Highlights

Identity-by-descent segments in large samples

Seth D. Temple, Elizabeth A. Thompson

- We show the asymptotic normality of the detectable identity-by-descent rate, a mean of correlated binary random variables that arises in population genetics studies.
- We generalize our main central limit theorem to cover scenarios of nonconstant population sizes, multi-way identity-by-descent segments, and identity-by-descent rates of multiple samples from the same population.
- In enormous simulation studies, we use an efficient algorithm to characterize distributional properties of the detectable identity-by-descent rate.

Identity-by-descent segments in large samples

Seth D. Temple[a,b,c,*], Elizabeth A. Thompson[a]

[a]*Department of Statistics, University of Washington, Seattle, Washington, USA*
[b]*Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA*
[c]*Michigan Institute for Data Science, University of Michigan, Ann Arbor, Michigan, USA*

Abstract

If two haplotypes share the same alleles for an extended gene tract, these haplotypes are likely to be derived identical-by-descent from a recent common ancestor. Identity-by-descent segment lengths are correlated via unobserved ancestral tree and recombination processes, which commonly presents challenges to the derivation of theoretical results in population genetics. We show that the proportion of detectable identity-by-descent segments around a locus is normally distributed when the sample size and the scaled population size are large. We generalize this central limit theorem to cover flexible demographic scenarios, multi-way identity-by-descent segments, and multivariate identity-by-descent rates. We use efficient simulations to study the distributional behavior of the detectable identity-by-descent rate. One consequence of non-normality in finite samples is that a genome-wide scan looking for excess identity-by-descent rates may be subject to anti-conservative control of family-wise error rates.

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

[*]Corresponding author
Email address: `sethtem@umich.edu` (Seth D. Temple)

January 3, 2025

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 that overlap 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.

Using modern methods, long IBD segments can be detected with high accuracy from genetic data [21, 39, 44, 57]. 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, 23, 35]. They have also been used to estimate relatedness [21, 39, 57], haplotype phase [2, 33], mutation rates [37, 51, 52], recombination rates [55], gene conversion rates [6, 37], demographic changes [4, 8, 36], and positive selection [50]. We will study a sample mean of indicators if an IBD segment is long enough to be reliably detected. The binary random variables are correlated via unobserved recombinations and a random ancestral tree.

For independent, identically distributed data, maximum likelihood estimators are asymptotically consistent, efficient, and normally distributed under regularity conditions [10]. Composite likelihood approaches are commonly used in genetics when it is analytically intractable or computationally expensive to address dependencies in the data [30]. To what extent consistency, efficiency, and asymptotic normality extend to maximum composite likelihood estimators is generally unknown [30]. Studying maximum composite likelihood estimators can be especially challenging if their maxima do not have a closed form [36, 51]. In our work, the composite likelihood will be the binomial likelihood, which is maximized by the sample mean of binary random variables. The statistical property we care the most about is asymptotic normality, which is that the estimator's distribution converges to a Gaussian distribution as the sample size tends to infinity [10].

Without theoretical results, some authors assume that their estimators are distributed within some parametric family. In one example, Palamara et al. [38] assume without proof that their estimator of coalescent rates within the past tens of generations is Gamma distributed. In another example, Carmi et al. [9] observe that the Gaussian distribution is a good fit for the average fraction of the genome shared IBD by an individual with any other individual. Still, this observation is not the same as a theoretical result. When the sampling distribution is not sub-normal [53], statistical inference assuming normality may understate the probability of extreme values.

Creating valid confidence intervals can be more straightforward when an estimator is asymptotically normally distributed. The parametric bootstrap approach proposed in Temple et al. [50] gives adequate coverage in selection coefficient estimation for numerous simulation studies. Their technique implicitly assumes that the rate of detectable IBD segments around a locus, and certain functions[1] thereof, are normally distributed in large samples. In contrast, bootstrap resampling [16] has been employed in IBD-based estimation procedures [4, 6, 8, 36, 51]. For significance level α, these existing works do not demonstrate that their $(1-\alpha)\%$

[1] Functions that satisfy the Delta method conditions [10]: the first derivative at the expected number of detectable IBD segment around a locus exists and is nonzero.

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 [34].

Here, we derive sufficient conditions under which the proportion of detectable IBD segments around a locus is asymptotically normally distributed. 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 times the population size from which a large sample is drawn. The large population size requirement, in particular, indicates that most of the branch lengths in the ancestral tree must be long for the result to hold. The overall contribution of this work is to support IBD-based statistical inference with rigorous theory and extensive simulation studies.

The outline of the paper is as follows. In Section 2, we formally define our probability model for IBD segments that overlap a fixed location. In Section 3, we present and prove our main result for the asymptotic normality of the detectable IBD rate around a fixed location. In Section 4, we generalize our central limit theorem to cover nonconstant population sizes, multi-way IBD segments, and IBD rates between samples from the same population. In Section 5, we use simulation to investigate the statistical properties of IBD-based estimators and IBD graphs around a locus. Many calculations of covariance terms are left to the Appendix.

2. Preliminary material

First, we define our mathematical notation. The notation in Sections 2.1 and 2.2 follow the notation used in Temple et al. [49]. We use the Kingman coalescent [26, 27] as a model for the times until recent common ancestors, and we use the

Poisson process to model recombination without interference. The probability that an IBD segment is longer than a detection threshold is derived by integrating over these two waiting time distributions.

2.1. The time until a common ancestor

Let n be the haploid sample size and $k \leq n$ be the size of a subsample. Define N to be the constant population size and $N(t)$ the population size t generations ago. Let the random variable T_k denote the time until a common ancestor is reached for any two of k haploids, which we measure in units of N generations. In the discrete-time Wright-Fisher (WF) process, each haploid has a haploid ancestor in the previous generation, and if haploids have the same haploid ancestor, their lineages join.

The Kingman coalescent comes from the continuous-time limit of the WF process when subsample sizes are much smaller than constant population sizes. Specifically, T_k converges weakly to Exponential($\binom{k}{2}$) for $k \ll N$ and $N \to \infty$ [26, 27], where $\binom{k}{2}$ is the rate parameter. We focus on the times $T_4 \sim$ Exponential$(6), T_3 \sim$ Exponential(3), and $T_2 \sim$ Exponential(1) until any two of four, three, and two haploids reach a common ancestor, respectively.

2.2. The distance until crossover recombination

The genetic distance (in Morgans) between two loci is the number of crossovers expected to occur in an offspring gamete. Assuming no interference in double-stranded breaks and independent crossovers, Haldane [24] derives that the genetic distance until a crossover recombination is exponentially distributed. This result leads to modeling crossover points along the genome as a Poisson process. Browning [3] considers crossover models without and with interference [28] when studying transitions between IBD states, whereas we exclusively use the model without interference.

From a fixed point, the Morgans distance in one direction until a gamete offspring crossover is exponentially distributed with rate parameter 1. After t independent meioses, the surviving haplotype segment length to the right of the focal location is distributed as Exponential(t), where t is the rate parameter. Let a and b be sample haplotypes in the current generation, and define $L_a, R_a \,|\, t \sim$ Exponential(t) to be sample haplotype a's recombination endpoints to the left and right of a focal location after t generations. Because crossovers to the left and right of the focal location are independent, the extant width from the ancestor at time t is $W_a := L_a + R_a \,|\, t \sim \text{Gamma}(2, t)$. Since the t meioses descend independently to a and b from their most recent common ancestor, the IBD segments that are shared by a and b are $L_{a,b}, R_{a,b} | t \sim$ Exponential$(2t)$ and $W_{a,b} \,|\, t \sim \text{Gamma}(2, 2t)$.

2.3. The presence of detectable IBD segments

Relative to a focal point, we consider 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 detection threshold w Morgans. The binary random variables $\{X_{a,b}\}$ are identically distributed with the same mean $\mathbb{E}_2[X_{a,b}]$ and 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.

Our central limit theorem concerns a mean of the IBD segment indicator random variables. Namely, the detectable IBD rate to the right of a fixed location is

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

Let $Z_{a,b} := X_{a,b} - \mathbb{E}_2[X_{a,b}]$ be the mean-centered binary random variable, and let the sum of all except one of these mean-centered random variables be $\mathbf{Z}_{-a,b} := \sum_{(c,d)} Z_{c,d} - Z_{a,b}$. The sum of variances of all IBD segment indicators is

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

Finally, the mean-centered and suitably scaled detectable IBD rate to the right of a locus is

$$\bar{\mathbf{Z}}_{\binom{n}{2}} := \Omega_{\binom{n}{2}}^{-1/2} \times (\bar{\mathbf{X}}_{\binom{n}{2}} - \mathbb{E}_2[X_{a,b}]). \tag{3}$$

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. We drop the subscript $\binom{n}{2}$ when it is clear that the aggregation is over $\binom{n}{2}$ pairs of haplotypes. Figure 1 provides a conceptual example calculating $\bar{\mathbf{Y}}$ for four sample haplotypes.

We use additional subscript indices when segments are IBD among multiple haplotypes, which we refer to as multi-way IBD segments. For instance, $Y_{a,b,c}$ indicates if the IBD segment around a locus that is shared between haplotypes a, b, and c is longer than w Morgans. The corresponding sample mean over $\binom{n}{3}$ haplotype triplets is denoted $\bar{\mathbf{Y}}_{\binom{n}{3}}$, and the related sums, means, and variances are defined similarly. This notation is important to extend our main central limit theorem to multi-way IBD segment indicators.

We use the superscript l to denote the sample label when different population samples are considered. For example, $X^0_{a,b}$ and $X^1_{c,d}$ indicate if the IBD segments around a locus that are shared between haplotypes a and b in population sample 0 and c and d in population sample 1 are longer than w Morgans, respectively. Mean-centered and bold-faced terms are defined analogously for these extensions. For example, the mean in population sample 0 of 2-way IBD segment indicators overlapping a focal location is denoted $\bar{\mathbf{Y}}^0$. This notation is important to extend our main univariate central limit theorem to a multivariate Gaussian version.

3. Main central limit theorem

If $U_1, \ldots, U_n \sim^{iid} G$ for some model G, the Lindeberg-Lévy central limit theorem says that the standardized sample mean weakly converges to the standard normal distribution (under some regularity conditions) [31]. The special case of this result for binary random variables [15] is more closely related to our work. The result does not apply in our case because the IBD segment indicators $\{X_{a,b}\}$ to the right of a focal point are not independent. We start by focusing on the mean-centered and suitably scaled detectable IBD rate $\bar{\mathbf{Z}}_{\binom{n}{2},N}$ to the right of a focal location, where the subscript N clarifies that the haplotypes are sampled from a population of constant size N.

Our central limit theorems concern large sample size n and large population size N scaled by the Morgans detection threshold w. The intuition for our weak law is that the covariance between IBD segment indicators $\sum_{(a,b)\neq(c,d)} \text{Cov}(X_{a,b}, X_{c,d})$ is small relative to the sum of the variances of the individual IBD segment indicators $\Omega_{\binom{n}{2}}$. The sum of covariances between random variables being negligible compared to the sum of variances of the random variables themselves is the basis of the general central limit theorem for dependent data that is given in Chandrasekhar and Jackson [11] and Chandrasekhar et al. [12].

Theorem 3.1. *For n and Nw tending to infinity, the mean-centered and suitably scaled detectable IBD rate $\bar{\mathbf{Z}}_{\binom{n}{2},N}$ to the right of a focal location converges in distribution to the standard normal distribution 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. [12]. 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.

$$\mathbb{E}_2[X_{a,b}] = \int \exp(-2Nt_2w)\exp(-t_2)\,dt_2 = (2Nw+1)^{-1} = O((Nw)^{-1}). \tag{4}$$

It is easy to show that $\mathbb{E}_2[X_{a,b}] \to 0$ uniformly for large scaled population size (Lemma A.1). The second condition implies that $\Omega_{\binom{n}{2}} \to \infty$. The assumption in Chandrasekhar et al. [12] that $\mathbb{E}[|Z_{a,b}|^3]/\mathbb{E}[|Z_{a,b}|^2]^{3/2}$ is bounded above is true for nondegenerate Bernoulli random variables [11] (Lemma A.2). Lastly, given $n = o(Nw)$, we show that

$$\sum_{(a,b)\neq(c,d)} \mathrm{Cov}(\, X_{a,b}, X_{c,d}\,) = o(\Omega_{\,\binom{n}{2}}\,). \tag{5}$$

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.

$$\Omega_{\binom{n}{2}} \sim n^2 \cdot O((Nw)^{-1}) = o((Nw)^2) \cdot O((Nw)^{-1}) = o(Nw); \tag{6}$$

$$\sum_{a,b,c} \text{Cov}_3(X_{a,b}, X_{a,c}) \sim n^3 \cdot O((Nw)^{-2}) = o((Nw)^3) \cdot O((Nw)^{-2}) = o(Nw); \tag{7}$$

$$\sum_{a,b,c,d} \text{Cov}_4(X_{a,b}, X_{c,d}) \sim n^4 \cdot O((Nw)^{-3}) = o((Nw)^4) \cdot O((Nw)^{-3}) = o(Nw). \tag{8}$$

The covariance between IBD segment indicators (Equations 7 and 8) is controlled by the covariance within IBD segment indicators (Equation 6). □

The first two conditions have appealing interpretations. First, $Nw = o(n^2)$ says that the sample size is large enough relative to the scaled population size such that we observe many IBD segments to the right of a focal location that are longer than the Morgans 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 Morgans 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 Morgans 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 nonnegative covariance, which we show in Appendix A.1. Moreover, one intuits that the posterior distribution of $X_{a,b}|\mathbf{X}_{-a,b}$ is more likely to come from a tree with long branches than the unconditional distribution of $X_{a,b}$ is when $\mathbf{X}_{-a,b} < \mathbb{E}_2[X_{a,b}] \times (\binom{n}{2} - 1)$, and vice versa when $\mathbf{X}_{-a,b} > \mathbb{E}_2[X_{a,b}] \times (\binom{n}{2} - 1)$.

One can show that the small sample size $n = 3$ is a pathological example where the third condition breaks down (Lemma A.6). 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.2), concluding that this condition likely holds in large samples.

The asymptotic normality of $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ follows from the same arguments as those of the proof in Theorem 3.1. We show in Appendix A.1 that $\text{Cov}_2(Y_{a,b}, Y_{a,b})$, $\text{Cov}_3(Y_{a,b}, Y_{a,c})$, and $\text{Cov}_4(Y_{a,b}, Y_{c,d})$ are $O((Nw)^{-1})$, $O((Nw)^{-2})$, and $O((Nw)^{-3})$, respectively.

Theorem 3.2. *For n and Nw tending to infinity, the mean-centered and suitably scaled detectable IBD rate $\bar{\tilde{\mathbf{Z}}}_{\binom{n}{2},N}$ around a locus converges in distribution to the standard normal distribution 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 $\tilde{\mathbf{Z}}_{-a,b}$.*

4. Extensions

4.1. Flexible demographic scenarios

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

Theorem 4.1. *For n, $N_1 w$, and $N_2 w$ tending to infinity, rhe mean-centered and suitably scaled detectable IBD rate $\bar{\mathbf{Z}}_{\binom{n}{2},N(t)}$ to the right of a focal location converges in distribution to the standard normal distribution 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 $\bar{\bar{\mathbf{Z}}}_{\binom{n}{2},N(t)}$.

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.

$$\Omega_{\binom{n}{2}} \sim n^2 \cdot O((N_2 w)^{-1}) = o(N_2 w); \tag{9}$$

$$\sum_{a,b,c} \mathrm{Cov}_3(X_{a,b}, X_{a,c}) \sim n^3 \cdot O((N_2 w)^{-2}) = o(N_2 w); \tag{10}$$

$$\sum_{a,b,c,d} \mathrm{Cov}_4(X_{a,b}, X_{c,d}) \sim n^4 \cdot O((N_2 w)^{-3}) = o(N_2 w). \tag{11}$$

□

Theorem 3.1 is a special case of Theorem 4.1 when $N_1 = N_2$. The conditions in Theorem 4.1 are unlikely to hold in real data examples and are more difficult to interpret. Note that the proof of Theorem 4.1 does not make use of the entire curve $N(t)$. The population sizes at the most recent coalescent times impact the covariance of and between IBD segments around a focal location the most. As in Theorem 3.2, we can extend Theorem 4.1 to address IBD segments overlapping a focal location.

4.2. Multi-way IBD segments

To calculate the probability that an m-way IBD segment indicator is 1, we integrate over $m - 1$ coalescent times and the recombination processes at these common ancestors. Here, we consider $m > 2$ but m much smaller than the sample size n. For example, we compute the expected value of the 3-way IBD segment indicator to the right of a focal location

$$\begin{aligned}\mathbb{E}_3[X_{a,b,c}] = \int & \exp(-2Nt_2w)\exp(-3Nt_3w) \\ & \exp(-t_2)\exp(-3t_3)\,dt_2 \\ & = 3(2Nw+1)^{-1}(Nw+1)^{-1} = O((Nw)^{-2}).\end{aligned} \tag{12}$$

Note in this derivation and that of Equation 4 fall under the general result that $\mathbb{E}_m[X_{\ldots m}] = O((Nw)^{-(m-1)})$, where $\ldots m$ denotes m labeled haplotypes. To observe many m-way IBD segment indicators, we require $(Nw)^{m-1} = o(n^m)$ because the sums are over $\binom{n}{m} \sim n^m$ identically distributed random variables.

Theorem 4.2. *For n and Nw tending to infinity and bounded $m = O(1)$, the mean-centered and suitably scaled detectable IBD rate $\bar{\mathbf{Z}}_{\binom{n}{m},N}$ to the right of a focal location converges in distribution to the standard normal distribution when the following are true:*

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

The weak convergence result holds for $\bar{\bar{\mathbf{Z}}}_{\binom{n}{m},N}$ *under the same conditions.*

Proof. The proof is again to show that the three conditions are sufficient to apply Corollary 1 in Chandrasekhar et al. [12]. The strategy is to calculate the relevant integrals $\mathbb{E}_m[\cdot], \dots, \mathbb{E}_{2m}[\cdot]$, count the number of occurrences of each covariance type, and then observe that the condition $n = o(Nw)$ is sufficient to control the total covariance. In Appendix A.1.2, we give a full proof for the 3-way IBD rate, from whose covariances and combinatorics it is straightforward to see a pattern as m increases.

□

Theorems 3.1 and 3.2 are special cases of Theorem 4.2 when $m = 2$. We remark that $n = o(Nw)$, which does not involve m, is a condition shared between Theorems 3.1 and 4.2. Recall that this condition maintains that covariances between IBD segment indicators are small, which is governed by large scaled population size Nw.

4.3. Multivariate IBD rates

We now show that the conditions $n = o(Nw)$ and $Nw = o(n^2)$ are also sufficient to apply the multivariate version of the Chandrasekhar et al. [12] central limit theorem. From the multivariate result, we can derive the asymptotic distribution of the difference in IBD rates between case and control sample sets. To extend our main result to multivariate random vectors, we consider the example of two disjoint sample sets labeled 0 and 1. Each sample consists of n samples from the same population of size N.

Let $(\bar{\mathbf{X}}^0, \bar{\mathbf{X}}^1)' \in \mathbb{R}^{2\times n}$ be the column vector of two sample means, where $'$ is transpose. The detectable identity-by-descent segment rates around a locus are denoted $(\bar{\mathbf{Y}}^0, \bar{\mathbf{Y}}^1)'$, and the standardized sample means are denoted $(\bar{\mathbf{Z}}^0, \bar{\mathbf{Z}}^1)'$ and $(\bar{\bar{\mathbf{Z}}}^0, \bar{\bar{\mathbf{Z}}}^1)'$. In general, we denote $\bar{\mathbf{X}}^{1:\ell}$ and $\bar{\mathbf{Y}}^{1:\ell}$ and $\bar{\mathbf{Z}}^{1:\ell}$ and $\bar{\bar{\mathbf{Z}}}^{1:\ell}$ as the unstandardized and standardized IBD rates to the right of and overlapping a focal location for ℓ distinct samples of n haplotypes from N. The mean-centered sums of IBD segment indicators excluding $Z^l_{a,b}$ and $\tilde{Z}^l_{a,b}$ are denoted $\mathbf{Z}^l_{-a,b}$ $\tilde{\mathbf{Z}}^l_{-a,b}$, respectively.

Theorem 4.3. *For bounded $\binom{\ell}{2}$ and n and Nw tending to infinity, the mean-centered and suitably scaled IBD rates $\bar{\mathbf{Z}}^{1:\ell}$ converge in distribution to the standard normal distribution $N_\ell(\mathbf{0}, \mathbf{I}_{\ell\times\ell})$ when the following are true:*

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

The weak convergence result holds for $\bar{\bar{\mathbf{Z}}}^{1:\ell}$ under the same conditions.

(The proof is in Appendix A.1.3 using the result from Chandrasekhar et al. [12].)

One important consequence of Theorem 4.3 is that affine transformations of the sample means column vector are asymptotically normally distributed. In particular, for the example of two samples and the row vector $(1, -1)$, the difference in standardized IBD rates around a locus $\bar{\bar{\mathbf{Z}}}^0 - \bar{\bar{\mathbf{Z}}}^1$ is asymptotically normally distributed. When there are ℓ sample sets, for each pair of the ℓ sample means, a row vector exists such that the dot product gives the difference in their IBD rates.

To apply Corollary 1 of Chandrasekhar et al. [12], we restrict our result to equally sized samples of n haplotypes. In case-control studies, there may be samples of unequal sizes n_1 and n_0. We conjecture that the difference in IBD rates will still asymptotically normally distributed, so long as $Nw = o(n_1^2)$ and $Nw = o(n_0^2)$ and $\max(n_0, n_1) = o(Nw)$. The conditions $Nw = o(n_1^2)$ and $Nw = o(n_0^2)$ maintain that we detect many IBD segments in both samples. The condition $\max(n_0, n_1) = o(Nw)$ maintains that covariances are vanishing both in the diagonal terms $\text{Cov}(\bar{\bar{\tilde{\mathbf{Z}}}}^1, \bar{\bar{\tilde{\mathbf{Z}}}}^1)$ and $\text{Cov}(\bar{\bar{\tilde{\mathbf{Z}}}}^0, \bar{\bar{\tilde{\mathbf{Z}}}}^0)$ and the off-diagonal term $\text{Cov}(\bar{\bar{\tilde{\mathbf{Z}}}}^0, \bar{\bar{\tilde{\mathbf{Z}}}}^1)$.

Another limitation is our restriction to distinct sample sets, which is necessary to make the covariance calculations analytically tractable. Browning and Thompson [7] study the IBD rates between case-case, case-control, and control-control haplotype pairs, resulting in a sample means vector that does not fall under our mathematical framework. We conjecture that the empirical distributions of such vectors may be similar to those of vectors of nonoverlapping sample sets when the samples come from a large population. The reason for our conjecture is the same as before: the large scaled population size leads to vanishing covariances in the diagonal and off-diagonal terms.

5. Simulation studies

The theoretical results in Sections 3 and 4 rely on asymptotic conditions, not finite sample conditions. Using simulation, we explore the finite sample empirical distributions and percentiles of detectable IBD rate-based statistics around a fixed location. To investigate normality, we require massive simulations to form tens of thousands of empirical distributions.

We use the algorithm in Temple et al. [49] to simulate detectable IBD segments overlapping a fixed location. Despite the speed of the algorithm, the enormous scope of our simulations takes hundreds of days of computing time, which we spread across core processing units. If not for the algorithm's efficiency, we would be limited in our ability to study the distributional behavior of the standardized detectable IBD rate $\bar{\tilde{\mathbf{Z}}}$ and the difference in IBD rates $\bar{\tilde{\mathbf{Z}}}^0 - \bar{\tilde{\mathbf{Z}}}^0$.

We consider sample sizes of five and ten thousand "diploid" individuals. To implement "diploids", we use a haploid model with two times the sample size of diploids (and likewise for demographic models). We consider the same demographic scenarios described in Temple et al. [50] and Temple et al. [49]: constant population sizes ranging from ten thousand to ten million diploid individuals and examples of exponential growth phases and a population bottleneck. Both complex demographic scenarios amount to population sizes $\geq 10^6$ in the most recent tens of generations and population sizes $\leq 10^4$ more than a few hundred generations ago. Figure S1 from Temple et al. [49] illustrates some of these demographic scenarios.

5.1. Identity-by-descent rates in finite samples

5.1.1. Constant population sizes

Using the Shapiro-Wilk test [41, 42, 43], we investigate if empirical distributions of $\sum_{a,b} Y_{a,b}$ resemble normal distributions as sample size n, population size N, and the Morgans length threshold w increase. We partition simulated IBD rates into five hundred empirical distributions based on one thousand observations. The null hypothesis is that the empirical distribution of detectable IBD rates 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 significance level 0.05.

Figure 2 shows the proportion of rejected tests for increasing population size and Morgans length threshold with sample size fixed at five and ten thousand diploid individuals. The trend is that the proportion of rejected tests decreases with the increasing population size and Morgans length threshold. Figure S2 shows that this trend does not depend on the significance 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 Morgans 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).

Figure S3 shows the proportion of rejected tests for increasing sample size and Morgans length threshold with population size fixed at fifty and one hundred thousand diploid 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).

5.1.2. Flexible demographic scenarios

Figure S4 shows the proportion of rejected tests for increasing sample size and Morgans length threshold in the three phases of exponential growth and population bottleneck demographic scenarios. For Morgans 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 Morgans 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 diploid individuals (Figure 2).

The conditions on the global extrema of population sizes in Theorem 4.1 are very stringent. The most recent population sizes have the strongest impact on the covariances of IBD segment indicators. One interpretation of the results in Figure S4 is that the detectable IBD rate around a locus may behave like a normal distribution in demographic scenarios with large recent population sizes, regardless of the not-so-recent population sizes.

5.1.3. Difference of identity-by-descent rates in two samples

We compute the difference of detectable IBD rates around a locus by splitting five thousand diploid individuals into two equally sized subsets. Then, under different experimental conditions, we perform two hundred and fifty Shapiro-Wilk tests based on five hundred simulations of the test statistic.

Figure S4 shows the proportion of rejected tests for increasing population size and Morgans length threshold. At the significance level 0.05, and for all scaled population sizes, between 0.05 and 0.15 percent of tests are rejected. At the significance level 0.10, and for all scaled population sizes, between 0.10 and 0.30 percent of tests are rejected. There is no apparent trend as either population size or Morgans length threshold increases. One explanation is that any potential overdispersion of $\bar{\bar{\mathbf{Z}}}^0$ and $\bar{\bar{\mathbf{Z}}}^1$, relative to the standard normal distribution, may be partially balanced out by considering the difference of the rates. Another explanation is the limited power to reject the Shapiro-Wilk null hypothesis in the scope of our computationally feasible experiments.

Across all simulation experiments in Sections 5.1.1, 5.1.2, and 5.1.3, we reject normality at rates greater than the Type 1 error rate 0.05 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 validity of our central limit theorems.

5.2. Percentiles of the finite sample distributions

Next, we investigate possible explanations for rejecting the nominal significance levels at elevated rates. We focus on the upper percentiles of the empirical distribution of our test statistics $\bar{\tilde{\mathbf{Z}}}$ and $\bar{\tilde{\mathbf{Z}}}^0 - \bar{\tilde{\mathbf{Z}}}^1$. For each batch of simulations, we compute a mean, a standard deviation, and the mean plus three or four standard deviations. Then, we calculate the 99.86501^{th} and 99.99683^{th} percentiles of the test statistic over all batches. (These percentiles correspond to the standard normal quantiles of three and four.) We multiply the reciprocal of these quantiles by their corresponding estimated upper bounds, which we refer to as the relative upper bounds.

5.2.1. The identity-by-descent rate in one sample

Browning and Browning [5], Temple et al. [50], and Temple [48] conduct hypothesis tests to evaluate if the detectable IBD rate $\bar{\tilde{\mathbf{Z}}}$ around any specific locus exceeds a genome-wide mean IBD rate. When our central limit theorems hold, we can interpret their hypothesis test as a one-sample one-sided z tests. Our estimated upper bounds, the mean plus some standard deviations, are meant to mimic their hypothesis tests [5, 50].

Figures 3 and S6 show the average relative upper bounds by increasing population size and Morgans length threshold. The average estimated upper bounds are less than the simulated percentile threshold for all sample sizes, population sizes, Morgans length thresholds, and quantiles considered. The average estimated upper bound is proportionally closer to the percentile threshold as population size and Morgans length threshold increase, which is a result consistent with Section 5.1.1 and our central limit theorems.

Figure S4 shows that the average estimated upper bound is also less than the simulated percentile threshold for all sample sizes and Morgans length thresholds in the complex demographic scenarios. The average estimated upper bound is proportionally closer to the percentile threshold for the population bottleneck scenario compared to the three phases of exponential growth scenario, which is the complex demographic scenario with larger recent population sizes (Figure S1).

These experiments suggest that one reason why we reject the Shapiro-Wilk null hypothesis at elevated rates is because the test statistic's upper tail probability is heavier than that of the standard normal distribution.

5.2.2. Difference of identity-by-descent rates in two samples

Analogous to the excess IBD rate test, the difference in IBD rates $\bar{\tilde{\mathbf{Z}}}^0 - \bar{\tilde{\mathbf{Z}}}^0$ may be used as a hypothesis test for equality of means between two labeled subgroups. We perform the same experiment, except for the difference in IBD rates as our test statistic.

Figure S7 shows the average relative upper bounds by increasing population size and Morgans length threshold. We see no trend between the average relative upper bounds and sample size, population size, and Morgans length threshold, respectively. Compared to our observation in the one-sample experiment, the test statistic's upper tail probability is not noticeably different from that of the standard normal distribution. These empirical observations are consistent with our Type 1 error experiment in Section 5.1.3.

5.3. Identity-by-descent graphs around a locus

Clusters of detectable IBD haplotypes overlapping a focal point indicate non-negligible covariance between segments. These cluster covariances 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 as 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. We calculate the average, variance, minimum, and maximum for each feature over replicate simulations. We also conduct Shapiro-Wilk tests by splitting the simulated data as described in Section 5.1.1.

Note that the number of trees of order m is not the same as the m-way IBD rate around a locus. For example, in a complete connected component of four nodes, there are $\binom{4}{3}$ counts of 3-way detectable IBD. As a result, Theorem 4.2

does not apply to the following experiments on tree orders. However, we might expect to see some approximately normally distributed data if most components of degree m are trees.

5.3.1. 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 [17]. 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 Morgans length threshold [36]. 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 [18], and 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 is that almost all nodes are in trees of small order or a single "giant" component [18].

Figure 4 shows that some empirical distributions of graph features resemble normal distributions in a sample size of five thousand diploid individuals from a population of one hundred thousand diploid individuals. 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 nonzero covariance of IBD edges[2]. 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 we have presented.

5.3.2. Flexible demographic scenarios

Figure S8 shows that the apparent normality of some graph features extends to the three phases of exponential growth and population bottleneck demographic scenarios. Table S1 reports that the proportions of rejected hypothesis tests for numbers of trees of order 2 are close to 0.05 for both demographic scenarios. We also 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 that the limiting distributional behaviors of graph features in detectable IBD graphs around a locus can be similar for large constant populations and demographic scenarios with large recent population sizes.

5.3.3. The impact of strong positive selection

Strong directional selection increases the detectable IBD rate around a locus [50] and the probability of IBD alleles [1], but less is known about how this phenomenon alters the feature distributions of detectable IBD graphs. In a hard

[2] The expected number of edges should be the same, if not for some approximations [36, 49].

selective sweep, a single allele increases in frequency at a rate of change that depends on a selection coefficient [14, 20, 25, 54]. The selection coefficient parameterizes the advantage that the sweeping allele has relative to alternative alleles, in so much as the gradient of the allele frequency trajectory is larger when the selection coefficient is larger.

We conduct more simulations of detectable IBD graphs for selection coefficients between 0.01 and 0.04 and the three phases of exponential growth and population bottleneck scenarios. Tables S2 and S3 demonstrate multiple trends as the selection coefficient increases. The apparent normality of the number of trees of order 2 does not noticeably change as we change the selection coefficient. Compared to our simulations with no selection, 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 more apparent under the selection models with more detectable IBD segments. The main effect of strong positive selection appears to be the growth of the 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. [50].

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 around a locus whose regularity conditions have intuitive interpretations in population genetics. The sample size squared must be large enough such that there are many

IBD segments long enough to be accurately detected by existing methods [5, 21, 44, 56]. The population size must be large enough that there are few to no large IBD clusters about a locus.

The conceptual framework for these conditions involves envisioning a coalescent tree with long internal branches, but numerous coalescent events occur near the leaves. The internal branches are long because of the large population size, and there are numerous coalescent events near the leaves because of the large sample size. The large Morgans threshold further decreases the probability of a detectable IBD segment and the correlations between IBD segment indicators.

The techniques we use might be useful to other studies involving coalescent *and* recombination processes. For instance, to generalize our main central limit theorem, we take a formulaic approach. First, we derive covariances for a finite set of classes. Second, we count the number of covariance terms of each class that occur in the total covariance of the sample mean. Third, we determine a “little-o” condition such that the sum of covariances of one specific class is asymptotically equivalent to the sum of covariances of all the other classes. We use a particular central limit theorem for dependent data [11, 12], which is derived using Stein’s moment-based method—a more general technique to demonstrate weak convergences to Gaussian or non-Gaussian random variables [29, 40, 46, 47].

One counterexample in which our particular proof strategy does not work concerns the density of recent coalescent events (DRC) [38]. This test statistic is the sample mean of indicators if a haplotype pair has a common ancestor within a given time threshold [3]. For constant population size N, each covariance type is

[3] The DRC and the detectable IBD rate have both been used in selection scans [5, 50]. Intuitively, they capture a similar population genetics signal.

the survival function of a hypoexponential random variable. The main terms in all these survival functions take the form $\exp(-C \cdot N)$, where C is a constant. However, the combinatorics are the same as those in our IBD-based statistics. As a result, the sum of covariances of one specific class is *not* asymptotically equivalent to the sum of covariances of all the other classes. This observation points out the role that integrating over *shared recombinations* plays in reducing the covariance.

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 nonnegligible covariance may come from the accumulation of IBD clusters. Based on the tail behavior of simulated distributions, we expect that a one-sample z test for excess IBD rates may inflate the number of false positives.

Our regularity conditions concern a balance between sample size and scaled population size that is unlikely to hold in practical settings. In our experiments, we observe neither a trend between sample size and the proportion of rejected tests nor between sample size and the relative upper tail probability. We advocate that the collected sample size should always be as large as is feasible and that the smallest Morgans length threshold for which IBD segment detection is accurate should be chosen.

Our theoretical results and simulation studies support ongoing methodological developments based on IBD segments. Existing genome-wide scans for excess IBD rates [5, 50] or differences in IBD rates between groups [7] lack formal or exact hypothesis testing frameworks. Motivated in part by this work, Temple [48] controls the family-wise error rate (FWER) in their selection scan by modeling the IBD rate process as an Ornstein-Uhlenbeck process, thereby assuming that the IBD rate is normally distributed at any given spatial position. Consistent with this work, they show anti-conservative control of FWER. Combining an FWER control technique [19, 45] with our multivariate central limit theorem, we indicate that a modification of the Temple [48] method may apply to a test for equality of detectable IBD rates in case-control studies. In these examples and others [13, 22, 32] from statistical and population genetics, assuming reasonable asymptotic models is often vital when adjusting for many correlated tests.

Data and code availability

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.

Acknowledgements

S.D.T. acknowledges funding support from the US National Defense Science and Engineering Graduate Fellowship, the US National Institute of Health T32 GM081062 Pre-doctoral Training Grant in Statistical Genetics, and Schmidt Sciences LLC. Computational resources were supported in part 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. We thank Sharon R. Browning for helpful discussions.

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] A. Albrechtsen, I. Moltke, and R. Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186:295–308, 2010.

[2] 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, 2021.

[3] S. Browning. A Monte Carlo approach to calculating probabilities for continuous identity by descent data. *J. Appl. Prob.*, 37(3):850–864, 2000.

[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, 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, 2020.

[6] S. R. Browning and B. L. Browning. Biobank-scale inference of multi-individual identity by descent and gene conversion. *Am. J. Hum. Genet.*, 111(4):691–700, 2024.

[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, 2012.

[8] R. Cai, B. L. Browning, and S. R. Browning. Identity-by-descent-based es-

timation of the X chromosome effective population size with application to sex-specific demographic history. *G3*, 13(10), 2023.

[9] S. Carmi, P. R. Wilton, J. Wakeley, and I. Pe'er. A renewal theory approach to IBD sharing. *Theor. Popul. Biol.*, 97:35–48, 2014.

[10] G. Casella and R. L. Berger. *Statistical Inference*. Thomson Learning, 2002.

[11] A. G. Chandrasekhar and M. O. Jackson. A network formation model based on subgraphs. *arXiv*, 2016. doi: 10.48550/arXiv.1611.07658.

[12] A. G. Chandrasekhar, M. O. Jackson, T. H. McCormick, and V. Thiyageswaran. General covariance-based conditions for central limit theorems with dependent triangular arrays. *arXiv*, 2023. doi: 10.48550/arXiv.2308.12506.

[13] K. N. Conneely and M. Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *Am. J. Hum. Genet.*, 81(6):1158–1168, 2007.

[14] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, New York, NY, 1970.

[15] A. De Moivre. *The Doctrine of Chances, Or, A Method of Calculating the Probabilites of Events in Play...* Pearson, 1718.

[16] B. Efron. Better bootstrap confidence intervals. *J. Am. Stat. Assoc.*, 82(397): 171–185, 1987.

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

[18] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

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

[20] R. A. Fisher. XXI.—on the dominance ratio. *Proc. R. Soc. Edinb.*, 42:321–341, 1923.

[21] 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, 2021.

[22] K. E. Grinde, L. A. Brown, A. P. Reiner, T. A. Thornton, and S. R. Browning. Genome-wide significance thresholds for admixture mapping studies. *Am. J. Hum. Genet.*, 104(3):454–465, 2019.

[23] 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, 2011.

[24] 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.

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

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

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

[28] D. D. Kosambi. The estimation of map distances from recombination values. *Ann. Eugen.*, 12(1):172–175, 1943.

[29] K. Lange. *Applied Probability*. Springer, New York, NY, 2nd edition, 2010.

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

[31] J. W. Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Math. Z.*, 15(1):211–225, 1922.

[32] Y. Liu, S. Chen, Z. Li, A. C. Morrison, E. Boerwinkle, and X. Lin. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.*, 104(3):410–421, 2019.

[33] 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, 2016.

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

[35] 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, 2020.

[36] 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, 2012.

[37] 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, 2015.

[38] P. F. Palamara, J. Terhorst, Y. S. Song, and A. L. Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.*, 50(9):1311–1317, 2018.

[39] 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.

[40] N. Ross. Fundamentals of Stein's method. *Probab. Surv.*, 8:210–293, 2011.

[41] S. S. Shapiro and R. S. Francia. An approximate analysis of variance test for normality. *J. Am. Stat. Assoc.*, 67(337):215–216, 1972.

[42] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[43] S. S. Shapiro, M. B. Wilk, and H. J. Chen. A comparative study of various tests for normality. *J. Am. Stat. Assoc.*, 63(324):1343–1372, 1968.

[44] 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, 2021.

[45] D. O. Siegmund and B. Yakir. *The Statistics of Gene Mapping*. Statistics for Biology and Health. Springer, New York, NY, 2007.

[46] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Berkeley Symp. Math. Statist. Probab.*, pages 583–602, 1972.

[47] C. Stein. *Approximate Computation of Expectations*, volume 7. Institute of Mathematical Statistics, 1986.

[48] S. D. Temple. *Statistical Inference Using Identity-by-Descent Segments: Perspectives on Recent Positive Selection*. PhD thesis, University of Washington, 2024.

[49] S. D. Temple, S. R. Browning, and E. A. Thompson. Fast simulation of identity-by-descent segments. *bioRxiv*, 2024. doi: 10.1101/2024.12.13.628449.

[50] S. D. Temple, R. K. Waples, and S. R. Browning. Modeling recent positive selection using identity-by-descent segments. *Am. J. Hum. Genet.*, 111(11): 2510–2529, 2024.

[51] 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, 2019.

[52] 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, 2022.

[53] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[54] S. Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931.

[55] 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, 2020.

[56] 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, 2020.

[57] 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, 2020.

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 as* $Nw \to \infty$.

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

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

Proof.

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

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

□

Lemma A.3. *Cov*$_3(Z_{a,b}, Z_{a,c}) \equiv$ *Cov*$_3(X_{a,b}, X_{a,c}) = O((Nw)^{-2})$.

Proof. Up to reordering three sample haplotypes, there is one possible bifurcating tree (Figure S9). 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 four sample haplotypes, there are two possible bifurcating trees (Figure S10). 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.

$$\begin{aligned}\mathbb{E}_4[X_{a,c}] &= \mathbb{E}_4[X_{b,d}] \\ &= 6 \cdot 3 \int\int \exp(-2N(t_4 + t_3 + t_2)w) \exp(-(6t_4 + 3t_3 + t_2))\, dt_4 dt_3 dt_2 \\ &= 18(2Nw + 6)^{-1}(2Nw + 3)^{-1}(2Nw + 1)^{-1}.\end{aligned} \tag{A.7}$$

$$\begin{aligned}\mathbb{E}_4[X_{a,c}X_{b,d}] = 6 \cdot 3 \int\int\int & \exp(-(4Nt_4 + 3Nt_3 + 2Nt_2)w) \\ & \exp(-(6t_4 + 3t_3 + t_2))dt_4 dt_3 dt_2 \\ = 18(4Nw + 6)^{-1}&(3Nw + 3)^{-1}(2Nw + 1)^{-1}.\end{aligned} \tag{A.8}$$

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

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,

$$\begin{aligned}\mathbb{E}_4[X_{a,c}] &= 6 \cdot 3 \int\int \exp(-2N(t_4 + t_3)w) \exp(-(6t_4 + 3t_3 + t_2))\, dt_4 dt_3 dt_2 \\ &= 18(2Nw + 6)^{-1}(2Nw + 3)^{-1}.\end{aligned} \tag{A.10}$$

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

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*

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

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 4 and Lemmas A.3 and A.4.

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

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

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

□

Lemma A.6. *For a sample of three haplotypes a, b, and c, when $\mathbb{E}_2[X_{a,c}] < 1/2$, the conditional expectation $\mathbb{E}[Z_{a,c} \times \mathbf{Z}_{-a,c} | \mathbf{Z}_{-a,c}] \not\geq 0$ for all $\mathbf{Z}_{-a,c}$.*

Proof. Define $q =: \mathbb{E}_2[X_{a,c}]$, and fix $\mathbf{X}_{-a,c} = 1$.

$$\begin{aligned}\mathbb{E}[Z_{a,c} \times \mathbf{Z}_{-a,c} | \mathbf{Z}_{-a,c}] &= \mathbb{E}[(X_{a,c} - q) \times (X_{a,b} + X_{b,c} - 2q) | X_{a,b} + X_{b,c} = 1] \\ &= \mathbb{E}[X_{a,c} \times (1 - 2q) | X_{a,b} + X_{a,c} = 1] - q + 2q^2\end{aligned}$$

Because of IBD transitivity, $X_{a,c} = 0$ with probability 1. Then, the equation simplifies to $-q(1 - 2q) < 0$. □

A.1.2. Multi-way IBD segments

Proof of Theorem 4.2. We give the general argument for 3-way IBD segment indicators. To begin, we calculate bounds on the relevant integrals $\mathbb{E}_k[\cdot], \ldots, \mathbb{E}_{2k}[\cdot]$. Recall that $\mathbb{E}_k$ is the expected value with respect to a coalescent tree of k haplotypes.

$$\begin{aligned}\mathbb{E}_3[X_{a,b,c}X_{a,b,c}] &= O((Nw)^{-2}) \\ \mathbb{E}_4[X_{a,b,c}X_{a,b,d}] &= O((Nw)^{-3}) \\ \mathbb{E}_5[X_{a,b,c}X_{a,d,e}] &= O((Nw)^{-4}) \\ \mathbb{E}_6[X_{a,b,c}X_{d,e,f}] &= O((Nw)^{-5}).\end{aligned} \tag{A.16}$$

These are also the covariance bounds because $\mathbb{E}_3[X_{a,b,c}X_{a,b,c}] \geq 0$ and $\mathbb{E}_3[X_{a,b,c}X_{a,b,c}] \geq \mathbb{E}_3[X_{a,b,c}]^2$ and so on for the other $\mathbb{E}_k$ relations.

Next, we take sums over these covariance bounds and substitute in the $n = o(Nw)$ condition.

$$\begin{aligned}\Omega_{\binom{n}{3}} &\sim n^3 \cdot O((Nw)^{-2}) \\ &= o((Nw)^3) \cdot O((Nw)^{-1}) \\ &= o(Nw);\end{aligned} \tag{A.17}$$

$$\begin{aligned}\sum_{a,b,c,d} \mathrm{Cov}_4(X_{a,b,c}, X_{a,b,d}) &\sim n^4 \cdot O((Nw)^{-3}) \\ &= o((Nw)^4) \cdot O((Nw)^{-3}) \\ &= o(Nw);\end{aligned} \tag{A.18}$$

$$\begin{aligned}\sum_{a,b,c,d,e} \mathrm{Cov}_5(X_{a,b,c}, X_{a,d,e}) &\sim n^5 \cdot O((Nw)^{-4}) \\ &= o((Nw)^5) \cdot O((Nw)^{-4}) \\ &= o(Nw);\end{aligned} \tag{A.19}$$

$$\begin{aligned}\sum_{a,b,c,d,e,f} \mathrm{Cov}_6(X_{a,b,c}, X_{d,e,f}) &\sim n^6 \cdot O((Nw)^{-5}) \\ &= o((Nw)^6) \cdot O((Nw)^{-5}) \\ &= o(Nw).\end{aligned} \tag{A.20}$$

The covariance within IBD segment indicators $\Omega_{\binom{n}{3}}$ controls the sum of covariances $\sum_{(a,b,c)\neq(d,e,f)} \mathrm{Cov}(X_{a,b,c}, X_{d,e,f})$. Using the bounding argument in Equation A.15, the result extends to IBD segment indicators around a focal location.

□

A.1.3. Multivariate IBD rates

Proof of Theorem 4.3. To use Corollary 1 from Chandrasekhar et al. [12] in multiple dimensions, we now require defining their notion of an affinity set. These are subsets $\mathcal{A}^l_{a,b}$ containing the haplotype pair a and b from sample l such that $\text{Cov}(X^l_{a,b}, X^{l^\star}_{c,d})$ is high if the haplotype pair c and d from sample $l^\star$ are in the affinity set and low if they are not. We consider the singleton affinity sets $\{X^l_{a,b}\}$. (We remark that singletons are the affinity sets we use in all of our proofs for the one-dimensional results.)

We use the example of two sample means to concretely calculate covariances. Let $\Omega_{2\times 2}$ be the covariance matrix such that

$$\begin{aligned}\Omega_{0,0} = \Omega_{1,1} &= \sum \text{Cov}(X^0_{a,b}, X^0_{a,b}) \sim n^2 (Nw)^{-1};\\ \Omega_{0,1} = \Omega_{1,0} &= 0.\end{aligned}$$

$\Omega_{0,1}$ and $\Omega_{1,0}$ concern the sum of covariances of IBD segment indicators within affinity sets but in different samples, which is zero because the affinity set of a haplotype pair in one sample includes no haplotype pairs in a different sample.

The term that controls the sum of covariances across affinity sets is the Frobenius norm $||\Omega_{2\times 2}||_F$. We calculate this norm as

$$\begin{aligned}||\Omega_{2\times 2}||_F &= \sqrt{\Omega^2_{0,0} + 2\cdot\Omega^2_{0,1} + \Omega^2_{1,1}}\\ &\sim \sqrt{2n^4 (Nw)^{-2} + 0} \qquad \text{(A.21)}\\ &= \sqrt{2} n^2 (Nw)^{-1}.\end{aligned}$$

Under the condition $n = o(Nw)$, Equation A.21 is $o(Nw)$, and under the condition

$Nw = o(n^2)$, the variance term $||\Omega_{2\times 2}||_F$ tends to infinity.

The first condition from Corollary 1 in Chandrasekhar et al. [12] is

$$\sum_{(l^\star,a,b)\neq(l,c,d)} \mathrm{Cov}(X^{l^\star}_{a,b}, X^{l}_{c,d}) = o(||\Omega_{2\times 2}||_F) = o(Nw). \tag{A.22}$$

We compute the sums of covariances of IBD segment indicator types $\{(a,b),(a,e)\}$ and $\{(a,b),(c,d)\}$, where a, b, and e are haplotypes in one sample and c and d are haplotypes in the other sample. By the same calculations as in the previous proofs, these sums are asymptotically equivalent to $n^3(Nw)^{-2} = o(Nw)$ and $n^4(Nw)^{-3} = o(Nw)$.

Since the column vector is now multi-dimensional, we must also show that

$$\sum_{(l^\star,a,b),(l,c,d)} \mathrm{Cov}((X^{l^\star}_{a,b})^2, (X^{l}_{c,d})^2) = o(||\Omega_{2\times 2}||^2_F). \tag{A.23}$$

This calculation is simplified as

$$\begin{aligned}\sum_{(l^\star,a,b),(l,c,d)} \mathrm{Cov}((X^{l^\star}_{a,b})^2, (X^{l}_{c,d})^2) &= \sum_{(l^\star,a,b),(l,c,d)} \mathrm{Cov}(X^{l^\star}_{a,b}, X^{l}_{c,d}) \\ &= \Omega_{0,0} + \Omega_{1,1} + \sum_{(l^\star,a,b)\neq(l,c,d)} \mathrm{Cov}(X^{l^\star}_{a,b}, X^{l}_{c,d}) \\ &= o(Nw).\end{aligned} \tag{A.24}$$

We get the general result by extending these calculations for sums and norms over covariances of two samples to those of ℓ samples. The term in Equation A.22 involves sums of covariances of $\binom{\ell}{2}$ pairs of samples. This term is why we require the bound on $\binom{\ell}{2}$, because in Equation A.21 we have the multiplicative factor $\sqrt{\ell}$.

Using the bounding argument in Equation A.15, the result extends to IBD segment indicators around a focal location.

□

A.2. Verifying an assumption of the central limit theorem

We take a Monte Carlo approach to examine the conditional expectation assumption $\mathbb{E}[\tilde{Z}_{a,b} \times \tilde{\mathbf{Z}}_{-a,b}|\tilde{\mathbf{Z}}_{-a,b}] \geq 0$ for all $\tilde{\mathbf{Z}}_{-a,b}$ because $\mathbb{E}[\tilde{Z}_{a,b}|\tilde{\mathbf{Z}}_{-a,b}]$ is analytically intractable. Namely, by replacing the expected value $\mathbb{E}[Y_{a,b}|\mathbf{Y}_{-a,b}]$ with an average over a large number of simulations, we assess if $\mathbb{E}[Y_{a,b}|\mathbf{Y}_{-a,b}] \geq \mathbb{E}[Y_{a,b}]$ when $\mathbf{Y}_{-a,b} \geq (\binom{n}{2} - 1) \cdot \mathbb{E}[Y_{a,b}]$ and vice versa that $\mathbb{E}[Y_{a,b}|\mathbf{Y}_{-a,b}] \leq \mathbb{E}[Y_{a,b}]$ when $\mathbf{Y}_{-a,b} \leq (\binom{n}{2} - 1) \cdot \mathbb{E}[Y_{a,b}]$. (Recall that $Z_{a,b}$ is the binary random variable $Y_{a,b}$ after mean-centering.) The intuition is that if the observed sum $\mathbf{Y}_{-a,b}$ is larger than the expected sum $\mathbb{E}[\mathbf{Y}_{-a,b}]$ then the held out $Y_{a,b}$ is more likely to be 1 than it would be if the observed sum equaled the expected sum.

We run the Temple et al. [49] algorithm one hundred and twenty million times, recording the value of $Y_{a,b}$ and the sum $\mathbf{Y}_{-a,b}$ for some fixed haplotype pair a and b. Then, we calculate the difference between the empirical average $\bar{Y}_{a,b}$ and $\mathbb{E}[Y_{a,b}]$, stratified into eight quantile bins depending on the sum $\mathbf{Y}_{-a,b}$. The sample sizes are limited to two to four hundred diploid individuals to keep runtime modest.

Figure S11 shows the results of this simulation study. For each bin, the average count is less than and greater than $\mathbb{E}[Y_{a,b}]$ when the sum $\mathbf{Y}_{-a,b}$ is less than and greater than $\mathbb{E}[\mathbf{Y}_{-a,b}]$, respectively. This trend is especially apparent for $\mathbf{Y}_{-a,b}$ far from the mean IBD count $(\binom{n}{2} - 1) \times \mathbb{E}[Y_{a,b}]$. These findings provide empirical evidence that the theorem assumption may be true for moderate to large sample sizes.

924 ## Figures

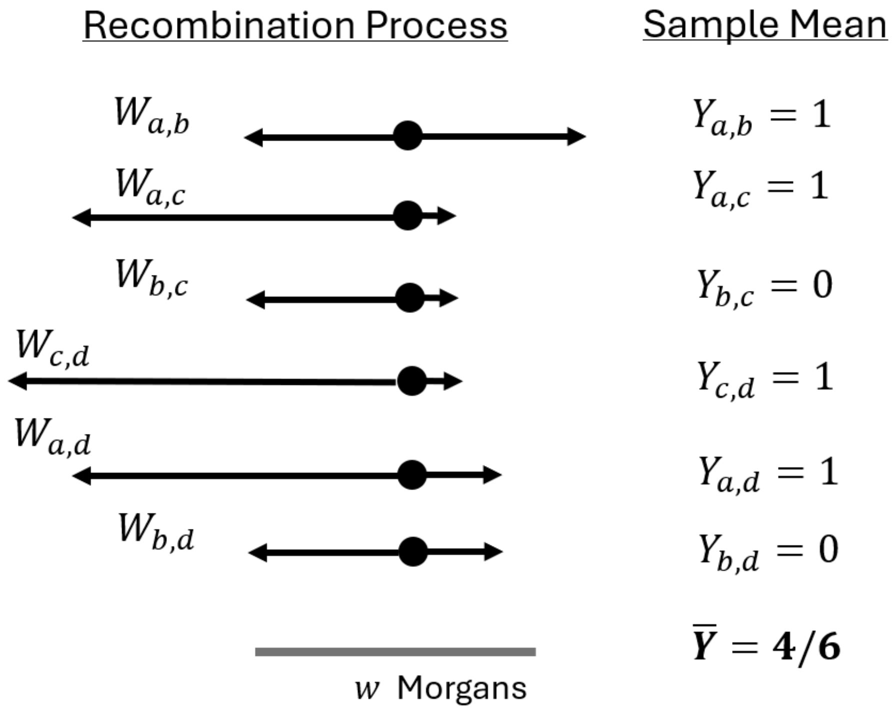

Figure 1: Example calculation of the detectable IBD rate. IBD segment lengths overlapping a focal point for sample haplotypes a, b, c, d are shown. The IBD segment indicators ($Y_{i,j}$'s) are 1 if their IBD segment lengths ($W_{i,j}$'s) exceed w Morgans and otherwise 0. The detectable IBD rate $\bar{\mathbf{Y}}$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

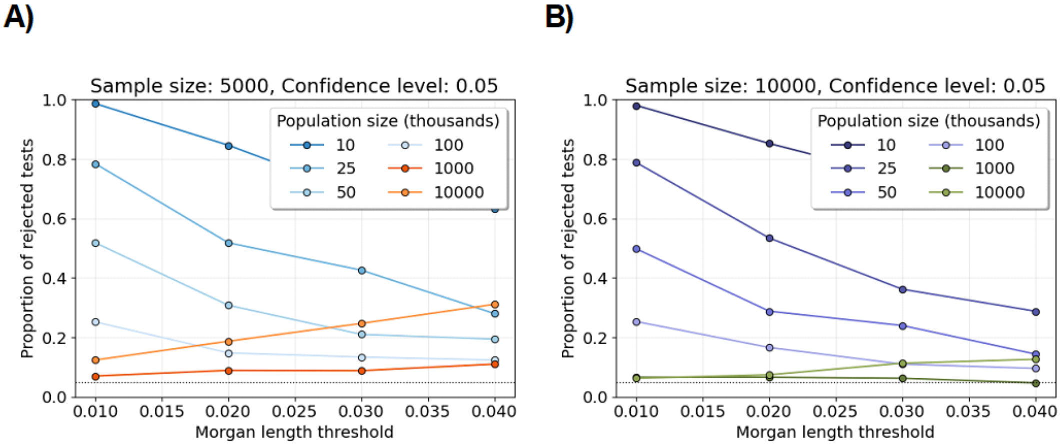

Figure 2: Shapiro-Wilk tests for varying population sizes. Line plots show the proportions of Shapiro-Wilk tests rejected at the significance level 0.05 (y-axis) for varying population size and fixed sample size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgans length threshold (x-axis). A) The sample size is five thousand individuals. B) The sample size is ten thousand. The legends assign colors to different population sizes. The horizontal dotted line is at 0.05.

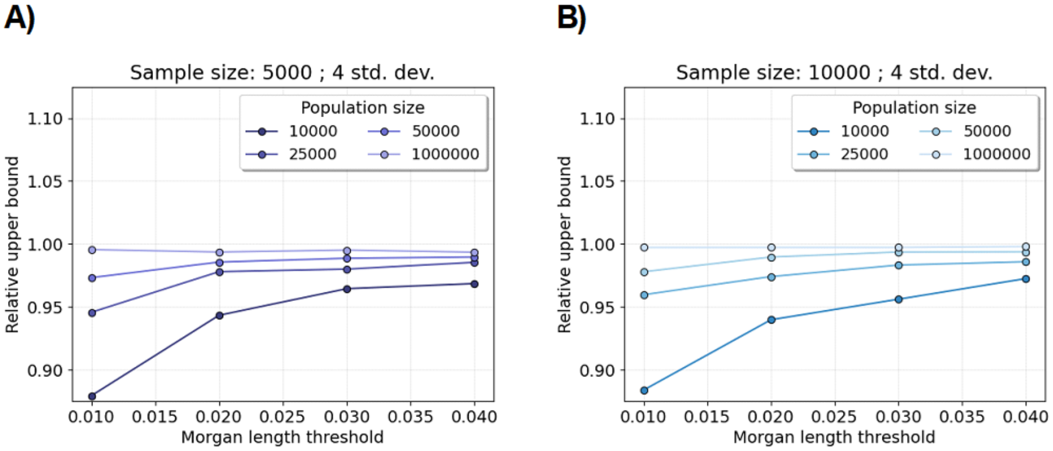

Figure 3: Relative upper bound for excess IBD scan. Line plots show the average mean plus four standard deviations divided by the 99.99683 percentile over two million simulations (y-axis). (The standard normal survival function of four is 0.9999683.) Each average relative upper bound is computed over one thousand tests. Each test is based on two thousand simulations of the number of identity-by-descent lengths longer than a specified Morgans length threshold (x-axis). A) The sample size is five thousand diploid individuals. B) The sample size is ten thousand diploid individuals. The legends assign colors to different constant population sizes.

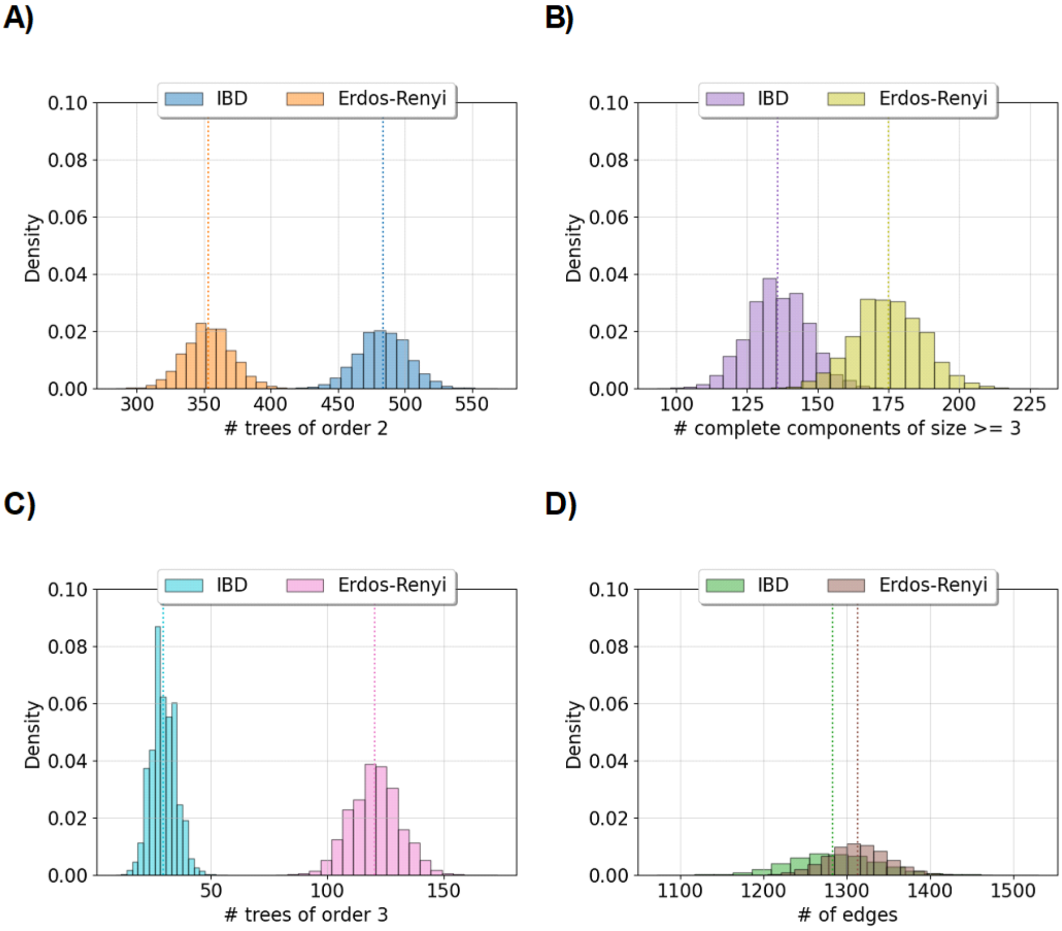

Figure 4: Comparing features between IBD and Erdős-Rényi graphs. Histograms compare the density of graph features between IBD and Erdős-Rényi graphs. Each histogram summarizes the results of one hundred and twenty-five thousand simulations. A) and C) show the number of trees of order 2 and 3, respectively. B) shows the number of complete components with more than three nodes. D) shows the total number of edges. The legends give color assigned to the IBD and Erdős-Rényi graphs. IBD graphs are simulated using the constant one hundred thousand diploid individuals' demography and the 0.03 Morgans length threshold. Erdős-Rényi graphs are simulated using the same success probability as in the IBD graph. The sample size is two thousand diploid individuals. Vertical lines show the means.

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

926 Supplementary figures

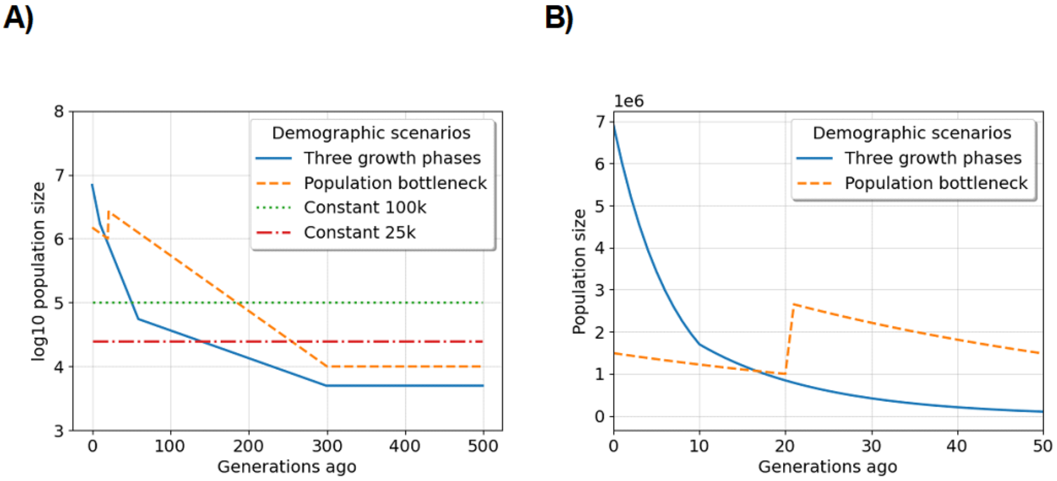

Figure S1: Demographic scenarios we consider in simulation studies: A) coalescent time in generations ago by the log 10 population size, and B) the most recent fifty generations by population size for examples of exponential growth. The legends specify the color and line style for each scenario. As opposed to coalescent time used in the main text, we describe the scenarios forward in time here. 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 a rate of two percent starting three hundred generations ago. Otherwise, the demographic scenarios we explore here are populations of constant size twenty-five and one hundred diploids.

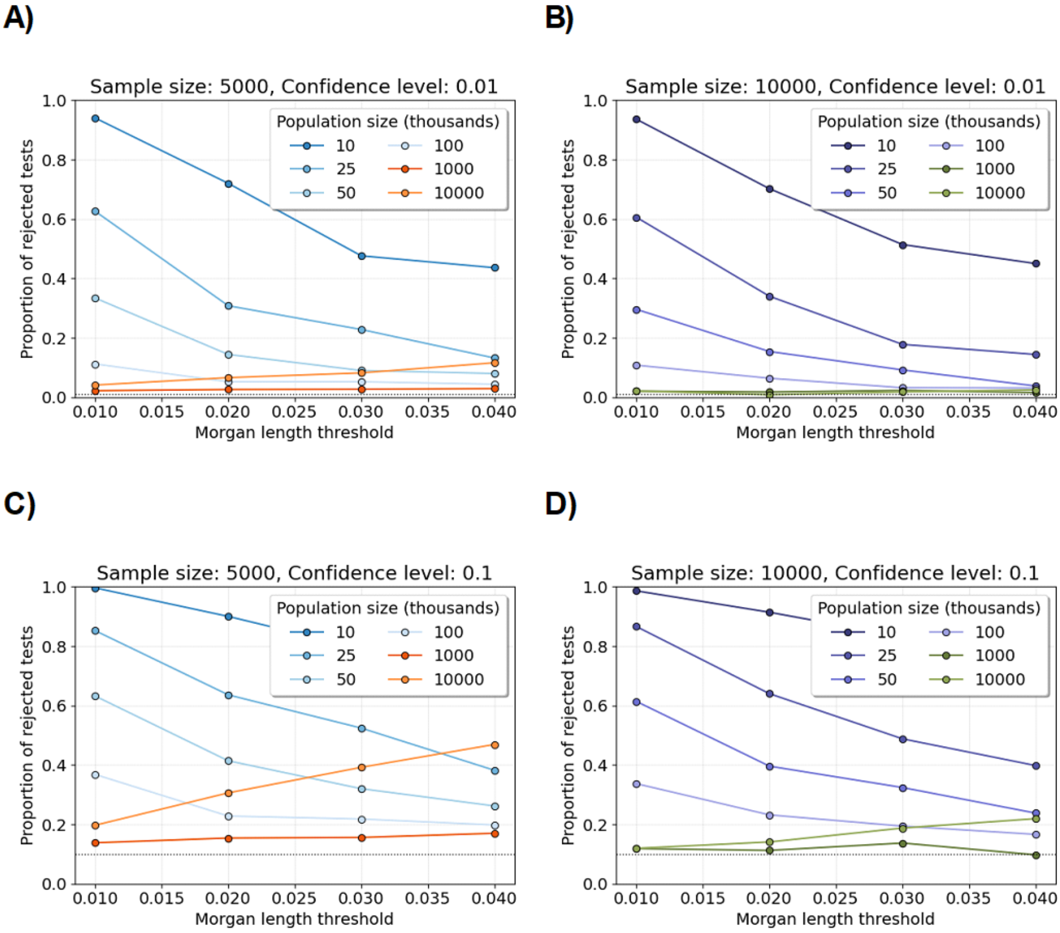

Figure S2: Shapiro-Wilk tests for varying population sizes and significance levels. Line plots show the proportions of Shapiro-Wilk tests rejected at significance levels A,B) 0.01 and C,D) 0.1 (y-axis) for varying population size and fixed sample size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgans length threshold (x-axis). A,C) The sample size is five thousand diploid individuals. B,D) The sample size is ten thousand diploid individuals. The legends assign colors to different population sizes. The horizontal dotted lines are significance levels.

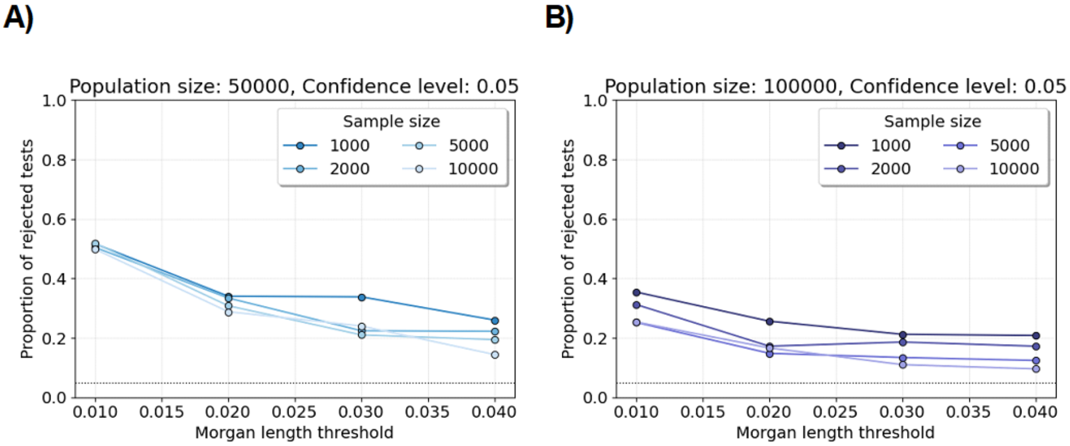

Figure S3: Shapiro-Wilk tests for varying sample sizes. Line plots show the proportions of Shapiro-Wilk tests rejected at the significance level 0.05 (y-axis) for varying sample size and fixed population size. Each proportion is computed over five hundred tests. Each test is based on one thousand simulations of the number of identity-by-descent lengths longer than a specified Morgans length threshold (x-axis). A) The population size is fifty thousand diploid individuals. B) The population size is one hundred thousand diploid individuals. The legends assign colors to different sample sizes. The horizontal dotted line is at 0.05.

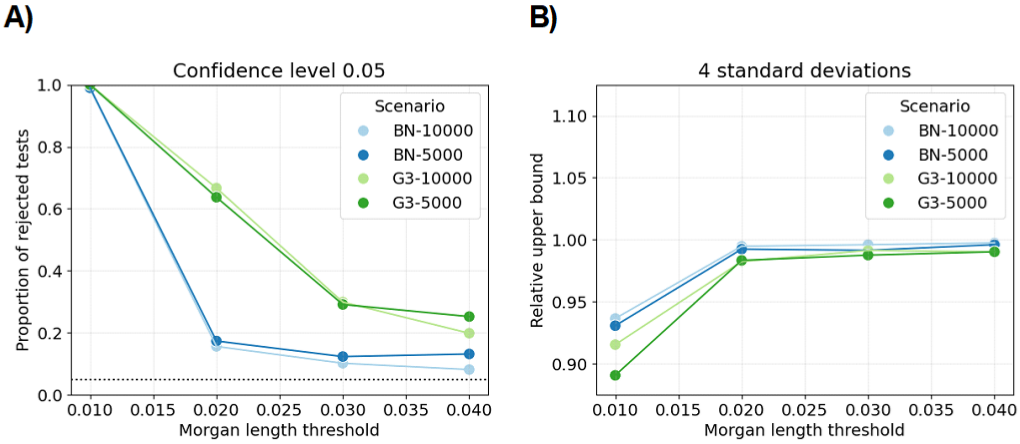

Figure S4: Shapiro-Wilk tests and relative upper tail bounds for complex demography scenarios. A) Line plots show the proportions of Shapiro-Wilk tests rejected at the sigificance level 0.05 (y-axis) for the population bottleneck (BN) or three phases of exponential growth (G3) demographic scenarios and sample sizes of five or ten thousand diploid 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 Morgans length threshold (x-axis). B) Line plots show the average mean plus four standard deviations divided by the 99.99683 percentile over two million simulations (y-axis). Plot designs are identical to Figures 2 and 3.

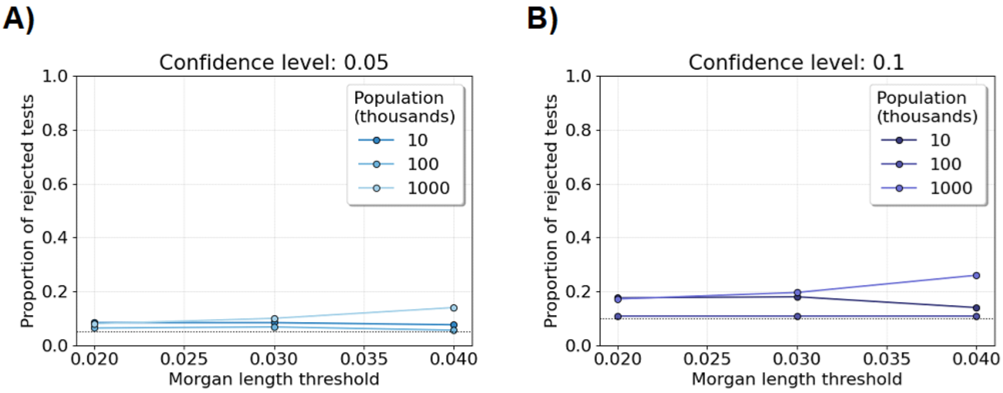

Figure S5: Shapiro-Wilk tests for difference in IBD rates between groups. Line plots show the proportions of Shapiro-Wilk tests rejected at the significance level 0.05 (y-axis) for increasing constant population sizes (in thousands). The sample size is five thousand diploid individuals. Each proportion is computed over two hundred and fifty tests. Each test is based on five hundred simulations of the difference between groups in IBD rates longer than a specified Morgans length threshold (x-axis). The significance threshold is either A) 0.05 or B) 0.10, shown as horizontal dotted black lines.

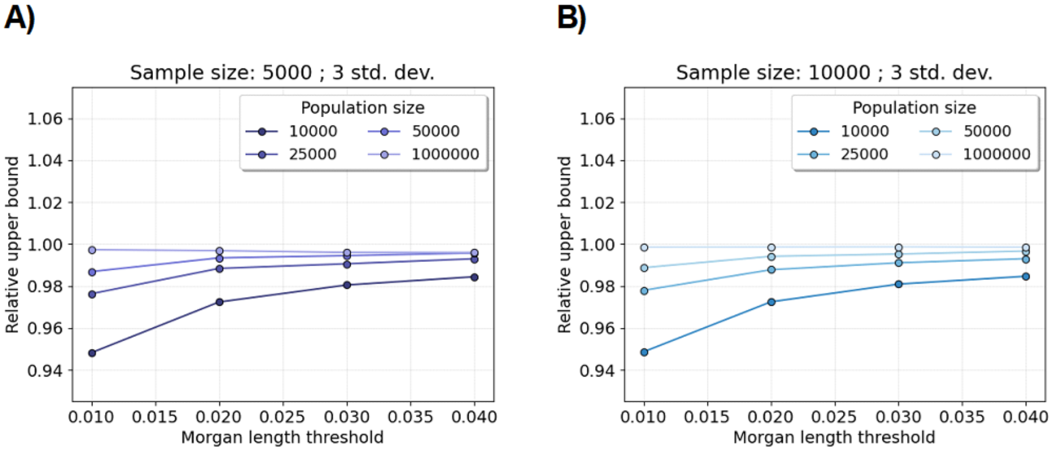

Figure S6: Relative upper bound for excess IBD scan. Line plots show the average mean plus three standard deviations divided by the 99.86501 percentile over two million simulations (y-axis). (The standard normal survival function of three is 0.9986501.) Each average relative upper bound is computed over one thousand tests. Each test is based on two thousand simulations of the number of identity-by-descent lengths longer than a specified Morgans length threshold (x-axis). A) The sample size is five thousand diploid individuals. B) The sample size is ten thousand diploid individuals. The legends assign colors to different constant population sizes.

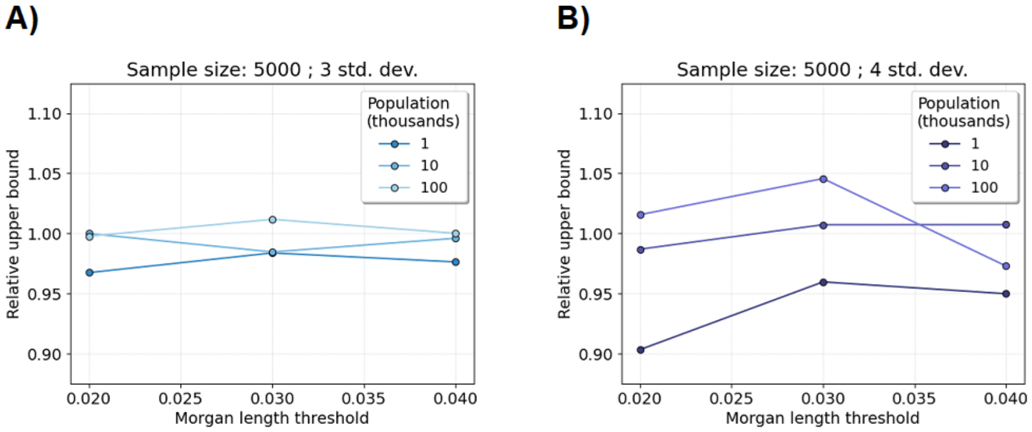

Figure S7: Relative upper bound for the difference in IBD rates test. Line plots show the average mean plus A) three or B) four standard deviations divided by the standard normal corresponding percentiles over one hundred and twenty-five thousand simulations (y-axis). Each average relative upper bound is computed over two hundred and fifty tests. Each test is based on five hundred simulations of the number of identity-by-descent lengths longer than a specified Morgans length threshold (x-axis). The sample size is five thousand diploid individuals. The legends assign colors to increasing constant population sizes (in thousands).

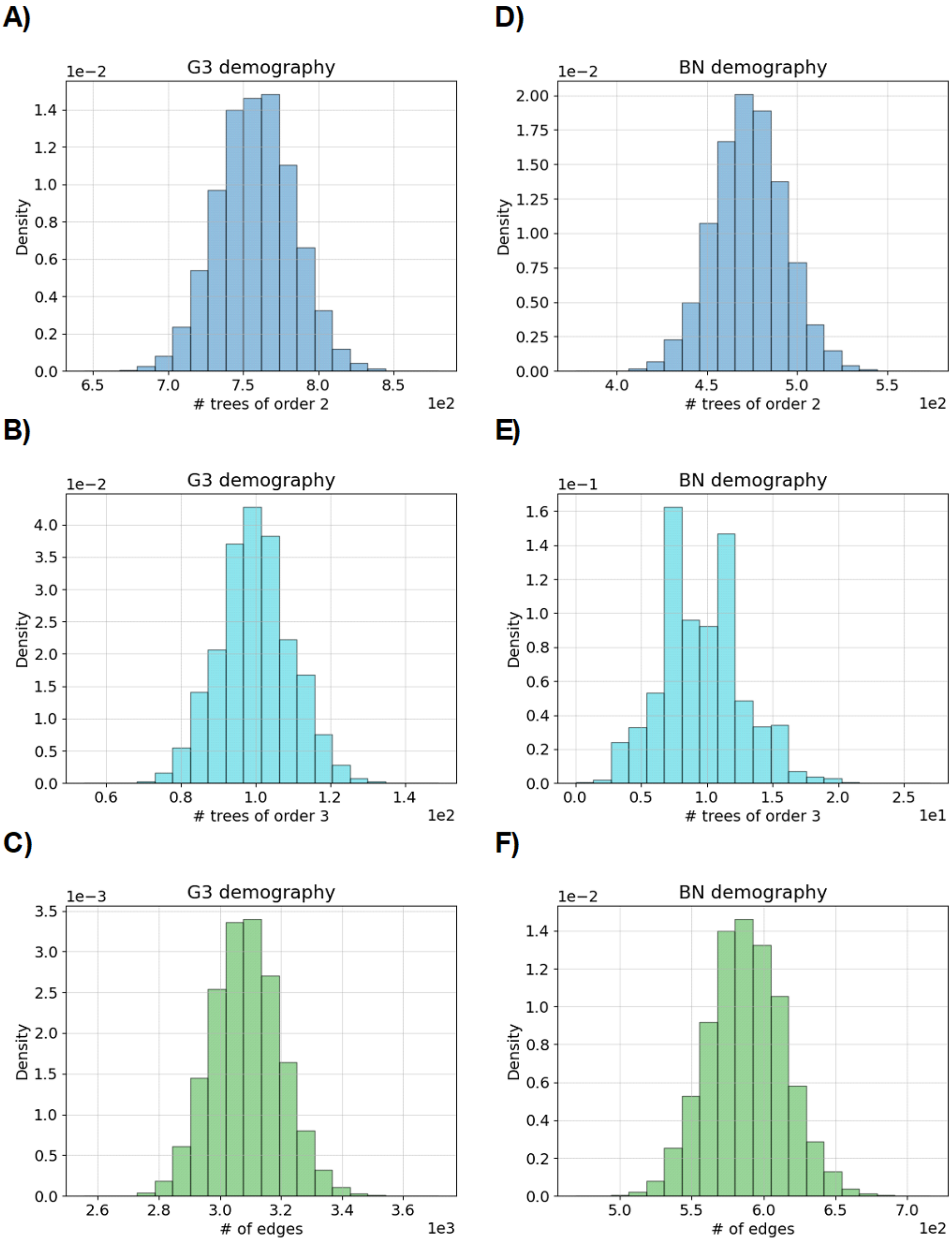

Figure S8: Comparing features between IBD graphs for complex demographic scenarios. Histograms show the density of IBD graph features between A-C) the three phases of exponential growth (G3) and D-F) the population bottleneck (BN) demographic scenarios. Each histogram is based on at least six hundred 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 Morgans length threshold is 0.03. The sample size is five thousand diploid individuals.

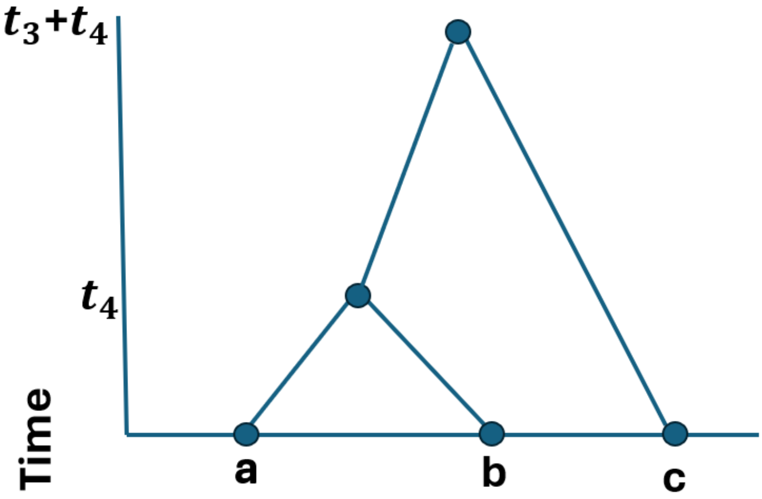

Figure S9: Illustration of the one possible coalescent tree used to calculate Cov_3 terms in Appendix A.1.

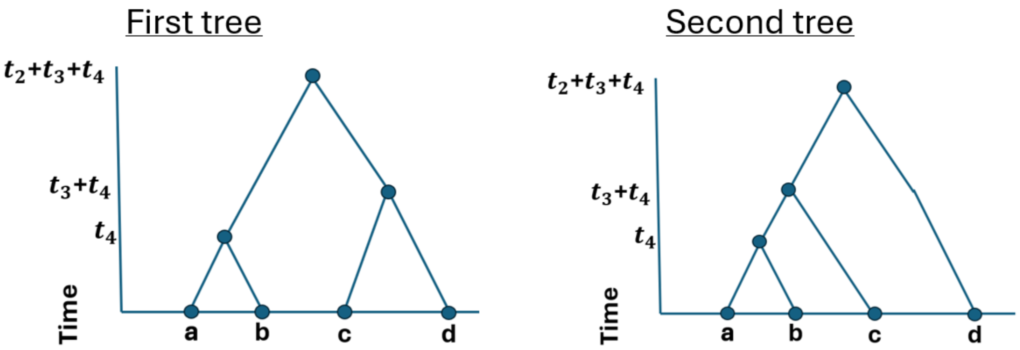

Figure S10: Illustration of the two possible coalescent trees used to calculate Cov_4 terms in Appendix A.1.

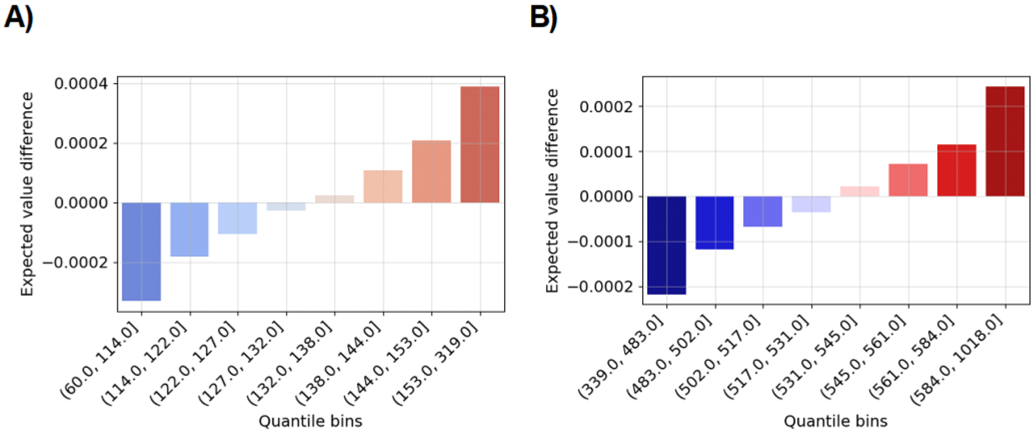

Figure S11: Monte Carlo verification of the conditional expectation condition in our central limit theorem. Bar charts show the difference between the proportion of simulations where two specific haplotypes share an IBD segment longer than 0.03 Morgans and the true success probability (y-axis). This 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 diploid individuals. The population size is ten thousand diploid individuals. The expectation is 132.78 in A) and 531.78 in B).

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 thousand simulations. Shapiro-Wilk tests at the significance level 0.05 are performed with 1000 replicates for at least 600 simulations, and the proportion of rejected null hypotheses is reported as S.W.t. The sample size is five thousand diploid individuals. The Morgans 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 $s \in [0.01, 0.02, 0.03, 0.4]$. The same description of IBD graph features as in Table 1. Shapiro-Wilk tests at the significance level 0.05 are performed with 250 replicates for 150 simulations, and the proportion of rejected null hypotheses reported as S.W.t. The sample size is five thousand diploid individuals. The Morgans length threshold is 0.03.

Type	**Structure**	**Avg**	**Var**	**Min**	**Max**	**S.W.t.**
$s = 0.01$	Edges	612.05	753.44	504.00	736.00	0.06
	Largest	**4.71**	0.75	3.00	14.00	0.97
	Tree2	481.32	400.48	397.00	566.00	0.06
	Tree3	11.33	11.24	1.00	25.00	0.90
	Complete	39.25	37.75	15.00	66.00	0.19
$s = 0.02$	Edges	722.33	1,349.58	582.00	967.00	0.38
	Largest	**9.79**	20.27	4.00	56.00	0.97
	Tree2	497.56	407.99	416.00	581.00	0.03
	Tree3	16.38	16.05	3.00	34.00	0.72
	Complete	50.79	48.02	24.00	81.00	0.15
$s = 0.03$	Edges	1,090.00	16,537.54	808.00	2,360.00	0.97
	Largest	**40.15**	456.43	8.00	172.00	0.97
	Tree2	501.78	424.81	409.00	592.00	0.06
	Tree3	20.80	20.37	4.00	43.00	0.47
	Complete	61.55	58.15	33.00	93.00	0.14
$s = 0.04$	Edges	2,177.58	284,697.22	1,219.00	7,591.00	0.97
	Largest	**122.45**	2,833.45	18.00	354.00	0.97
	Tree2	492.44	425.42	412.00	578.00	0.01
	Tree3	22.28	21.94	6.00	44.00	0.46
	Complete	66.05	63.26	36.00	99.00	0.19

Table S3: Summary statistics of IBD graphs for different selection coefficients and the population bottleneck demographic scenario. There is directional selection with different selection coefficients $s \in [0.01, 0.02, 0.03, 0.4]$. The same description of IBD graph features as in Table 1. Shapiro-Wilk tests at the significance level 0.05 are performed with 250 replicates for 150 simulations, and the proportion of rejected null hypotheses reported as S.W.t. The sample size is five thousand diploid individuals. The Morgans length threshold is 0.03.