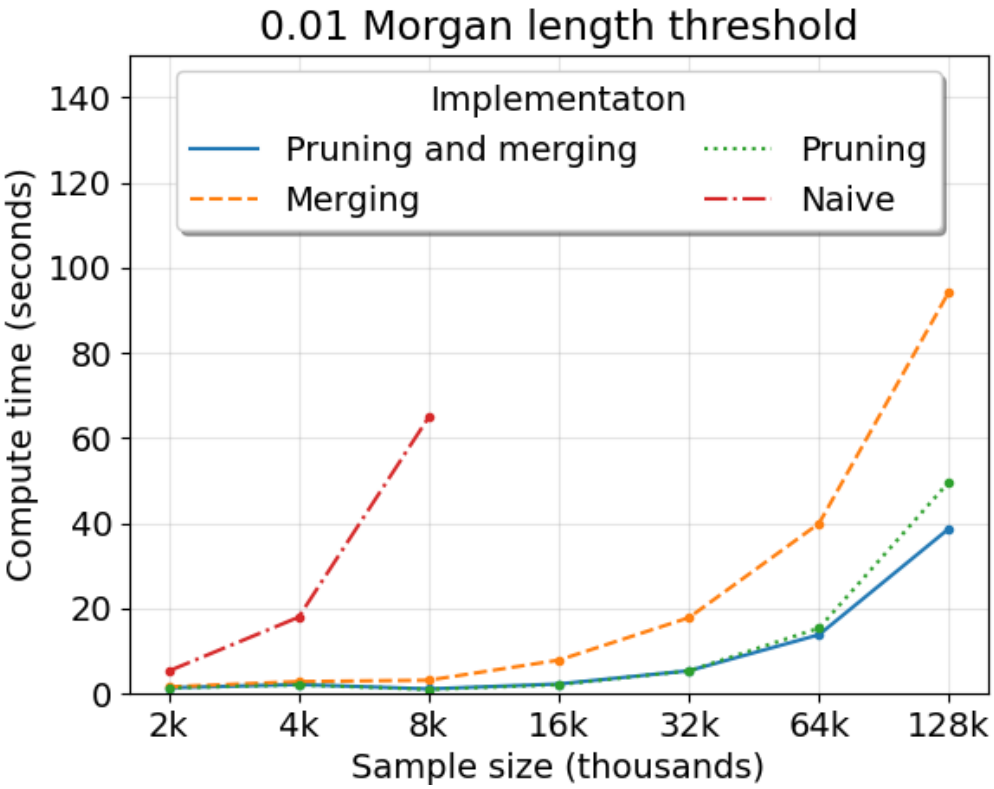

Figure 2: Compute time to simulate IBD segment lengths depending on algorithm implementation. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legend denotes colored line styles for implementations using Algorithm 1 as is, without pruning, without merging, and without pruning and merging. Main text describes "merging" and "pruning" techniques. The demography is the population bottleneck. The Morgan length threshold is 0.01.

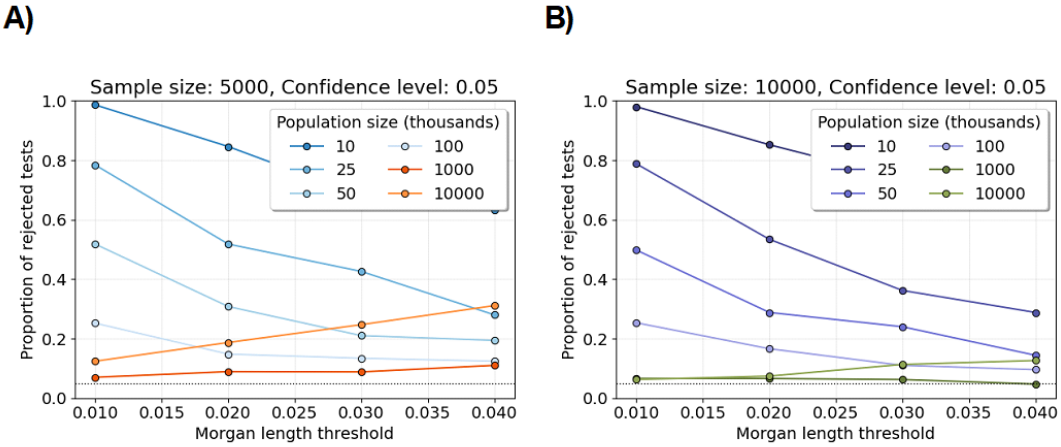

Figure 3: Shapiro-Wilk tests for varying population sizes. Line plots show the proportions of Shapiro-Wilk tests rejected at confidence level 0.05 (y-axis) for varying population size and fixed sample size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). A) The sample size is five thousand individuals. B) The sample size is ten thousand. The legends give colors assigned to different population sizes. The horizontal dotted line is at 0.05.

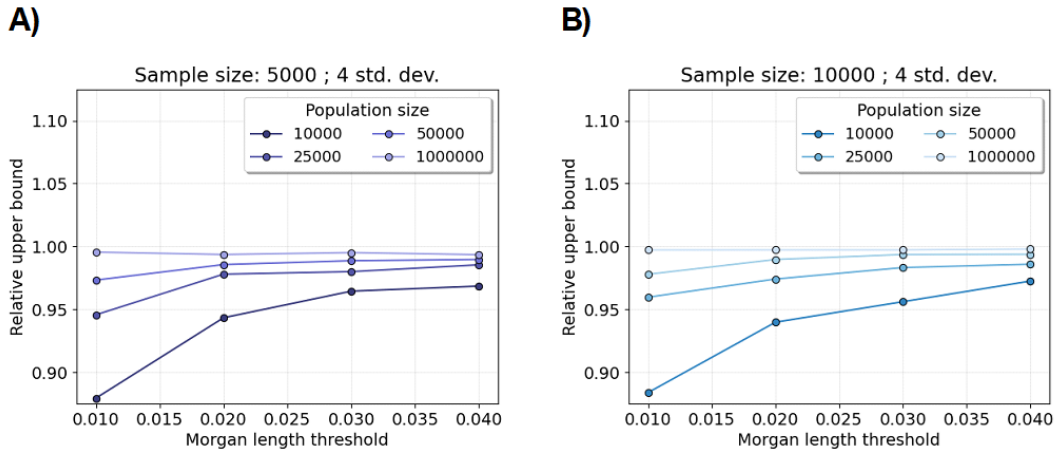

Figure 4: Relative upper bound for excess IBD scan. Line plots show the average mean plus four standard deviations divided by the 99.99683 percentile over two million simulations (y-axis). (The standard normal survival function of four is 0.9999683.) Each average relative upper bound is computed over one thousand tests. Each test is based on two thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). A) The sample size is five thousand diploid individuals. B) The sample size is ten thousand diploids individuals. The legends give colors assigned to different constant population sizes sizes.

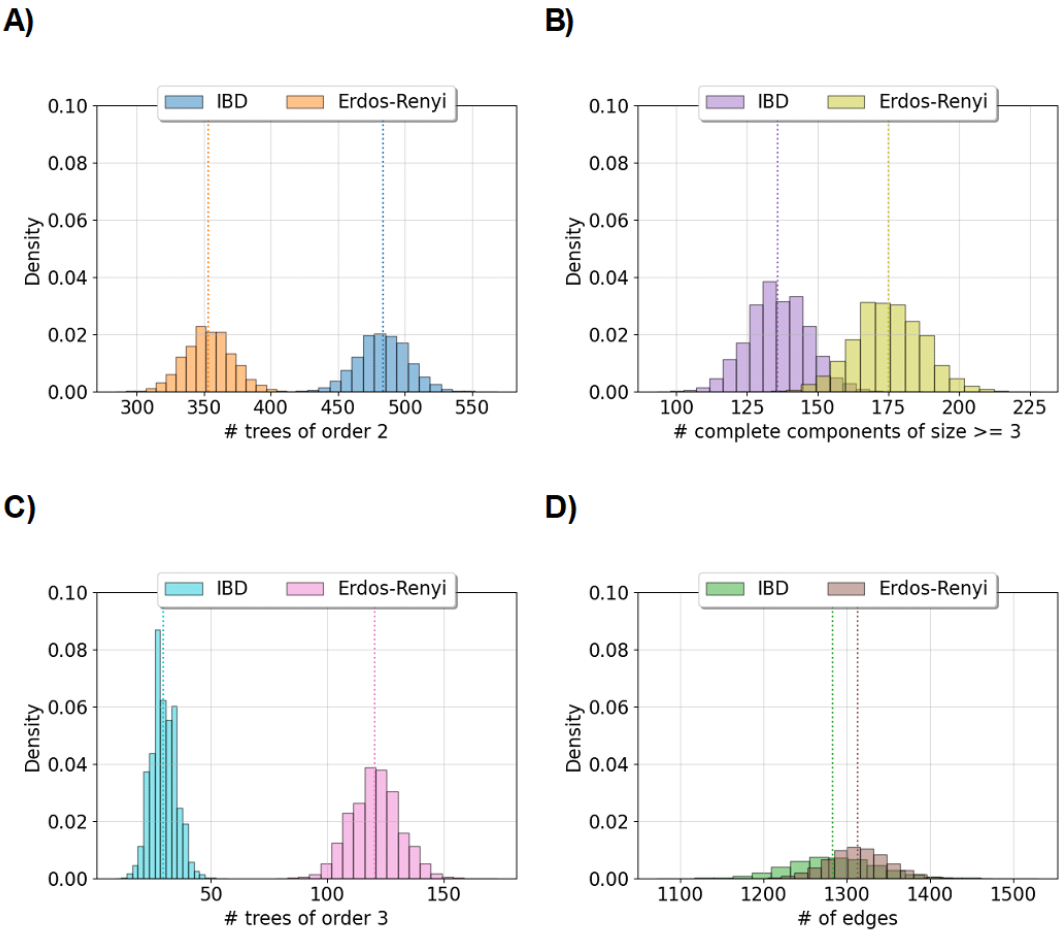

Figure 5: Comparing features between IBD and Erdős-Rényi graphs. Histograms compare the density of graph features between IBD and Erdős-Rényi graphs. Each histogram summarizes the results of one hundred and twenty-five thousand simulations. A) and C) show the number of trees of order 2 and 3, respectively. B) shows the number of complete components with more than three nodes. D) shows the total number of edges. The legends gives color assigned to the IBD and Erdős-Rényi graphs. IBD graphs are simulated using the constant one hundred thousand individuals 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 individuals. Vertical lines show the means.

Tables

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 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 confidence level 0.05 are performed with 500 replicates for 250 simulations, the proportion of rejected null hypotheses reported as S.W.t. The population size is one hundred thousand individuals. The sample size is two thousand individuals. The Morgan length threshold is 0.03.

Supplementary Figures

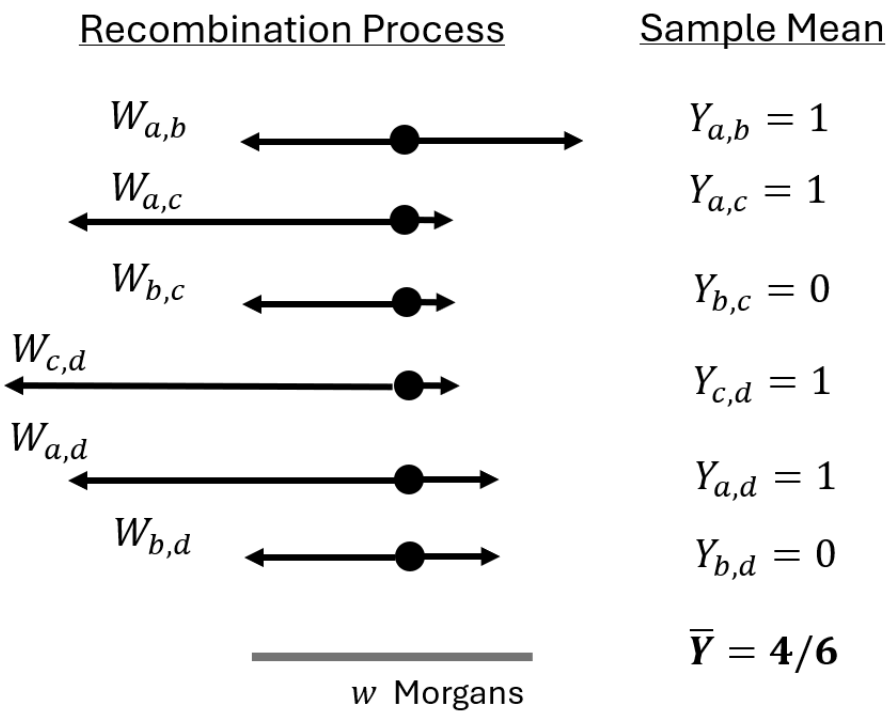

Figure S1: Example calculation of the detectable IBD rate. IBD segment lengths overlapping a focal point for sample haplotypes a, b, c, d are shown. The IBD segment indicators ($Y_{i,j}$'s) are 1 if their IBD segment lengths ($W_{i,j}$'s) exceed w Morgans and otherwise 0. The detectable IBD rate $\bar{\mathbf{Y}}$ 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.

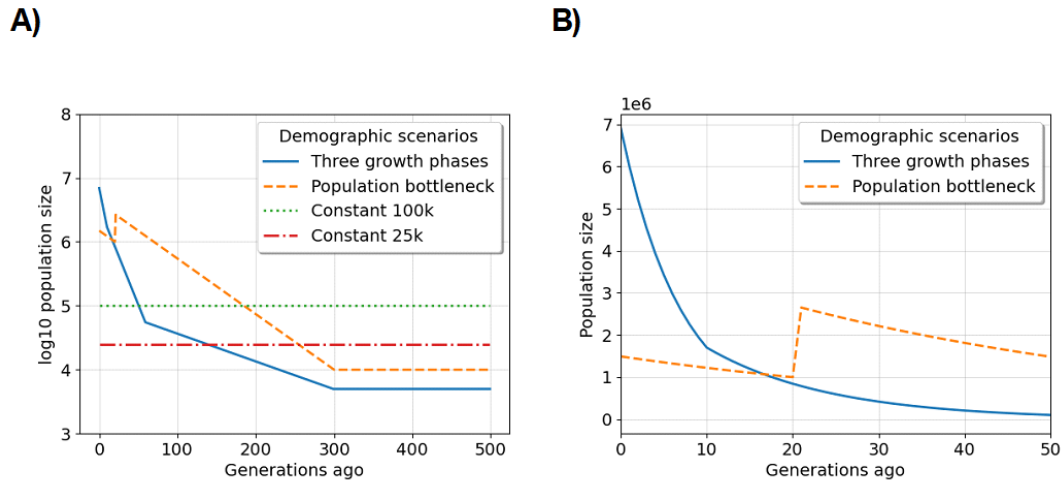

Figure S2: Demographic scenarios we consider in simulation studies: A) coalescent time in generations ago by the log 10 population size, and B) the most recent fifty generations by population size for examples of exponential growth. The legends specify the color and line style for each scenario. As opposed to coalescent time used in the main text, we describe here the scenarios forward in time. Three phases of exponential growth: 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. Population bottleneck: a population of ancestral size ten thousand diploids increases exponentially each generation at rate two percent starting three hundred generations ago. Otherwise, demographic scenarios we explore here are populations of constant size twenty five and one hundred diploids.

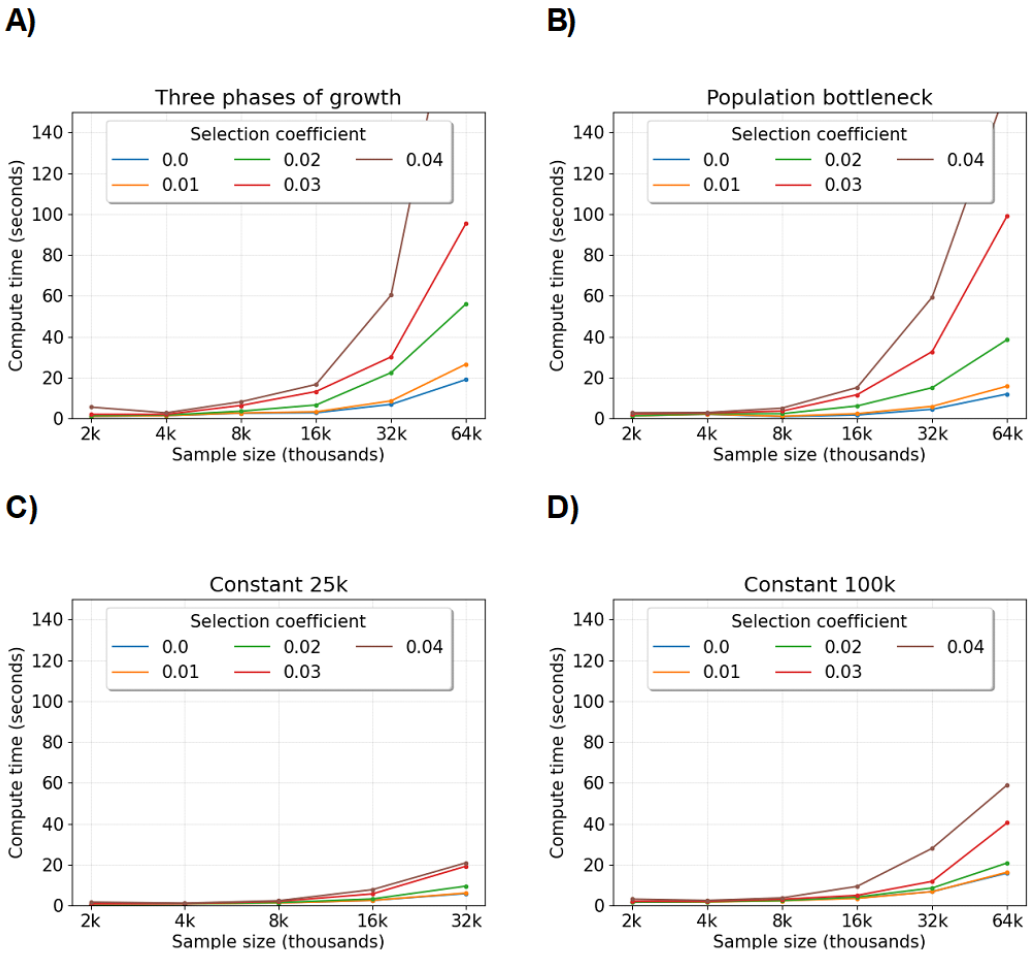

Figure S3: Compute time to simulate IBD segment lengths depending demography and selection. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legends denote colored line styles for different selection coefficients. A), B), C), and D) show results for demographic scenarios of three phases of exponential growth, a population bottleneck, and constant population sizes of twenty-five and one hundred thousand individuals, respectively. The Morgan length threshold is 0.01.

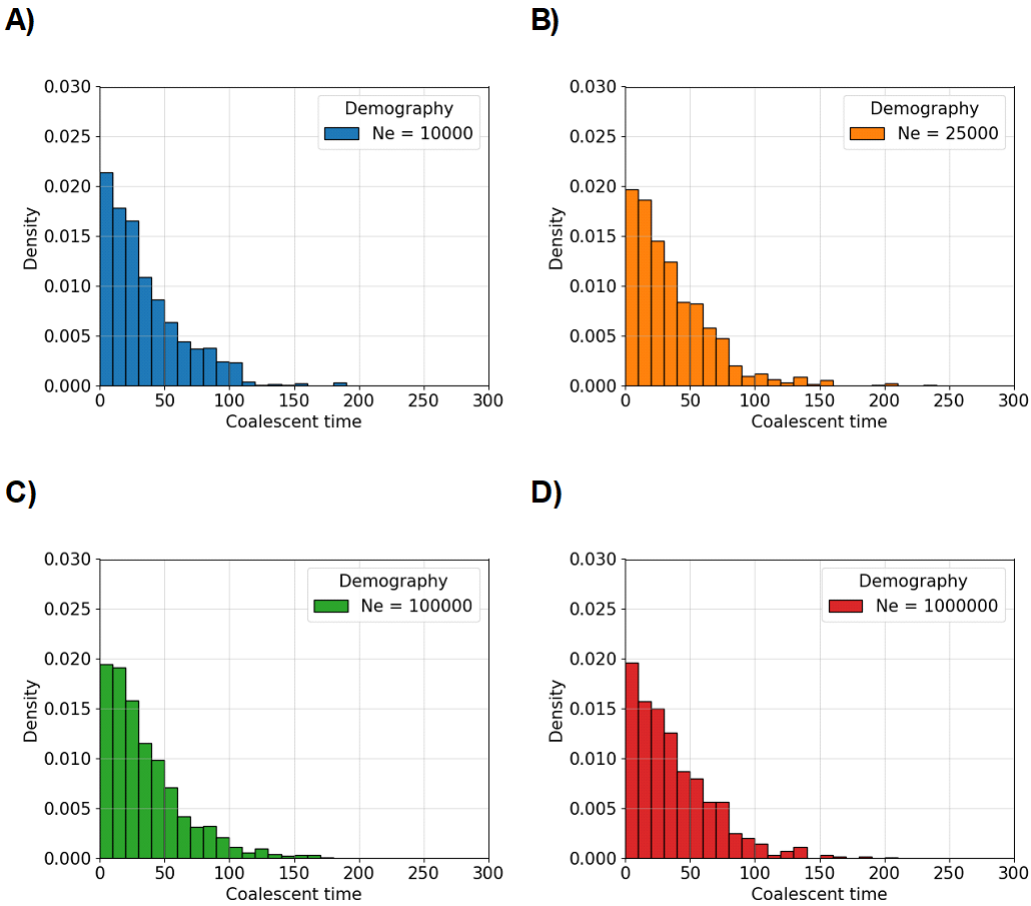

Figure S4: Empirical distributions of coalescent times conditional on IBD segments for varying population size. Histograms show the density of coalescent times split into fifty bins. Each panel represents one simulation. The population size is set to A) ten thousand, B) one hundred thousand, C) one million, and D) ten million individuals. The Morgan length threshold is 0.02. The sample size is five thousand individuals. There is no selection.

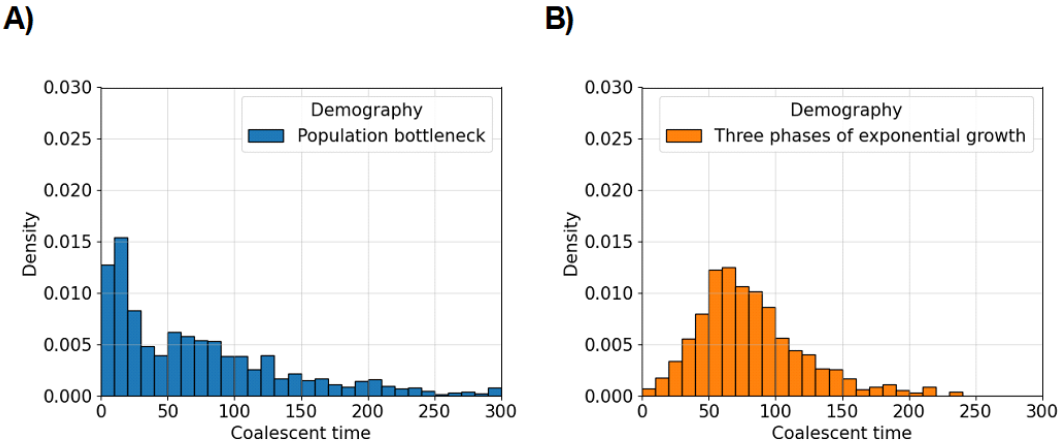

Figure S5: Empirical distributions of coalescent times conditional on IBD segments for different demographic scenarios. Histograms show the density of coalescent times split into fifty bins. Each panel represents one simulation. The demographic scenario is set to A) a population bottleneck and B) three phases of exponential growth. The Morgan length threshold is 0.02. The sample size is five thousand individuals. There is no selection.

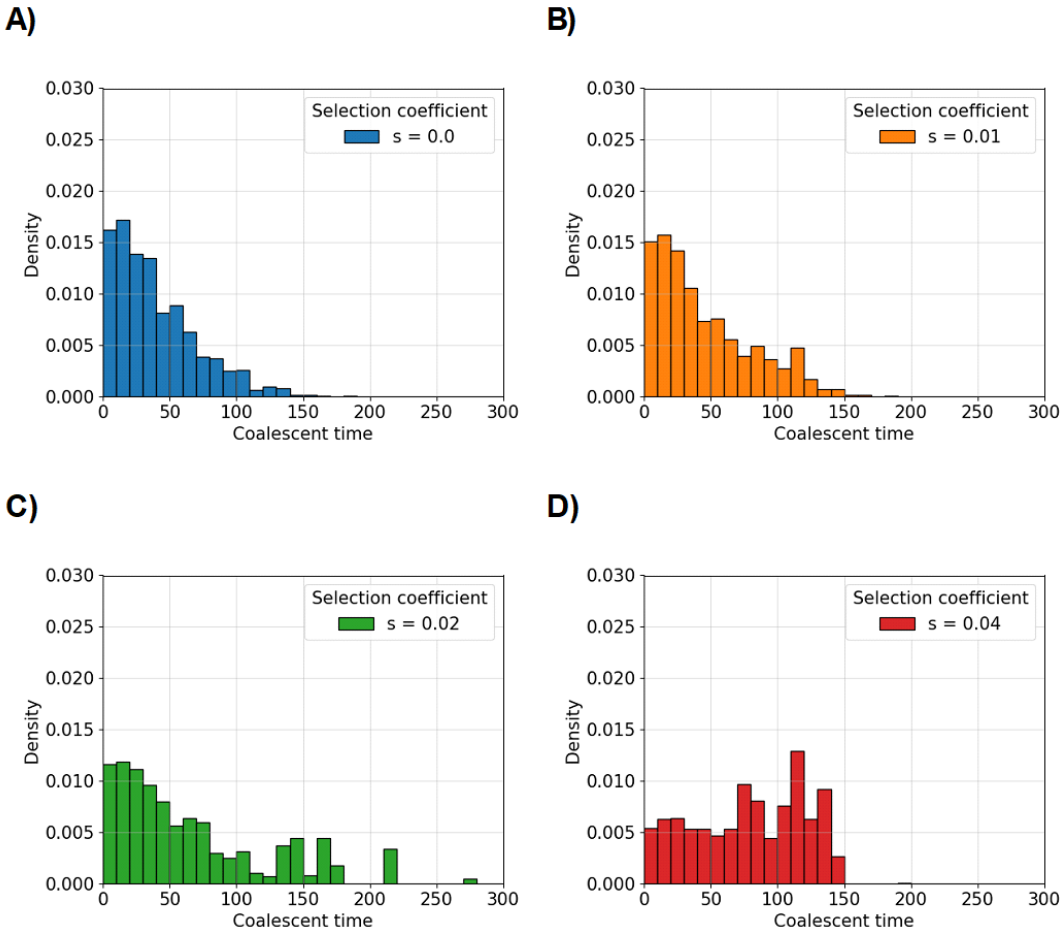

Figure S6: Empirical distributions of coalescent times conditional on IBD segments for different strength of selection and low common allele frequency. Histograms show the density of coalescent times split into fifty bins. Each panel represents one simulation. The selection coefficient is set to A) s=0.00, 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 individuals. The sample size is five thousand individuals.

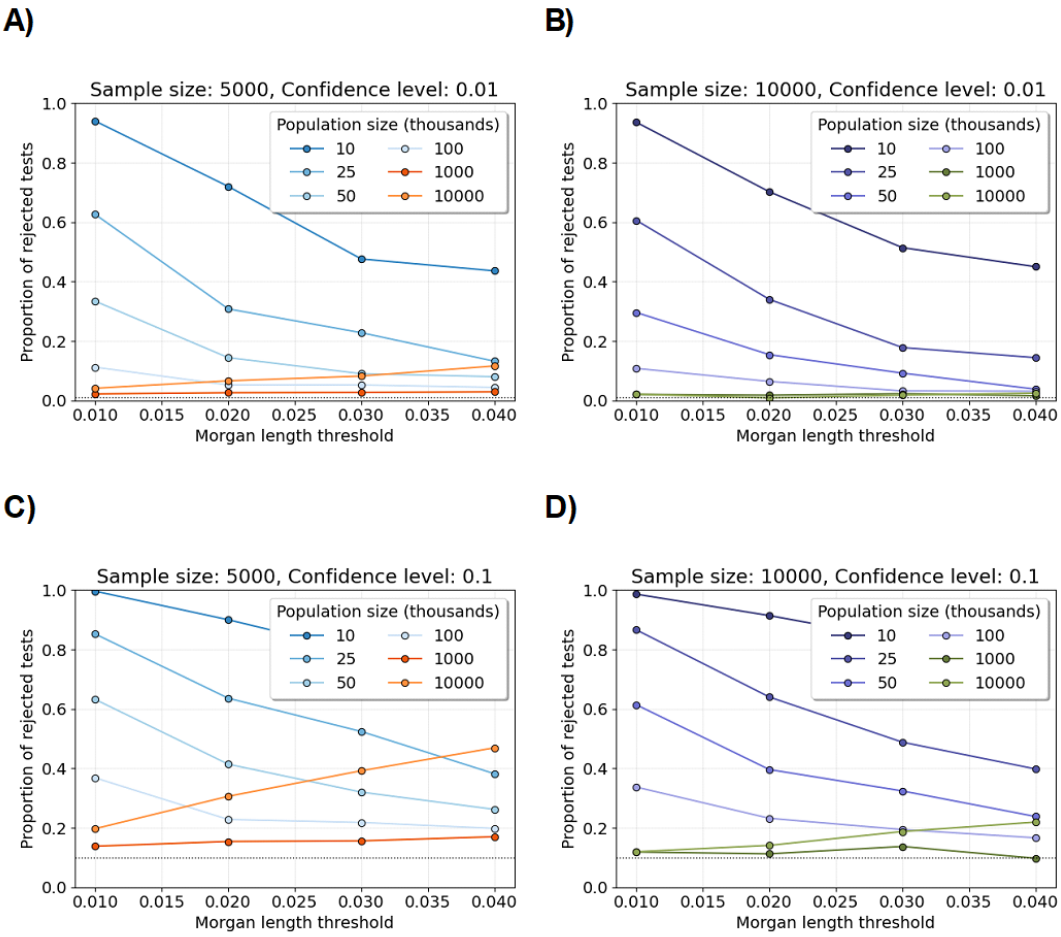

Figure S7: Shapiro-Wilk tests for varying population sizes and confidence levels. Line plots show the proportions of Shapiro-Wilk tests rejected at confidence levels A,B) 0.01 and C,D) 0.1 (y-axis) for varying population size and fixed sample size. Each proportion is computed over five hundred tests. 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 individuals. B,D) The sample size is ten thousand individuals. The legends give colors assigned to different population sizes. The horizontal dotted lines are confidence levels.

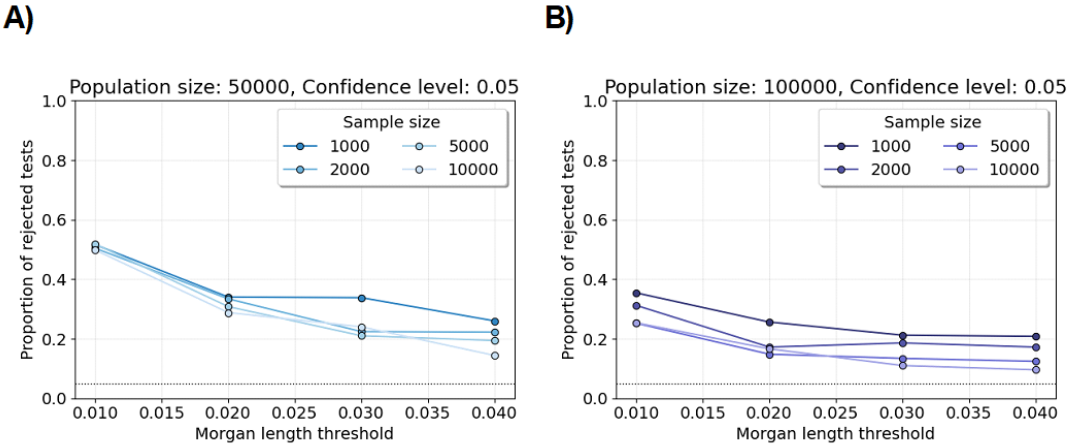

Figure S8: Shapiro-Wilk tests for varying sample sizes. Line plots show the proportions of Shapiro-Wilk tests rejected at confidence level 0.05 (y-axis) for varying sample size and fixed population size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgan length threshold (x-axis). A) The population size is fifty thousand individuals. B) The population size is one hundred thousand individuals. The legends give colors assigned to different sample sizes. The horizontal dotted line is at 0.05.

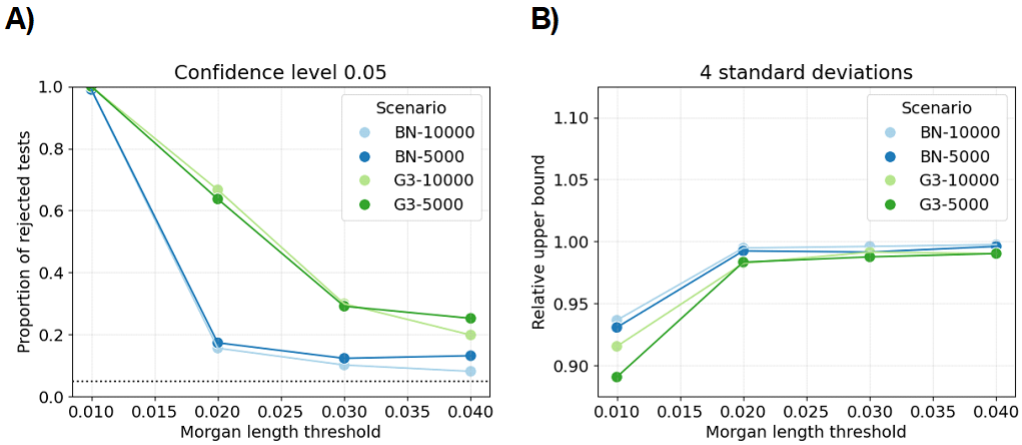

Figure S9: Shapiro-Wilk tests and relative upper tail bounds for complex demography scenarios. A) Line plots show the proportions of Shapiro-Wilk tests rejected at confidence level 0.05 (y-axis) for the population bottleneck (BN) or three phases of exponential growth (G3) demographic scenarios and sample sizes of five or ten thousand individuals. 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 standard deviations divided by the 99.99683 percentile over two million simulations (y-axis). Plot designs are identical to Figures 3 and 4.

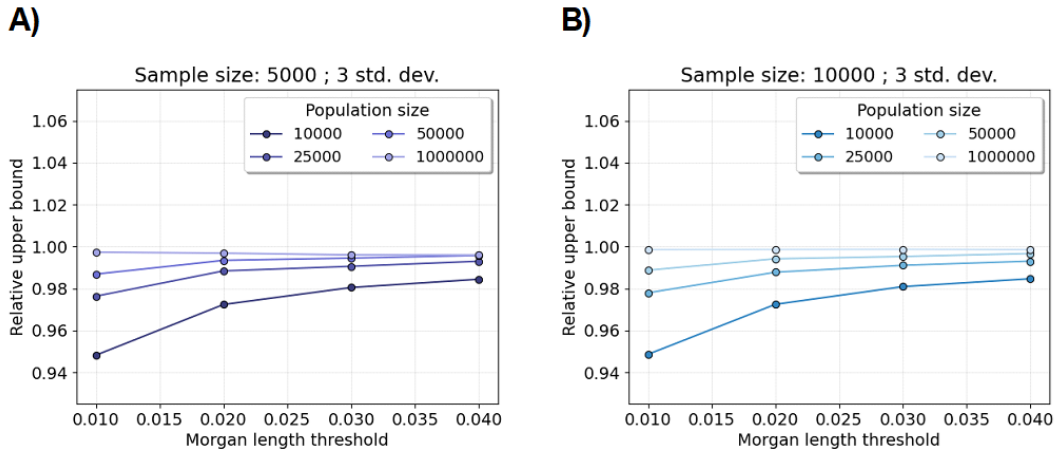

Figure S10: Relative upper bound for excess IBD scan. Line plots show the average mean plus three standard deviations divided by the 99.86501 percentile over two million simulations (y-axis). (The standard normal survival function of three is 0.9986501.) Each average relative upper bound is computed 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 diploid individuals. B) The sample size is ten thousand diploids individuals. The legends give colors assigned to different constant population sizes sizes.

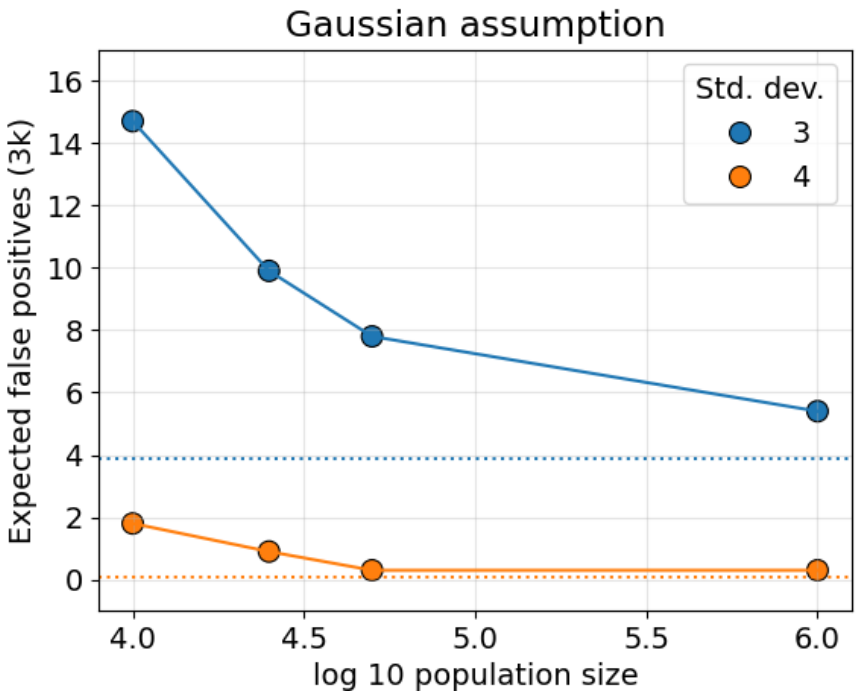

Figure S11: Expected false positives for excess IBD scan. Line plots show the expected false positive tests out of three thousand (x-axis) by the log 10 constant population size (y-axis). The z-tests use three (blue) or four (orange) estimated standard deviations above the estimated mean. The horizontal 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.

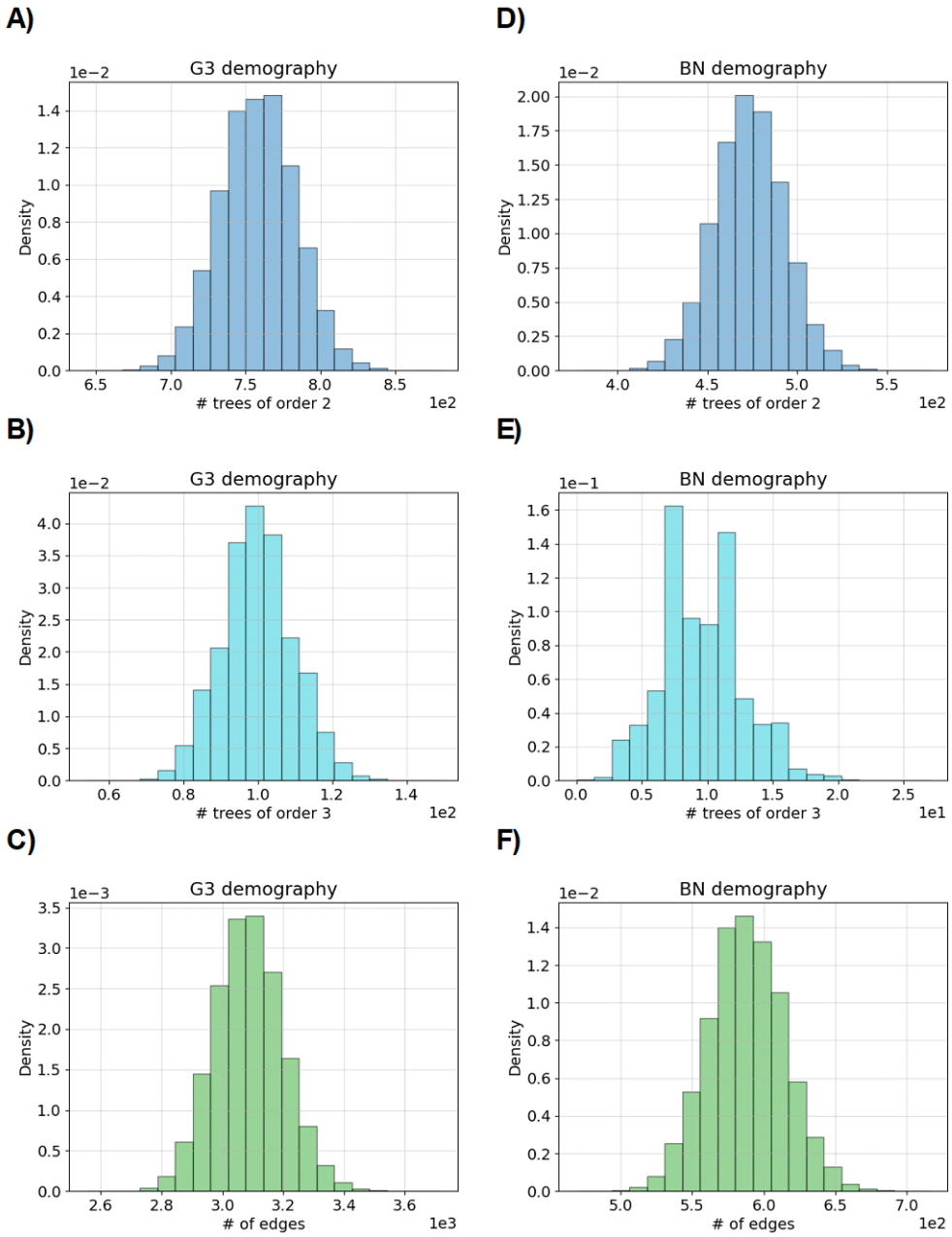

Figure S12: Comparing features between IBD graphs for complex demographic scenarios. Histograms show the density of IBD graph features between A-C) the three phases of exponential growth (G3) and D-F) the population bottleneck (BN) demographic scenarios. Each histogram is based on at least six hundred 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 individuals.

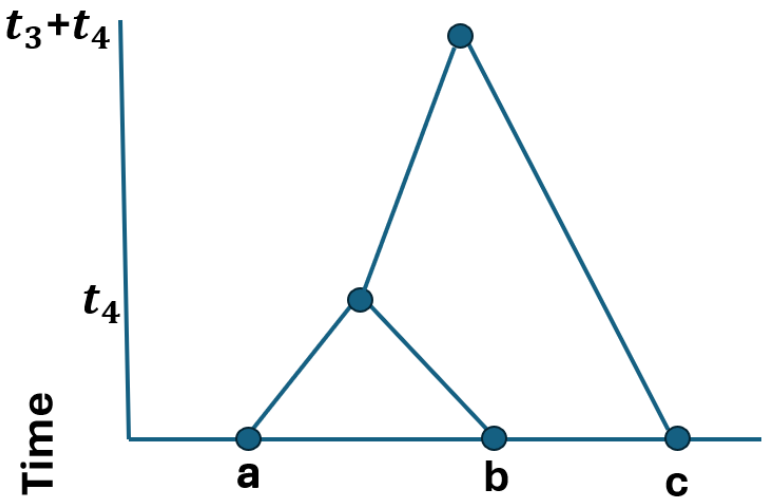

Figure S13: Illustration of the one possible coalescent tree used to calculate Cov_3 terms in Appendix A.1.

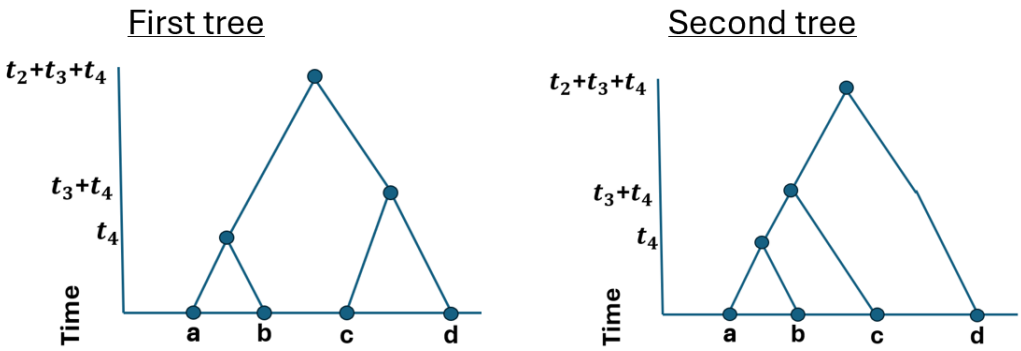

Figure S14: Illustration of the two possible coalescent trees used to calculate Cov_4 terms in Appendix A.1.

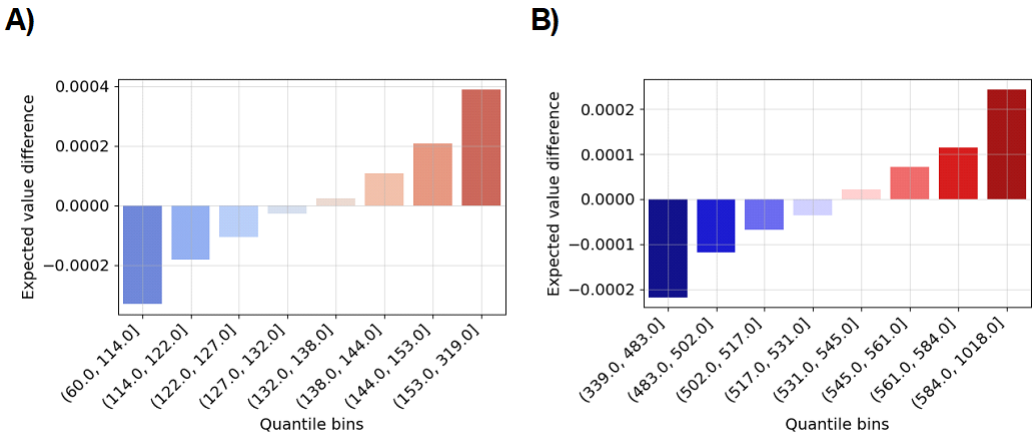

Figure S15: Monte Carlo verification of the conditional expectation condition in our central limit theorem. Bar charts show the difference between the proportion of simulations where two specific haplotypes share an IBD segment longer than 0.03 Morgans and the true success probability (y-axis). This 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 individuals. The population size is ten thousand individuals. The expectation is 132.78 in A) and 531.78 in B).

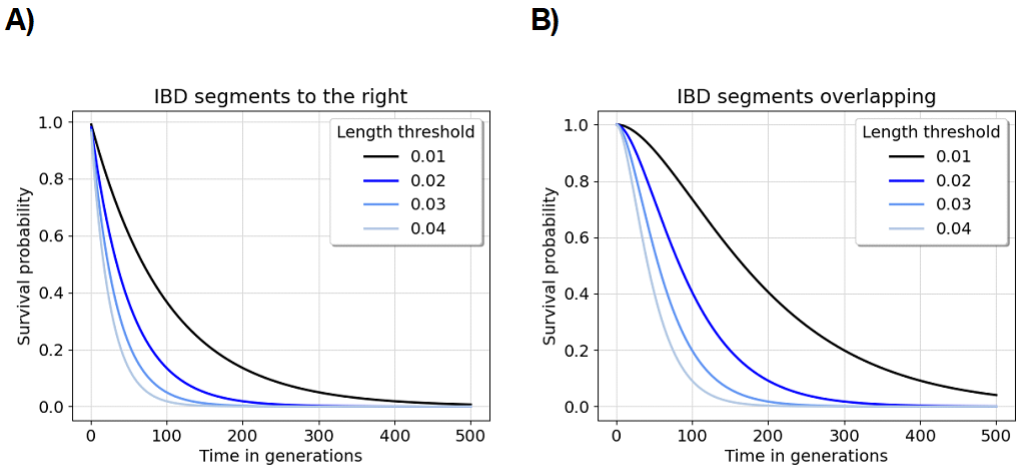

Figure S16: The survival probabilities of Erlang random variables. Subplots A) and B) show the survival probabilities for shape parameters 1 and 2, respectively. The rate of the random variables is the coalescent time in generations (x-axis). The survival probability (y-axis) comes from Equations A.4 and A.5. The length thresholds are denoted by different colors and line styles, defined in the legend.

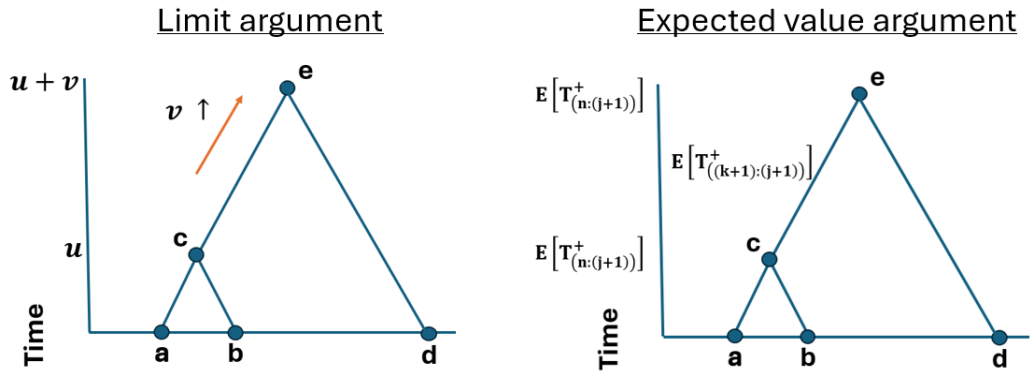

Figure S17: Illustration of coalescent tree branch lengths for merging arguments. (Left) The branch length connecting common ancestor c to its common ancestor e increases to infinity. (Right) The branch lengths of the tree are the expected values of time after $(k+1)$, $(j+1)$ coalescent events.

Supplementary Tables

Type	Structure	Avg	Var	Min	Max	S.W.t.
G3	Edges	3,085.66	12,827.06	2,554.00	3,716.00	0.29
	Largest	15.60	9.22	9.00	50.00	1.00
	Tree2	757.96	635.16	644.00	880.00	0.08
	Tree3	99.71	94.29	54.00	149.00	0.42
	Complete	251.55	230.65	185.00	3,118.00	0.14
BN	Edges	587.73	694.55	469.00	716.00	0.12
	Largest	4.32	0.39	3.00	11.00	1.00
	Tree2	473.24	393.50	377.00	574.00	0.09
	Tree3	9.66	9.57	0.00	27.00	1.00
	Complete	35.53	34.76	11.00	488.00	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 thousands simulations. Shapiro-Wilk tests at confidence level 0.05 are performed with 1000 replicates for at least 600 simulations, the proportion of rejected null hypotheses 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	2,916.00	4,143.00	0.36
	Largest	**24.33**	50.80	11.00	89.00	0.97
	Tree2	737.77	626.12	636.00	842.00	0.05
	Tree3	95.81	92.23	57.00	138.00	0.07
	Complete	242.41	215.76	187.00	305.00	0.05
$s = 0.02$	Edges	4,693.51	140,436.48	3,579.00	8,212.00	0.95
	Largest	**73.97**	1,219.95	22.00	346.00	0.97
	Tree2	697.19	588.38	596.00	791.00	0.10
	Tree3	86.65	83.70	53.00	126.00	0.09
	Complete	220.37	199.88	161.00	281.00	0.10
$s = 0.03$	Edges	8,242.12	2,283,864.57	4,998.00	37,933.00	0.97
	Largest	**230.39**	12,224.19	39.00	819.00	0.97
	Tree2	659.10	565.21	562.00	759.00	0.07
	Tree3	78.43	74.69	46.00	119.00	0.11
	Complete	199.95	181.88	145.00	254.00	0.06
$s = 0.04$	Edges	16,486.56	24,295,227.62	7,747.00	72,775.00	0.97
	Largest	**484.92**	38,683.32	89.00	1,229.00	0.97
	Tree2	630.68	529.35	546.00	731.00	0.02
	Tree3	72.95	70.26	41.00	108.00	0.11
	Complete	185.76	167.85	135.00	241.00	0.07

Table S2: Summary statistics of IBD graphs for different selection coefficients and the three phases of exponential growth demographic scenario. There is directional selection with different selection coefficients $s \in [0.01, 0.02, 0.03, 0.4]$. Same description of IBD graph features as in Table 1. Shapiro-Wilk tests at confidence level 0.05 are performed with 250 replicates for 150 simulations, the proportion of rejected null hypotheses 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	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 $s \in [0.01, 0.02, 0.03, 0.4]$. Same description of IBD graph features as in Table 1. Shapiro-Wilk tests at confidence level 0.05 are performed with 250 replicates for 150 simulations, the proportion of rejected null hypotheses reported as S.W.t. The sample size is five thousand individuals. The Morgan length threshold is 0.03.

A.1. Derivations of theoretical results

A.1.1. Theorem 3.1 and its extensions

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

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 reordering of three sample haplotypes, there is one possible bifurcating tree (Supplementary Figure S13). 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 reordering of four sample haplotypes, there are two possible bifurcating trees (Supplementary Figure S14). 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.

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

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

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

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 marginal 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 marginal expected values in Equation 6 and Lemmas A.3 and A.4.

Calculating $\mathbb{E}_3[Y_{a,b}Y_{a,c}]$ is more involved. Up to 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 $\mathbb{E}[Z_{a,c} \times \mathbf{Z}_{-a,c} | \mathbf{Z}_{-a,c}] \ngeq 0$ 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 of the central limit theorem

The third condition in Theorems 3.1, 3.2, and 3.2 amounts to a notion of joint nonnegative 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 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 S15 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 $y_{1,2} | \mathbf{y}_{-1,2}$ observations into eight quantile bins because $\mathbb{E}[Y_{1,2} | \mathbf{Y}_{-1,2}]$ can be exceptionally small for some $\mathbf{Y}_{-1,2}$. For each bin, the average count is less than and greater than $\mathbb{E}[Y_{1,2}]$ when the sum $\mathbf{Y}_{-1,2}$ is less than and greater than $\mathbb{E}[\mathbf{Y}_{-1,2}]$, re-

spectively. This trend is especially apparent for $\mathbf{Y}_{-1,2}$ far from the mean IBD count $\binom{n}{2} \times \mathbb{E}[Y_{1,2}]$. These findings provide some empirical evidence that the theorem assumption may be true for moderate sample sizes.

A.3. Efficient simulation of identity-by-descent segments around a locus

A.3.1. Approximating the Wright-Fisher process for large samples

The Kingman coalescent assumes that sample size is much smaller than population size such that there being more than one coalescent event in a generation is improbable. This assumption 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 [2, 24, 33].

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

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

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

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

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.

Simulating the approximate number of coalescent events in one generation from $\text{Binomial}(\binom{k}{2}, N'^{-1})$ can be more efficient than drawing a parent for each child and then scanning a vector of size k for siblings. The last term being subtracted in Equation A.2 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.

$$
\begin{aligned}
&P(A_1 = a_1, A_2 = a_2) - P(A_1 = a_1) \times P(A_2 = a_2) \\
&= \binom{k}{a_1} \frac{1}{N'} \cdots \frac{1}{N' - a_1} \times \binom{k - a_1}{a_2} \frac{1}{N' - a_1 - 1} \cdots \frac{1}{N' - a_1 - a_2} \\
&- \binom{k}{a_1} \frac{1}{N'} \cdots \frac{1}{N' - a_1} \times \binom{k}{a_2} \frac{1}{N'} \cdots \frac{1}{N' - a_2} \\
&\leq \binom{k}{a_1} \frac{1}{N'} \cdots \frac{1}{N' - a_1} \times \binom{k}{a_2} \frac{1}{N' - a_1 - 1} \cdots \frac{1}{N' - a_1 - a_2}.
\end{aligned}
\tag{A.3}
$$

Equation A.3 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 A.3 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. [2] have shown that the majority of simultaneous coalescent events in a generation are of this type.

A.3.2. The probability of detectable haplotype segment lengths

Within tens of generations most haplotype segment lengths are shrunk by crossovers to less than detection thresholds that are used in IBD-based analyses[13]. 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

$$1 - F_{R_a|t}(w) = \exp(-Ntw), \tag{A.4}$$

$$1 - F_{W_a|t}(w) = \exp(-Ntw) + Ntw \cdot \exp(-Ntw). \tag{A.5}$$

Supplementary Figure S16 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 descendant from an ancestor more than three hundred generations ago.

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

[13] A Morgan length threshold at least greater than 0.01 is typical in applied research [4, 5, 8, 29].

event can be derived with a telescoping sum argument:

$$\begin{aligned}\mathbb{E}[T^{+}_{n:k}] &= 2 \times \sum_{l=k}^{n} \frac{1}{l(l-1)} \\ &= 2 \times \sum_{l=k}^{n} \left(\frac{1}{l-1} - \frac{1}{l}\right) \\ &= 2 \times ((k-1)^{-1} - n^{-1}).\end{aligned} \tag{A.6}$$

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 time $\mathbb{E}[T_{n:400}]$ is 501.25.

A.3.3. The probability that recombination endpoints are shared between haplotypes

At some point in the past two sample haplotypes may share the same recombination endpoints. 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 the right $R_{a,d}, R_{b,d} \sim \text{Exponential}(u)$ and $R_{d,e} \sim \text{Exponential}(v)$. Supplementary Figure S17 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

$$\begin{aligned} P(\min(R_{a,d}, R_{b,d}, R_{d,e}) = R_{d,e}) &= \int_{R_{d,e}=r} P(R_{d,e} = r) \times P(\min(R_{a,d}, R_{b,d}, r) = r) \\ &= \int_{R_{d,e}=r} P(R_{d,e} = r) \times P(R_{a,d} \geq r)^2 \\ &= \int_r v \exp(-(2u+v)r) dr \\ &= \frac{v}{2u+v}. \end{aligned} \tag{A.7}$$

For fixed u, the limit of Equation A.7 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 $2 \times (1 - n^{-1})$ [17]. 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. Supplementary Figure S17 illustrates the difference between the limit of Equation A.7 and the following proposition in the context of branch lengths.

Proposition A.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 A.6). For* $j = o(k)$,

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

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

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

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

$\square$

The implication of Proposition A.1 is that haplotypes that share a recent common ancestor should have redundant endpoints at the most distant common ancestors. Since recombinations to the right and left of a focal location are independent, the result of Proposition A.1 extends to simulating IBD segments overlapping a focal location.