Fast simulation of identity-by-descent segments

Seth D. Temple[1,2,3*], Sharon R. Browning[4] and Elizabeth A. Thompson[1]

[1*]Department of Statistics, University of Washington, Seattle, WA, USA.
[2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA.
[3]Michigan Institute of Data Science, University of Michigan, Ann Arbor, MI, USA.
[4]Department of Biostatistics, University of Washington, Seattle, WA, USA.

*Corresponding author(s). E-mail(s): sethtem@umich.edu;
Contributing authors: sguy@uw.edu; eathomp@uw.edu;

Abstract

The worst-case runtime complexity to simulate haplotype segments identical by descent (IBD) is quadratic in sample size. We propose two main techniques to reduce the compute time, both of which are motivated by coalescent and recombination processes. We provide mathematical results that explain why our algorithm should outperform a naive implementation with high probability. In our experiments, we observe average compute times to simulate detectable IBD segments around a locus that scale approximately linearly in sample size and take a couple of seconds for sample sizes that are less than ten thousand diploid individuals. In contrast, we find that existing methods to simulate IBD segments take minutes to hours for sample sizes exceeding a few thousand diploid individuals. When using IBD segments to study recent positive selection around a locus, our efficient simulation algorithm makes feasible statistical inferences, e.g., parametric bootstrapping in analyses of large biobanks, that would be otherwise intractable.

Keywords: identity-by-descent, coalescent, computational runtime

MSC Classification: 60-08 , 92-04 , 92-08 , 92-10 , 92D15

1 Introduction

Simulation is a powerful tool in population genetics to forecast the genetic impact of evolutionary scenarios, perform statistical inference on models and their parameters, and develop and evaluate new methods (Hoban et al., 2012; Yuan et al., 2012). There are two main frameworks for population genetics simulations, each having its own use cases, advantages, and disadvantages. Forward simulation models the dynamics of entire populations over time regarding individuals and their interactions (Haller and Messer, 2019). This flexible approach can incorporate complex dynamics of selection, migration, and spatial context, among other features, at the cost of additional computation. Backward simulation models the genealogy of present-day samples strictly through their common ancestors and is less computationally intensive (Hoban et al., 2012).

The speed of backward simulation is in large part due to coalescent theory (Kingman, 1982a,b), which approximates the Wright-Fisher (WF) process (Wright, 1931) when the sample size is much smaller than the population size. The Kingman coalescent has been extended to address examples of migration (Nath and Griffiths, 1993), recombination (Hudson, 1983; Hudson and Kaplan, 1988), selection (Hudson and Kaplan, 1988; Kaplan et al., 1988), and demography (Hein et al., 2005). With recombination, the model becomes a sequence of correlated coalescent trees called the ancestral recombination graph (ARG). In recent years, numerous coalescent methods have been developed to simulate polymorphism data over large genomic regions efficiently (Ewing and Hermisson, 2010; Hudson, 2002; Kern and Schrider, 2016), having randomly placed mutations on tree branches at a fixed genome-wide rate. The `msprime` software is a popular and robust option for backward simulation that scales to entire chromosomes and thousands of individuals (Baumdicker et al., 2021). Hybrid frameworks with forward simulations (Haller et al., 2019) and standards set for species-specific simulations (Adrion et al., 2020; Lauterbur et al., 2023) have contributed to its widespread adoption.

Placing mutations on tree branches has linear complexity in sample size, which means analyses focusing on summary statistics of polymorphism data can be runtime inexpensive even in large samples. On the other hand, deriving the pairwise relationships between haplotypes is difficult for large sample sizes because the total number of computations scales quadratically in sample size. To be precise, two individuals share a haplotype segment identical-by-descent (IBD) if they inherit it from the same common ancestor. `msprime` has a feature to access IBD segments from the tree sequence, but its documentation warns that deriving and storing the IBD segments requires a lot of time and memory (Baumdicker et al., 2021). Another coalescent method `ARGON` simulates IBD segments as a feature within a much broader ARG-inference program (Palamara, 2016). These methods are the two current options to simulate IBD segments genome-wide in modestly sized samples. The runtime to simulate IBD segments with these programs has not been extensively benchmarked.

Long IBD segments can be informative about recent demographic changes (Browning and Browning, 2015; Browning et al., 2018; Cai et al., 2023; Palamara et al., 2012), recent positive selection (Browning and Browning, 2020; Temple et al., 2024), population-specific recombination rates (Zhou et al., 2020a), mutation rates (Tian et al., 2019), allelic conversions (Browning and Browning, 2024), rare variant association studies (Browning and Thompson, 2012; Chen et al., 2023), and close familial relatedness (Zhou et al., 2020c), whereas summary statistics like the fixation index F_{ST} (Weir and Cockerham, 1984) and Tajima's D (Tajima, 1989) or models like the sequentially Markovian coalescent (SMC) (Li and Durbin, 2011), and its extensions (Schiffels and Durbin, 2014), concern population divergences and old selection events (Tajima, 1989; Weir and Cockerham, 1984), among other things. Methods using IBD segments thus serve as an important complementary approach to summary statistics and coalescent-based methods.

Distinguishing between alleles that are identical-by-state versus those that are identical-by-descent from a common ancestor can be challenging. Only those haplotypes extending over multiple centiMorgans, a unit of genetic distance to be defined in Section 2, can be detected as IBD with high accuracy (Freyman et al., 2021; Nait Saada et al., 2020; Naseri et al., 2019; Shemirani et al., 2021; Zhou et al., 2020b). We refer to IBD segments longer than a fixed Morgans threshold as "detectable", where a user-defined threshold can depend on the dataset, the IBD segment detection method, and the tolerance to detection inaccuracies. Exceptionally long IBD segments are rare to observe outside of family studies, meaning that large sample sizes are required to observe enough for IBD-based analyses in outbred population studies.

Some methods require IBD data for the entire chromosomes (Browning and Browning, 2015; Palamara et al., 2012; Temple, 2024; Zhou et al., 2020a,c), which simulators like `msprime` (Baumdicker et al., 2021) and `ARGON` (Palamara, 2016) are suited for. Other statistical inferences concern estimator consistency (Temple et al., 2024), uncertainty quantification (Temple et al., 2024), and convergence to an asymptotic distribution (Temple and Thompson, 2024) around a single locus. Validating such theoretical results involves enormous simulations, for which `msprime` and `ARGON` are less suited.

In this work, we propose an algorithm to simulate IBD segments overlapping a focal location that is fast enough to validate asymptotic properties like consistency, confidence interval coverage, and weak convergence (Casella and Berger, 2002). We modify a naive approach (Temple et al., 2024), and then we argue that our modified approach should drastically decrease runtime with high probability. We demonstrate in some simulation examples that the modified algorithm's average runtime scales approximately linearly with sample size, not quadratically.

2 Preliminary material

Backward simulation of IBD segment lengths overlapping a focal location involves two waiting time distributions: the time until a common ancestor and the genetic length until a crossover. Figure 1 illustrates the coalescent and recombination processes. Here, we formally define a parametric model for IBD segments overlapping a specific locus in terms of these processes.

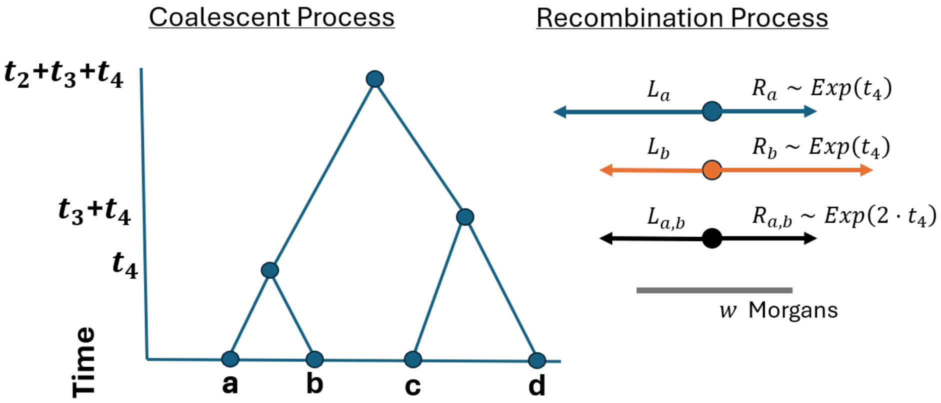

Fig. 1 Conceptual framework for IBD segment lengths. (Left) Sample haplotypes a, b, c, d trace their lineages back to common ancestors at times $t_4, t_4+t_3, t_4+t_3+t_2$. (Right) Relative to a focal point, the haplotype segments lengths R_a, R_b, L_a, L_b are independent, identically distributed Exponential(t_4).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

2.1 The time until a common ancestor

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

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

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

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

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

when k is much smaller than $\min_t N(t)$ (Hein et al., 2005). The geometric model assumes that multiple coalescent events in a single generation are improbable. Its rate $\binom{k}{2}/N(\tau)$ is the probability that any two of k haploids have the same haploid parent at generation τ.

The Kingman coalescent (Kingman, 1982b,a) comes from the continuous time limit of Equations 1 and 2 for large constant population size N. Specifically, T_k converges weakly to Exponential$(\binom{k}{2})$ for $k \ll N$, $N \to \infty$, and time is scaled in units of N generations. Henceforth, we consider the positive real-valued T_k in units of N generations. Varying population sizes $N(t)$ are implemented by rescaling time *post-hoc* in a coalescent with constant population size N (Hein et al., 2005).

2.2 The distance until crossover recombination

The genetic distance between two points is the expected number of crossovers between them in an offspring gamete. This unit of haplotype segment length is the Morgan. Assuming no interference in double-stranded breaks and that crossovers occur randomly and independently, Haldane (1919) derives that the genetic distance until crossover recombination is exponentially distributed, with the Poisson process modeling the crossover points along the genome. The number of crossovers between two points is then Poisson distributed with mean equal to the genetic distance between the two points, which leads to the Haldane map function connecting Morgans to the recombination frequency. (The Haldane map function is $\rho = 0.5(1 - \exp(-2d))$, where ρ is the recombination frequency and d is the genetic distance.)

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

Let a and b be sample haplotypes in the current generation. Define $L_a, R_a \,|\, t \sim$ Exponential(t) to be sample haplotypes a's recombination endpoints to the left and right of a focal location. Since crossovers to the left and right are independent, the extant width derived from the ancestor at time t is $W_a := L_a + R_a \,|\, t \sim \text{Gamma}(2, t)$. Because recombination events are independent in the t meioses descending to a and b from their common ancestor, the IBD segments that are shared by a and b are $L_{a,b}, R_{a,b} \,|\, t \sim$ Exponential($2t$) and $W_{a,b} \,|\, t \sim \text{Gamma}(2, 2t)$. Under this model, the lengths of IBD segments are thus shorter, with a higher probability the more removed its common ancestor is from the present day. This fact is a key motivation for the fast algorithm we develop.

3 An efficient algorithm to simulate identity-by-descent segments

Based on Sections 2.1 and 2.2, the blueprint to simulate IBD segment lengths around a locus is as follows: 1) simulate a coalescent tree for a sample from a population, 2) draw recombination endpoints to the left and right of a focal point at each coalescent event, and 3) derive from the recombination endpoints the haplotype segment lengths that are shared IBD. The third step involves calculating the minimum lengths to the right and left of a focal point for every pair of haplotypes, which is the computational bottleneck in simulating IBD segment lengths. Making fewer haplotype comparisons, without sacrificing the exactness of simulation, is the way to decrease compute times.

In Algorithm 1, we state the method to simulate long IBD segments around a single locus. We make four modifications to the naive simulation algorithm, which are designed to reduce compute times when the primary goal is to generate IBD segments longer than some detection threshold. These implementations reduce compute times due to the mathematical properties of the coalescent time and recombination endpoint distributions.

First, whenever there is likely to be more than one coalescent event in a Wright-Fisher (WF) generation, we approximate the sampling of haploid parents as a binomial random variable (Section 4). Second, we exchange the Kingman coalescent for the discrete-time WF model once the number of non-coalesced haploids is much smaller than the population sizes. This implementation is similar to the hybrid simulation approach in Bhaskar et al. (2014). Third, we do not consider a sample haplotype for IBD segment calculation at future coalescent events once its haplotype segment length is less than the specified detection threshold, which we refer to as “pruning”. In Section

5, we elaborate on the rare probability of long haplotype segments in large populations. Fourth, we combine two sample haplotypes for IBD segment calculation at future coalescent events if they share the same left and right recombination endpoints, which we refer to as "merging". In Section 6, we derive results concerning the probability of merging. We implement pruning and merging using object-oriented programming.

4 An approximation of the Wright-Fisher process in large samples

Simulating the Kingman coalescent is much faster than simulating the discrete-time WF process. The accuracy of the Kingman coalescent requires that the sample size is much smaller than the population size. This requirement is so that the probability of there being more than one coalescent event in a generation is small. The assumption that the sample size is small relative to the population size can be violated in analyses of human biobanks. Under this violation, the coalescent approximation can deviate significantly from the exact discrete-time WF model (Bhaskar et al., 2014; Palamara, 2016; Wakeley and Takahashi, 2003).

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

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

Algorithm 1 Efficient simulation of IBD segment lengths

```
Input: sample size $n$, population sizes $N(t)$, Morgans threshold $w$
Output: Detectable IBD segment lengths $\ell_{a,b} \geq w$ for $a, b \in \{1, \ldots, n\}$

Let current node set $\mathcal{N} = \{1, \ldots, n\}$
Initialize endpoints $l_a, r_a = \infty$ and the latest interior time $v_a$ for all $a \in \mathcal{N}$,
current sample size $k = n$, and current coalescent time $t = 0$

Simulate a coalescent tree
while $k > 2$ do
    Iterate $t$ up by 1
    if not $k^3 \ll N(t)^3$ then
        Draw $X \sim \text{Binomial}(\binom{k}{2}, N(t)^{-1})$ (or Poisson in the limit)
        for 1 ... $X$ do
            Choose haplotypes $a, b \in \mathcal{N}$ to coalesce
            Remove $a, b$ from $\mathcal{N}$
            Add a coalesced node to $\mathcal{N}$
            Iterate $k$ down by 1
    else
        Draw ancestors from $\{1, \ldots, N(t)\}$ for $a \in \mathcal{N}$
        if $a, b$ have a common ancestor then
            Coalesce them and iterate $k$ down

Simulate recombination endpoints
Initialize $\tau = 1$
while $\tau <= t$ do
    for coalescent event at time $\tau$ do
        for each sample $a$ under the subtree do
            Draw $l'_a, r'_a \sim \text{Exponential}(\tau - v_a)$.
            Update endpoints $l_a = \min(l_a, l'_j), r_j = \min(r_a, r'_a)$ and latest time $v_a = \tau$

            Pruning
            if $l_a + r_a < w$ then
                Ignore all future updates and comparisons for $a$

        Merging
        for each pair $a, b$ do
            if $l_a = l_b$ and $r_a = r_b$ then
                Merge nodes together

        for each pair $a, b$ under the subtree that is not yet compared do
            Record IBD segment length if $\min(l_a, l_b) + \min(r_a, r_b) > w$
    Iterate $\tau$ up by 1

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

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

$$\mathbb{E}[k'] \approx N'\left(k/N' - \binom{k}{2} \times N'^{-2}\right) = k - \binom{k}{2} \times N'^{-1}. \tag{4}$$

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

Compared to drawing a parent for each child and then scanning a vector of size k for siblings, simulating the number of coalescent events in one generation from Binomial$(\binom{k}{2}, N'^{-1})$ can be an efficient approximation. The last term being subtracted in Equation 4 is equal to the expected value of a Binomial random variable of $\binom{k}{2}$ trials with success probability N'^{-1}. Next, let A_1 and A_2 be the number of children from two specific haploid parents among the N' parents in the previous generation. If A_1 and A_2 are independent, then $P(A_1 = a_1, A_2 = a_2) = P(A_1 = a_1) \times P(A_2 = a_2)$. A_1 and A_2 are not independent, but the difference between the left term $P(A_1 = a_1, A_2 = a_2)$ and the right term $P(A_1 = a_1) \times P(A_2 = a_2)$ can be vanishingly small when N' is large. The probability $P(A_1 = a_1, A_2 = a_2)$ is derived by choosing a_1 among k samples to have the same parent and then choosing a_2 among $k - a_1$ samples to have a same parent distinct from the parent of the first a_1 samples.

$$\begin{aligned}
&P(A_1 = a_1, A_2 = a_2) - P(A_1 = a_1) \times P(A_2 = a_2) \\
&= \binom{k}{a_1}(N')^{-a_1} \times \binom{k-a_1}{a_2}(N'-1)^{-a_2} \\
&- \binom{k}{a_1}(N')^{-a_1} \times \binom{k}{a_2}(N')^{-a_2} \\
&\leq \binom{k}{a_1}(N')^{-a_1} \times \binom{k}{a_2}(N')^{-a_2} \\
&= O(k^{a_1+a_2} \cdot N'^{-(a_1+a_2)}).
\end{aligned} \tag{5}$$

If both A_1 and A_2 have two or more children ($\min(a_1, a_2) \geq 2$), then Equation 5 is $o(1)$ when the second order approximation $k^3 = o(N'^3)$ is accurate.

In Algorithm 1, we assume that all simultaneous coalescent events are the result of only two children having the same parent. Bhaskar et al. (2014) have shown that the majority of simultaneous coalescent events in a generation are of this type. Due to the coalescent and WF approximations, our method is not exact with respect to the time until a common ancestor.

5 The probability of detectable haplotype segment lengths

Within tens of generations, most haplotype segment lengths are shrunk by crossovers to a genetic length less than detection thresholds that are used in IBD-based analyses. A Morgans length threshold at least greater than 0.01 is typical in applied research (Browning and Browning, 2015, 2020; Temple et al., 2024; Tian et al., 2019; Zhou et al., 2020a). The probabilities of a detectable haplotype segment to the right of and overlapping a focal location, R_a and W_a, respectively, conditional on coalescent time Nt (in generations), are

$$1 - F_{R_a|t}(w) = \exp(-Ntw), \tag{6}$$

$$1 - F_{W_a|t}(w) = \exp(-Ntw) + Ntw \cdot \exp(-Ntw). \tag{7}$$

Figure S1 shows that the upper tail probabilities of R_a and W_a are decreasing exponentially over Nt generations. The probabilities of haplotype segment lengths greater than 0.01 can be far from zero when the haplotype is descendant from an ancestor within the last 100 generations. The probabilities of haplotype segments lengths greater than 0.02 are nearly zero when they are descendant from an ancestor more than 300 generations ago. (But exponential random variables have heavy upper tail probabilities, so, in large samples, we may detect some long IBD segments descendant from ancestors older than 300 generations.)

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

$$\begin{aligned}\mathbb{E}[T^{+}_{n:k}] = \sum_{l=k}^{n} \mathbb{E}[T_l] &= \sum_{l=k}^{n} \binom{l}{2}^{-1} \\ &= 2 \times \sum_{l=k}^{n} \Big(\frac{1}{l-1} - \frac{1}{l}\Big) \\ &= 2 \times ((k-1)^{-1} - n^{-1}),\end{aligned} \tag{8}$$

where $\binom{l}{2}$ is the rate parameter for the time until a common ancestor is reached for any two of l haploids. For $N = 10,000$ and $n \to \infty$, the expected coalescent time $\mathbb{E}[T^{+}_{n:40}]$ is 512.82 generations. For $N = 100,000$ and $n \to \infty$, the expected coalescent time $\mathbb{E}[T_{n:400}]$ is 501.25. If many recombination endpoint comparisons happen at the coalescence of common ancestors older than five hundred generations ago, many haplotypes can be pruned ahead of time. The pruning technique does not compromise the exactness of simulating IBD segment detectable beyond a length threshold.

6 The probability that recombination endpoints are shared between haplotypes

At some point in the past, two sample haplotypes may share the same recombination endpoints to the left and right of a fixed location. Without loss of generality, let haplotypes a and b coalesce to their common ancestor c at time u, and let haplotypes c and d coalesce to their common ancestor e at time $u+v$. Figures S2 and S3 illustrate the coalescent tree in this scenario. Observe that the recombination endpoints to the right $R_{a,c}, R_{b,c} \sim \text{Exponential}(u)$ and $R_{c,e} \sim \text{Exponential}(v)$.

The merging step in Algorithm 1 serves to avoid comparing both the endpoints of a and b with d when a and b have the same endpoints at time $u + v$. Specifically, if a and b's shared recombination endpoint $R_{c,e}$ is smaller than their separate endpoints $R_{a,c}$ and $R_{b,c}$, we can henceforth treat them as the same haplotype without loss of information (Figure S2). If either of the individual lengths $R_{a,c}$ or $R_{b,c}$ are smaller than the common length $R_{c,e}$, we cannot merge the haplotypes without losing information (Figure S3).

The probability that haplotypes a and b have the same recombination endpoint at time $u+v$ is $v(2u+v)^{-1}$. We derive a result that replaces arbitrary coalescent times u and $u+v$ with double the expected times after the $(n-k)^{\text{th}}$ and $(n-j)^{\text{th}}$ coalescent events, respectively.

Proposition 1. *Let $u/2 = \mathbb{E}[T^+_{n:(k+1)}] = 1/k - 1/n$ and $v/2 = \mathbb{E}[T^+_{n:(j+1)}] - \mathbb{E}[T^+_{n:(k+1)}] = 1/j - 1/k$ (Equation 8). For $j = o(k)$,*

$$P(\min(R_{a,c}, R_{c,d}, R_{c,e}) = R_{c,e} \,|\, u, v) \to 1.$$

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

$$\begin{aligned} P(\min(R_{a,c}, R_{b,c}, R_{c,e}) = R_{c,e} \,|\, u, v) &= \frac{1/j - 1/k}{1/j + 1/k - 2/n} \\ &= \frac{(k-j)n}{(nk + nj - 2kj)} \\ &= \frac{1 - j/k}{(1 + j/k - 2j/n)} \\ &\to 1. \end{aligned}$$

□

The implication of Proposition 1 is that haplotypes that share a recent common ancestor should have the same endpoints at the most distant common ancestors.

Since recombinations to the right and left of a focal location are independent, the result of Proposition 1 extends to simulating IBD segments overlapping a focal location. Figures S2 and S3 illustrate that merging occurs when the minimum recombination endpoints to the left and right of the focal location are drawn for the common ancestor c ($\min(R_{a,c}, R_{b,c}, R_{c,e}) = R_{c,e}$ and $\min(L_{a,c}, L_{b,c}, L_{c,e}) = L_{c,e}$).

7 The number of identity-by-descent comparisons

Pruning and merging should be most effective at reducing runtime if the majority of recombination endpoint comparisons happen at the oldest coalescent events. For these oldest coalescent events, we show that without pruning nor merging the expected number of IBD comparisons is of the same order as the worst-case number of IBD comparisons, which is asymptotically equivalent to the sample size squared.

Consider a random bifurcating tree. Here, and nowhere else, we work downward from the root of the tree. Throughout, we assume that n equals a power of 2 to simplify the floor and ceiling functions $\lfloor n/2^j \rfloor = \lceil n/2^j \rceil$ for $j \in \mathbb{N}$. At the coalescent event T_2, the tree bifurcates into two subtrees. At the coalescent event T_3, the scenario with the worst case number of comparisons is subtrees of size $n/2, n/4$, and $n/4$. In general, at each coalescent event, the worst case is to split in half the largest subtree, depicted in Figure S4A.

Let B_j be the size of one subtree randomly bifurcated from a subtree of size B_{j-1}. Figure S4 illustrates these subtree sizes in the context of a random bifurcating tree. The number of recombination endpoint comparisons is $B_j(B_{j-1} - B_j)$. In Theorem 2, we relate the expected value and covariance of $B_j(B_{j-1} - B_j)$ to the worst-case $n/2^{2j}$ computations. The result concerns a bounded number of standard deviations from the expected value, which is a stronger notion than the expected number of computations $\Theta(\cdot)$. The intuition is that a Binomial$(m, 1/2)$ random variable's coefficient of variation $m^{-1/2}$ converges to 0 as m gets large. The general proof strategy is to recursively apply the law of total covariance and identify the exponents in the dominating terms.

Theorem 2. *Let $B_j \sim$ Binomial$(B_{j-1}, 1/2)$ for bounded index $j \geq 1$ and $B_0 = n$.*

$$\lim_{n\to\infty} \frac{\mathbb{E}[B_j(B_{j-1} - B_j)] + O(1) \cdot Cov^{1/2}(B_j(B_{j-1} - B_j))}{n^2/2^{2j}} = 1. \tag{9}$$

Proof. We must calculate the expected value and the covariance in the numerator. Let $B \sim$ Binomial$(m, 1/2)$.

$$\begin{aligned} \mathbb{E}[B(B-1)] &= \mathbb{E}[B^2] - \mathbb{E}[B] \\ &= m/4 + m^2/4 - m/2 \\ &= m(m-1)/4 \\ &= m(m-1) \cdot 2^{-2\cdot 1}. \end{aligned} \tag{10}$$

Using the law of total expectation, we solve the expected value for $j = 2$.

$$\begin{aligned}\mathbb{E}[B_2(B_1 - B_2)] &= \mathbb{E}[\mathbb{E}[B_2(B_1 - B_2)|B_1]] \\ &= \mathbb{E}[B_1(B_1 - 1) \cdot 2^{-2\cdot 1}] \\ &= n(n-1) \cdot 2^{-2\cdot 1} \cdot 2^{-2\cdot 1} = n(n-1) \times 2^{-2\cdot 2}.\end{aligned} \tag{11}$$

Applying Equation 10 recursively, we derive the general formula

$$\mathbb{E}[B_j(B_{j-1} - B_j)] = n(n-1) \cdot 2^{-2j}. \tag{12}$$

The limit of Equation 12 divided by $n^2 \cdot 2^{-2j}$ is one. Next, we require that the standard deviation is of order less than n^2. Using the law of total covariance, we derive in Lemma 3 that $\mathrm{Cov}(B_j(B_{j-1} - B_j)) \sim n^3$, where $\sim$ means asymptotically equivalent. Consequently,

$$\lim_{n\to\infty} n^{-2} \cdot \mathrm{Cov}^{1/2}(B_j(B_{j-1} - B_j)) = 0.$$

□

We remark that our marginal calculations along one branching path are not the same as deriving the expected number of comparisons at the final j coalescent events, the latter of which depends on the tree topology. Harding (1971) discusses the intractability of calculating probability masses for a tree topology with many leaves, which is a limiting factor in deriving the expected number of comparisons at the final j coalescent events. In Appendix A, we give moment calculations from Dahmer and Kersting (2015) that offer a complementary perspective on the number of IBD comparisons, reiterating that a number of computations $\sim n^2$ should occur at and near the root of the coalescent tree.

8 Empirical results

Temple et al. (2024) and Temple and Thompson (2024) use our algorithm to conduct enormous simulation studies involving sample sizes as large as ten thousand individuals and tens of millions of runs. (Individuals are "diploids", which we implement as a haploid model with the number of haploids equal to the number of individuals times 2.) Their empirical studies are feasible because of the pruning and merging techniques, whose effects on runtime we benchmark in this section. We also benchmark runtimes for `msprime` and `ARGON`, showing that these existing methods to simulate IBD can take more than an hour to complete one run when sample size exceeds five thousand diploid individuals.

8.1 Experimental setup

8.1.1 Demographic scenarios

We consider two complex demographic scenarios and constant population sizes. Figure S5 shows the demographic scenarios graphically. These demographic scenarios are the same as those used in Temple et al. (2024), Temple and Thompson (2024), and Temple (2024). We refer to the complex demographic scenarios as examples of three phases of exponential growth and a population bottleneck. The three phases of exponential growth scenario involves an ancestral population of five thousand individuals that grew exponentially at different rates in three different time periods. This demographic model is similar to the "UK-like" model in Cai et al. (2023). The population bottleneck scenario involves an ancestral population of ten thousand individuals that grew exponentially at a fixed rate but experienced an instantaneous reduction in size twenty generations before the present day.

8.1.2 Hard selective sweeps

We also consider a genetic model for positive selection (Fisher, 1923; Haldane, 1924, 1932) that is described in Crow and Kimura (1970), Temple et al. (2024), and Temple (2024) as well as in many other articles. Briefly, the allele frequency $p_s(t)$ decreases backward in time as a function of a nonnegative selection coefficient s. The selection coefficient reflects the advantage the allele has relative to alternative alleles. The larger the selection coefficient is, the faster the allele frequency increased. Also, the larger the selection coefficient is, the more detectable IBD segments there are on average.

Positive selection around a locus is implemented via a coalescent with two subpopulations: one subpopulation has the sweeping allele, and one subpopulation does not have the sweeping allele. The population sizes are $N_e(t) \cdot p(t)$ and $N_e(t) \cdot (1 - p(t))$. Until the coalescent reaches the sweeping allele's time of *de novo* mutation, IBD segments are not possible between individuals in separate subpopulations.

8.2 Compute times

8.2.1 Simulating identity-by-descent segment lengths around a locus

To assess the effect of the pruning and merging rules, we evaluate four implementation strategies: merging and pruning (Algorithm 1), pruning only, merging only, and neither pruning nor merging (the naive approach). For each implementation, we run five simulations for sample sizes increasing by a factor of 2, recording the average wall clock compute time. The upper bound on sample size that we consider is 128,000 individiuals, which is of the same order as the UK Biobank data (Bycroft et al., 2018).

Figure 2 shows the average runtime per sample size between the implementations. Simulating IBD segment lengths without pruning nor merging takes more than one minute on average for eight thousand samples. Simulating IBD segment lengths with either pruning or merging can take less than one minute for sixty-four samples. Pruning appears to give a larger reduction in compute time than merging. Merging can further

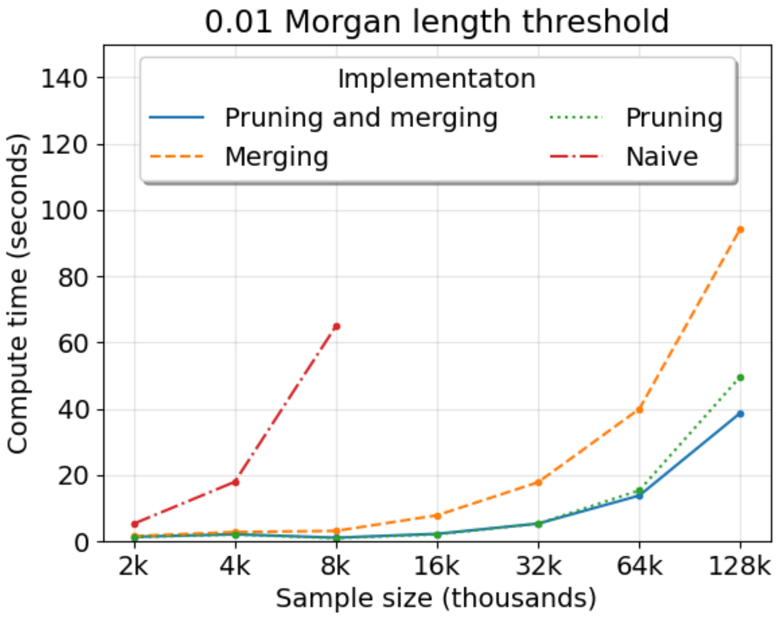

Fig. 2 Compute time to simulate IBD segment lengths around a locus depending on algorithm implementation. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legend denotes colored line styles for implementations using Algorithm 1 as is (blue), merging only (orange), pruning only (green), and neither pruning nor merging (red). The main text describes "merging" and "pruning" techniques. The demography is the population bottleneck. The Morgans length threshold is 0.01.

reduce runtime for sample sizes greater than one hundred thousand. The difference in five to ten seconds can be important when the number of simulations is enormous, as is the case in the Temple et al. (2024) and Temple and Thompson (2024) studies.

One important influence on runtime is the detection threshold. Figure 3A shows the algorithm's average runtime per sample size for different detection thresholds on segment length. With the 0.0025 Morgans cutoff, the quadratic behavior of runtime is visually apparent when more than twenty thousand samples are simulated, whereas the trend is less obvious for detection thresholds greater than 0.0050 Morgans. The algorithm is at least twice as fast on average for detection thresholds ≥ 0.02 Morgans versus those ≤ 0.0050 Morgans.

Another important influence on runtime is the population size. Figure 3B shows the algorithm's average runtime per sample size for different constant population sizes.

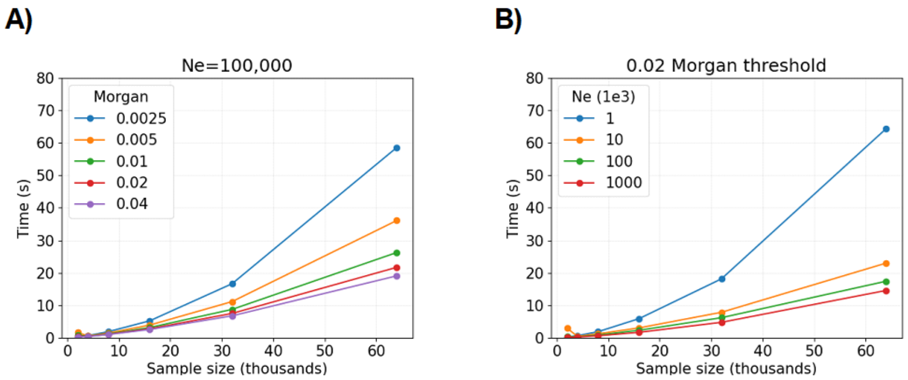

Fig. 3 Compute time to simulate IBD segment lengths around a locus depending on the detection threshold and population size. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legends denote colored line styles for A) different detection thresholds (in Morgans) with $N = 10^5$ fixed or B) different population sizes with 0.02 Morgans fixed.

The algorithm is at least twice as fast on average for population sizes $N \geq 10,000$ versus $N \leq 1,000$. Population sizes are estimated to be at least ten thousand for many model organisms (Adrion et al., 2020; Lauterbur et al., 2023).

Figure S6 shows the algorithm's average runtime per sample size for different demographic scenarios and varying selection coefficients. Simulating IBD segment lengths takes more time for the population bottleneck and three phases of exponential growth scenarios compared to constant-size population scenarios. Runtime increases with the selection coefficient. The highest average measurement is more than four minutes for sixty-four thousand samples, the population bottleneck scenario, and $s = 0.04$.

Now, we perform twenty simulations each for sample sizes $2 \cdot 10^4, 4 \cdot 10^4, 8 \cdot 10^4, 16 \cdot 10^4$, and $32 \cdot 10^4$ and regress on runtime. The linear models in runtimes $\mathbf{Y} \in \mathbb{R}$, sample sizes $\mathbf{X} \in \mathbb{R}$, and regression coefficients $\boldsymbol{\beta}$ be:

$$\mathbf{Y} = \beta_0 \mathbf{X}; \tag{13}$$

$$\mathbf{Y} = \beta_1 \mathbf{X} + \beta_2 \mathbf{X}^2. \tag{14}$$

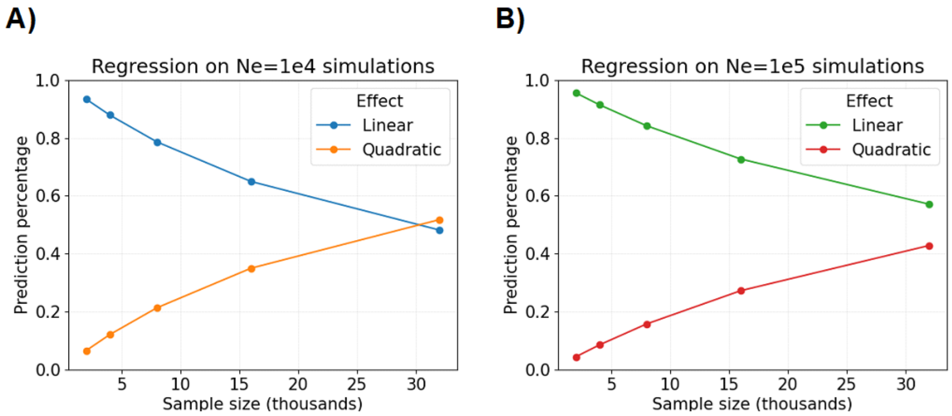

Fig. 4 Percentage of regression model predictions explained by linear and quadratic effects. The percentage of predicted compute time (y-axis) in seconds by sample size (x-axis) in thousands with respect to linear and quadratic effects. Plots show results for A) the constant population size $N \equiv N_e = 10,000$ versus B) $N \equiv N_e = 100,000$. The detectable IBD segments are simulated with a 0.02 Morgans threshold.

We measure the proportion of a fitted value explained by the linear effect as

$$\text{Percentage of prediction from linear component} = \frac{\hat{\beta}_1 x}{\hat{\beta}_1 x + \hat{\beta}_2 \cdot x^2}, \tag{15}$$

where x is a sample size. Figure 4 shows that the linear component in Equation 14 explains more than fifty percent of predictions for sample size $\leq 16 \cdot 10^4$, which we say demonstrates approximately linear computational complexity in this domain. Next, we estimate $\hat{\beta}_0$ equal to 0.0913 and 0.0707 in Equation 13 for population sizes $N = 10,000$ and $N = 100,000$. We interpret this to mean that the expected runtime increases by 0.0913 and 0.0707 seconds for each one unit increase to sample size (in thousands), at least up to $n \leq 16 \cdot 10^4$.

Overall, we benchmark that our simulation algorithm can be run tens of thousands of times within a day on one core processing unit of an Intel 2.2 GHz compute node. Despite performance savings, we observe that our simulation algorithm maintains quadratic behavior in sample size (Figure 2 and Figure S6). One explanation for this finding is that a sizeable fraction of all lineages coalesce at least once in the first few generations when the sample size exceeds ten thousand (Bhaskar et al., 2014).

8.2.2 Simulating identity-by-descent segment lengths from the ancestral recombination graph

We measure the times it takes `ARGON` (Palamara, 2016) and `msprime` (Baumdicker et al., 2021) with `tskibd` (Guo et al., 2024) to simulate detectable IBD segments around a locus. The `tskibd` program concatenates short IBD segments from `msprime` tree sequences into detectable IBD segment lengths. To measure the computing time of these approaches, we do not include the time to simulate an ARG.

We simulate IBD segments ≥ 0.02 Morgans in a 0.07 Morgans region, which is a large enough region to contain all IBD segments ≥ 0.02 Morgans around its central location. Both programs visit nodes in the ARG in small, non-overlapping sliding windows. We consider window sizes of 0.0001 and 0.00001 Morgans in benchmarking runtimes. We compare compute times to those of Algorithm 1 with the 0.02 Morgans detection threshold.

Table 1 reports the average runtimes of each method for increasing sample size in the population bottleneck demographic scenario. `ARGON` takes nearly an hour to simulate the detectable IBD segment lengths of two thousand diploids. We do not run it for more than two thousand diploids due to concerns surrounding quadratic runtimes. With 0.0001 Morgans windows, `tskibd` takes less than twenty minutes to simulate the detectable IBD segment lengths of four thousand diploids and a little over an hour to simulate the IBD length distribution of eight thousand diploids. To get more precise IBD segment endpoints with `tskibd`, we use 0.00001 Morgans windows, which increases runtime eightfold or more. Some true IBD segments will not be detected if the window size is too large, but decreasing the window size increases runtime.

Table 1 Average runtime to simulate detectable IBD segments with `ARGON` and `tskibd`

Method	Samples	Compute Time (s)
`ARGON`[1 2]	500	211.80
	1000	652.70 (≈ 11 min)
	2000	3292.90 (≈ 55 min)
`tskibd`	500	6.80
	1000	28.80
	2000	151.78
	4000	921.60 (≈ 15 min)
	8000	4409.40 (≈ 73 min)

[1] Runtimes to simulate ARGs are not included in the results, which are small to negligible percentages of total runtimes.

[2] Sliding non-overlapping windows are of size 0.0001 Morgans.

In comparison, for eight thousand diploids, our improved approach simulates IBD segments ≥ 0.01 Morgans around a locus in less than two seconds (Figure 2). Even our naive approach completes the same scope of simulations in less than two minutes. Our method is exact for a single locus, whereas `tskibd` may be inexact due to its windowing heuristic. Moreover, our method provides the full locus-specific length distribution whereas `ARGON` and `tskibd` provide a length distribution truncated by the size of the genomic region. Conversely, our method relies on a mathematical construct without regard to finite chromosome sizes, which can result in detectable IBD lengths exceeding the chromosome size.

9 Discussion

To efficiently simulate IBD segment lengths overlapping a focal location, we exploit the fact that small values occur with high probability in Gamma random variables. Fast simulation in population genetics is important for statistical methods like approximate Bayesian computation (Beaumont et al., 2002), importance sampling (Browning, 2000; Stern et al., 2019), and neural network learning (Korfmann et al., 2023). Our method was developed with the evaluation of statistical consistency (Casella and Berger, 2002), parametric bootstrapping (Efron, 1987), and asymptotic distributions (Temple and Thompson, 2024) in mind.

Existing methods ARGON (Palamara, 2016) and tskibd (Guo et al., 2024) simulate IBD segment lengths for genomic region sizes less than 0.10 Morgans and thousands of samples within hours to days. These runtime performances are insufficient for the aforementioned methods and analyses, in particular parametric bootstrapping (Temple et al., 2024; Efron, 1987). We benchmark that our average runtime scales approximately linear as the number of haplotype pairs scales quadratically in sample size, taking as little as a couple of seconds or tens of seconds for sample sizes of order 10^4 or 10^5, respectively. The pruning and merging techniques presented here for a single locus could motivate changes to ARGON and tskibd that improve the runtime of genome-wide IBD simulations.

Related studies have already used our algorithm to these ends. Running our algorithm tens of millions of times with samples sizes ≥ 5000, Temple and Thompson (2024) show simulation results that are consistent with the conditions of their central limit theorems. Running our algorithm millions of times with a sample size of five thouand diploids, Temple et al. (2024) show that ninety-five percent parametric bootstrap intervals for a selection coefficient estimator contain the true parameter in ninety percent of simulations. They also show that exploring the effects of sample size and detection threshold on selection coefficient estimation is feasible on a laptop. Temple (2024) assesses the tradeoffs between standard normal and percentile-based confidence intervals for the Temple et al. (2024) selection coefficient estimator. Temple (2024) also shows how to calculate the statistical power in an excess IBD rate scan as the magnitude of directional selection increases. These studies would otherwise have been computationally intractable using the existing methods ARGON and tskibd. Indeed, the scope of the Temple and Thompson (2024) simulations amounts to hundreds of days of computing time even with our efficient algorithm.

The algorithm may assist in developing IBD clustering methods as well. A previously published method to simulate IBD cluster sizes comparable to those observed in human data is based solely on heuristics (Shemirani et al., 2023), whereas our method is an exact simulation of the process. Temple et al. (2024) developed their method to find abnormally large IBD clusters by experimenting with our simulations. Generating IBD clusters, which are qualitatively different from Erdos-Renyi networks (Temple and Thompson, 2024), could be fruitful in IBD network analyses (Shemirani et al., 2023). The distribution of IBD cluster sizes could also help benchmark multi-way IBD segment detection (Browning and Browning, 2024).

Statements and Declarations

Funding

S.D.T. acknowledges funding support from the US Department of Defense National Defense Science and Engineering Graduate Fellowship, the US National Institutes of Health T32 GM081062 Predoctoral Training Grant in Statistical Genetics, and Schmidt Sciences, LLC. This research was supported in part by the US National Human Genome Research Institute of the National Institutes of Health under award number HG005701.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Data availability

Not applicable

Materials availability

Not applicable

Code availability

The algorithm is in a Python package (https://github.com/sdtemple/isweep), which is available under the CC0 1.0 Universal License.

Author contributions

S.D.T. proposed the study, planned the study, wrote the software, conducted the analysis, and wrote the manuscript. S.D.T. and S.R.B. developed the algorithm. S.D.T. and E.A.T. derived the theoretical results. All authors contributed to editing the manuscript.

Appendix A The number of identity-by-descent comparisons

A.1 Computations near the root of the coalescent tree

Lemma 3. *Recall that $B_j \sim$ Binomial$(B_{j-1}, 1/2)$ for $j \geq 1$ and $B_0 = n$. Then,*

$$Cov(B_j(B_{j-1} - B_j)) \sim n^3. \tag{A1}$$

Proof. We apply the laws of total covariance and expectation in a recursive fashion. Overall, we must control three terms:

$$\begin{aligned}\mathrm{Cov}(B_j(B_{j-1} - B_j)) = \mathrm{Cov}(B_j^2, B_j^2)& \\ + \mathrm{Cov}(B_j B_{j-1}, B_j B_{j-1})& \\ - 2 \cdot \mathrm{Cov}(B_j B_{j-1}, B_j^2).&\end{aligned} \tag{A2}$$

The first four moments of the conditional binomial random variable are useful:

$$\begin{aligned}\mathbb{E}[B_j|B_{j-1}] &= 0.5 \cdot B_{j-1}; \\ \mathbb{E}[B_j^2|B_{j-1}] &= 0.5^2(B_{j-1} + B_{j-1}^2); \\ \mathbb{E}[B_j^3|B_{j-1}] &= 1 \cdot 0.5 \cdot B_{j-1} \\ &\quad + 3 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\ &\quad + 1 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2); \\ \mathbb{E}[B_j^4|B_{j-1}] &= 1 \cdot 0.5 \cdot B_{j-1} \\ &\quad + 7 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\ &\quad + 6 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2) \\ &\quad + 1 \cdot 0.5^4 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2)(B_{j-1} - 3).\end{aligned} \tag{A3}$$

The following conditional covariances are also useful.

$$\mathrm{Cov}(B_j, B_j|B_{j-1}) = 0.5^2 \cdot B_{j-1}^1 = O(B_{j-1}^1). \tag{A4}$$

$$\begin{aligned}
\mathrm{Cov}(B_j, B_j^2|B_{j-1}) &= \mathbb{E}[B_j^3|B_{j-1}] - \mathbb{E}[B_j|B_{j-1}] \cdot \mathbb{E}[B_j^2|B_{j-1}] \\
&= 1 \cdot 0.5 \cdot B_{j-1} \\
&+ 3 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\
&+ 1 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2) \\
&- 0.5^3 \cdot B_{j-1}(B_{j-1} + B_{j-1}^2) \\
&= 0.5^2 \cdot B_{j-1}^2 \\
&= O(B_{j-1}^2).
\end{aligned} \tag{A5}$$

$$\begin{aligned}
\mathrm{Cov}(B_j^2, B_j^2|B_{j-1}) &= \mathbb{E}[B_j^4|B_{j-1}] - \mathbb{E}[B_j^2|B_{j-1}] \cdot \mathbb{E}[B_j^2|B_{j-1}] \\
&= 1 \cdot 0.5 \cdot B_{j-1} \\
&+ 7 \cdot 0.5^2 \cdot B_{j-1}(B_{j-1} - 1) \\
&+ 6 \cdot 0.5^3 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2) \\
&+ 1 \cdot 0.5^4 \cdot B_{j-1}(B_{j-1} - 1)(B_{j-1} - 2)(B_{j-1} - 3) \\
&- 0.5^4(B_{j-1}^2 + 2 \cdot B_{j-1}^3 + B_{j-1}^4) \\
&= -0.5^3 \cdot B_{j-1} - 0.5^3 \cdot B_{j-1}^2 + 0.5 \cdot B_{j-1}^3 \\
&= O(B_{j-1}^3).
\end{aligned} \tag{A6}$$

Notice that all of the conditional covariances are of an order of three or less. By recursively applying the law of total expectation, we derive $\mathbb{E}[B_j^3] \sim n^3$. Another important unconditional covariance term is

$$\begin{aligned}\mathrm{Cov}(B_j, B_j^2) &= \mathbb{E}[\mathrm{Cov}(B_j, B_j^2 | B_{j-1})] + \mathrm{Cov}(\mathbb{E}[B_j | B_{j-1}], \mathbb{E}[B_j^2 | B_{j-1}]) \\ &= \mathbb{E}[O(B_{j-1}^2)] + \mathrm{Cov}(O(B_{j-1}), O(B_{j-1}^2)),\end{aligned} \tag{A7}$$

which is asymptotically equivalent to n^2 when the total laws of expectation and covariance are applied recursively. Finally, we evaluate the asymptotic behavior of the three unconditional covariances in Equation A2.

$$\begin{aligned}\mathrm{Cov}(B_j^2, B_j^2) &= \mathbb{E}[\mathrm{Cov}(B_j^2, B_j^2 | B_{j-1})] + \mathrm{Cov}(\mathbb{E}[B_j^2 | B_{j-1}], \mathbb{E}[B_j^2 | B_{j-1}]) \\ &= \mathbb{E}[O(B_{j-1}^3)] + 0.5^4 \cdot \mathrm{Cov}(B_{j-1} + B_{j-1}^2, B_{j-1} + B_{j-1}^2)\end{aligned} \tag{A8}$$

$$\begin{aligned}\mathrm{Cov}(B_{j-1}B_j, B_{j-1}B_j) &= \mathbb{E}[B_{j-1}^2 \mathrm{Cov}(B_j, B_j | B_{j-1})] \\ &\quad + \mathrm{Cov}(B_{j-1} \cdot \mathbb{E}[B_j | B_{j-1}], B_{j-1} \cdot \mathbb{E}[B_j | B_{j-1}]) \\ &= 0.5^2 \cdot (\mathbb{E}[B_{j-1}^3] + \mathrm{Cov}(B_{j-1}^2, B_{j-1}^2))\end{aligned} \tag{A9}$$

$$\begin{aligned}\mathrm{Cov}(B_{j-1}B_j, B_j^2) &= \mathbb{E}[B_{j-1} \cdot \mathrm{Cov}(B_j, B_j^2 | B_{j-1})] \\ &\quad + \mathrm{Cov}(B_{j-1} \cdot \mathbb{E}[B_j | B_{j-1}], \mathbb{E}[B_j^2 | B_{j-1}]) \\ &= \mathbb{E}[O(B_{j-1}^3)] + 0.5^3 \cdot \mathrm{Cov}(B_{j-1}^2, B_{j-1} + B_{j-1}^2)\end{aligned} \tag{A10}$$

By recursively applying the total laws of expectation and covariance, we conclude that Equations A8, A9, and A10 are asymptotically equivalent to n^3. □

A.2 Computations near the leaves of the coalescent tree

A complementary perspective on joint subtree sizes we take from Dahmer and Kersting (2015). Now, we work upward from the leaves to the root. Dahmer and Kersting (2015) provide a lemma for the expected number of subtrees containing r sample haplotypes at the $(n-k)^{\text{th}}$ coalescent event. We can use their moment calculations together with the expected value of the hypoexponential random variable $T_{n:k}^{+}$ to build intuition for the average subtree sizes at a specified generation. In a toy example, Figure S7 shows the average number of sample haplotypes under subtrees of a given size at generations $N \cdot \mathbb{E}[T^{+}_{n:k}]$. Our main observation is that before the final coalescent events, which occur at expected times proportional to population size N, most sample haplotypes are expected to be under a subtree of size an order of magnitude smaller than the sample size.

References

Adrion, J.R., Cole, C.B., Dukler, N., Galloway, J.G., Gladstein, A.L., Gower, G., Kyriazis, C.C., Ragsdale, A.P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R.A., Durvasula, A., Gronau, I., Kim, B.Y., McKenzie, P., Messer, P.W., Noskova, E., Ortega-Del Vecchyo, D., Racimo, F., Struck, T.J., Gravel, S., Gutenkunst, R.N., Lohmueller, K.E., Ralph, P.L., Schrider, D.R., Siepel, A., Kelleher, J., Kern, A.D.: A community-maintained standard library of population genetic models. Elife **9** (2020)

Browning, S.R., Browning, B.L.: Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am. J. Hum. Genet. **97**(3), 404–418 (2015)

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

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

Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., Laurie, C.C.: Ancestry-specific recent effective population size in the Americas. PLoS Genet. **14**(5) (2018)

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A.P., Tsambos, G., Zhu, S., Eldon, B., Castedo Ellerman, E., Galloway, J.G., Gladstein, A.L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschmar, W.W., Lohse, K., Matschiner, M., Nelson, D., Pope, N.S., Quinto-Cortés, C.D., Rodrigues, M.F., Saunack, K., Sellinger, T., Thornton, K.R., Kemenade, H., Wohns, A., Wong, H.Y., Gravel, S., Kern, A., Koskela, J., Ralph, P.L., Kelleher, J.: Efficient ancestry and mutation simulation with msprime 1.0. Genetics **220**(3) (2021)

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

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.*: The UK Biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)

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

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

Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian computation in population genetics. Genetics **162**(4), 2025–2035 (2002)

Casella, G., Berger, R.L.: Statistical Inference, 2nd edn. Thomson Learning, Australia; Pacific Grove, CA (2002)

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

Crow, J.F., Kimura, M.: An Introduction to Population Genetics Theory. Harper & Row, New York, NY (1970)

Chen, H., Naseri, A., Zhi, D.: FiMAP: A fast identity-by-descent mapping test for biobank-scale cohorts. PLoS Genet. **19**(12) (2023)

Dahmer, I., Kersting, G.: The internal branch lengths of the Kingman coalescent. Ann. Appl. Probab. **25**(3), 1325–1348 (2015)

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

Ewing, G., Hermisson, J.: MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**(16), 2064–2065 (2010)

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

Freyman, W.A., McManus, K.F., Shringarpure, S.S., Jewett, E.M., Bryc, K., 23 and Me Research Team, Auton, A.: Fast and robust identity-by-descent inference with the templated positional Burrows-Wheeler transform. Mol. Biol. Evol. **38**(5), 2131–2151 (2021)

Guo, B., Borda, V., Laboulaye, R., Spring, M.D., Wojnarski, M., Vesely, B.A., Silva, J.C., Waters, N.C., O'Connor, T.D., Takala-Harrison, S.: Strong positive selection biases identity-by-descent-based inferences of recent demography and population structure in plasmodium falciparum. Nat Commun **15**(1), 2499 (2024)

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

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

Haldane, J.B.S.: The Causes of Evolution. Harper & Row, New York, NY (1932)

Harding, E.F.: The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Probab. **3**(1), 44–77 (1971)

Hoban, S., Bertorelle, G., Gaggiotti, O.E.: Computer simulations: tools for population and evolutionary genetics. Nat. Rev. Genet. **13**(2), 110–122 (2012)

Haller, B.C., Galloway, J., Kelleher, J., Messer, P.W., Ralph, P.L.: Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Mol. Ecol. Resour. **19**(2), 552–566 (2019)

Hudson, R.R., Kaplan, N.L.: The coalescent process in models with selection and recombination. Genetics **120**(3), 831–840 (1988)

Haller, B.C., Messer, P.W.: SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Mol. Biol. Evol. **46**(3), 632–637 (2019)

Hein, J., Schierup, M., Wiuf, C.: Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, New York, NY, USA (2005)

Hudson, R.R.: Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**(2), 183–201 (1983)

Hudson, R.R.: ms a program for generating samples under neutral models. Bioinformatics **18**(2), 337–338 (2002)

Kaplan, N.L., Darden, T., Hudson, R.R.: The coalescent process in models with selection. Genetics **120**(3), 819–829 (1988)

Korfmann, K., Gaggiotti, O.E., Fumagalli, M.: Deep learning in population genetics. Genome Biol. Evol. **15**(2) (2023)

Kingman, J.F.C.: The coalescent. Stoch. Process. Their Appl. **13**(3), 235–248 (1982)

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

Kern, A.D., Schrider, D.R.: Discoal: flexible coalescent simulations with selection. Bioinformatics **32**(24), 3839–3841 (2016)

Lauterbur, E.M., Cavassim, M.I.A., Gladstein, A.L., Gower, G., Pope, N.S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., Cury, J., Echevarria, I., Haller, B.C., Hasan, A.R., Huang, X., Iasi, L.N.M., Noskova, E., Obšteter, J., Pavinato, V.A.C., Pearson, A., Peede, D., Perez, M.F., Rodrigues, M.F., Smith, C.C.R., Spence, J.P., Teterina, A., Tittes, S., Unneberg, P., Vazquez, J.M., Waples, R.K., Wohns, A.W., Wong, Y., Baumdicker, F., Cartwright, R.A., Gorjanc, G., Gutenkunst, R.N., Kelleher, J., Kern, A.D., Ragsdale, A.P., Ralph, P.L., Schrider, D.R., Gronau, I.: Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. Elife **12** (2023)

Li, H., Durbin, R.: Inference of human population history from individual whole-genome sequences. Nature **475**(7357), 493–496 (2011)

Nath, H.B., Griffiths, R.C.: The coalescent in two colonies with symmetric migration. J. Math. Biol. **31**(8), 841–851 (1993)

Naseri, A., Liu, X., Tang, K., Zhang, S., Zhi, D.: RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. Genome Biol. **20**, 1–15 (2019)

Nait Saada, J., Kalantzis, G., Shyr, D., Cooper, F., Robinson, M., Gusev, A., Palamara, P.F.: Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. Nat. Commun. **11**(1), 6130 (2020)

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

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

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

Shemirani, R., Belbin, G.M., Burghardt, K., Lerman, K., Avery, C.L., Kenny, E.E., Gignoux, C.R., Ambite, J.L.: Selecting clustering algorithms for identity-by-descent mapping. In: Pacific Symposium on Biocomputing 2023, pp. 121–132 (2023)

Schiffels, S., Durbin, R.: Inferring human population size and separation history from multiple genome sequences. Nature genetics **46**(8), 919–925 (2014)

Stern, A.J., Wilton, P.R., Nielsen, R.: An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. **15**(9) (2019)

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

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

Temple, S.D.: Statistical inference using identity-by-descent segments: Perspectives on recent positive selection. PhD thesis, University of Washington (2024)

Temple, S.D., Thompson, E.A.: Identity-by-descent segments in large samples. bioRxiv (2024) https://doi.org/10.1101/2024.06.05.597656

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

Weir, B.S., Cockerham, C.C.: Estimating F-statistics for the analysis of population structure. Evolution **38**(6), 1358–1370 (1984)

Wright, S.: Evolution in mendelian populations. Genetics **16**(2), 97–159 (1931)

Wakeley, J., Takahashi, T.: Gene genealogies when the sample size exceeds the effective size of the population. Mol. Biol. Evol. **22**(2), 208–213 (2003)

Yuan, X., Miller, D.J., Zhang, J., Herrington, D., Wang, Y.: An overview of population genetic data simulation. J. Comput. Biol. **19**(1), 42–54 (2012)

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

Zhou, Y., Browning, S.R., Browning, B.L.: A fast and simple method for detecting identity-by-descent segments in large-scale data. Am. J. Hum. Genet. **106**(4), 426–437 (2020)

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

709 Supplementary figures

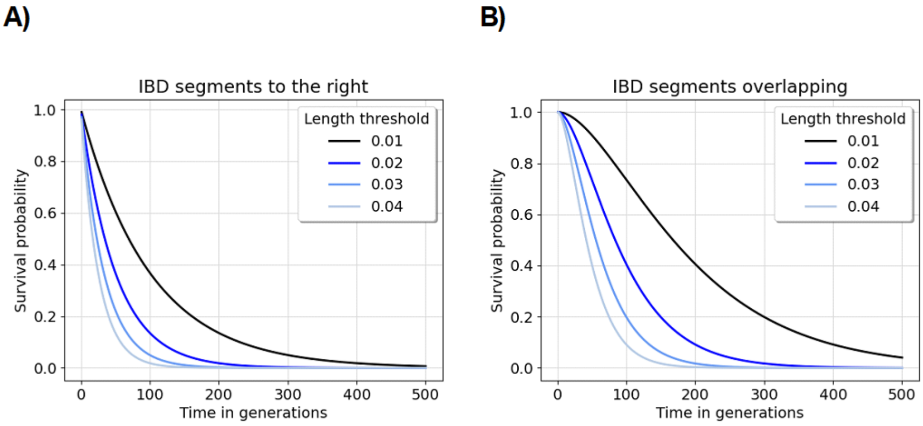

Fig. S1 The upper tail probabilities of Gamma random variables. Subplots A) and B) show the survival probabilities for shape parameters 1 and 2, respectively. The rate of the random variables is the coalescent time in generations (x-axis). The survival probability (y-axis) comes from Equations 6 and 7. The length thresholds are denoted by different colors and line styles, defined in the legend.

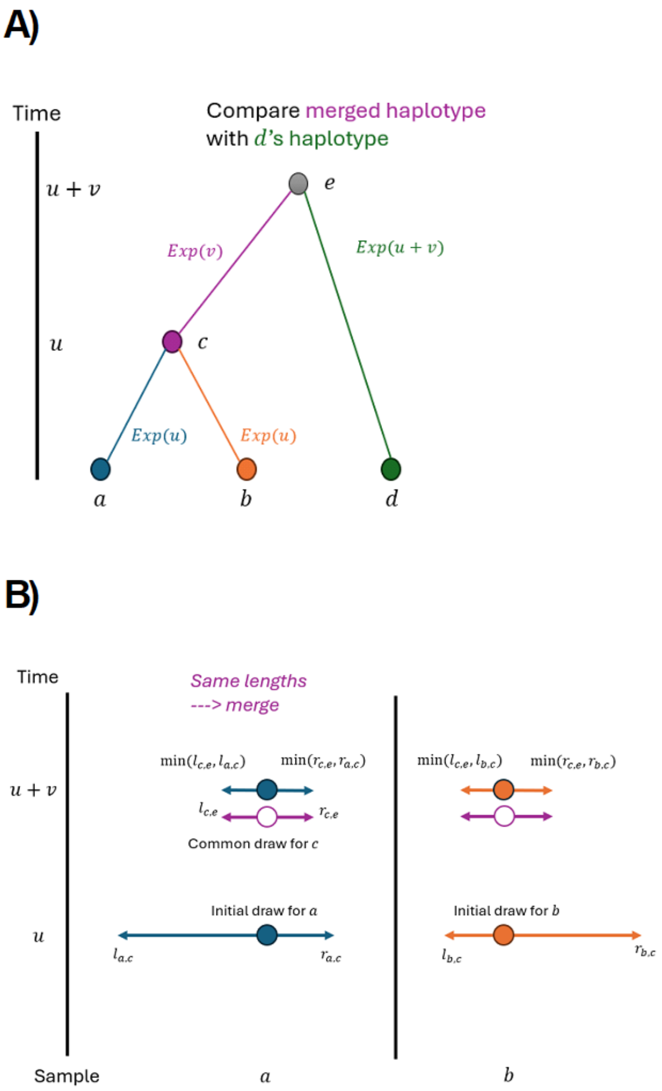

Fig. S2 Illustration of merging haplotypes. A) We draw recombination endpoints to the left and right of the focal location from Exponential(u) for both sample haplotypes a and b at coalescent time u. We draw recombination endpoints to the left and right of the focal location from Exponential(v) for the common ancestor c of a and b at coalescent time $u+v$. Colors denote branch lengths and recombination endpoints corresponding to a given haplotype. $l_{a,c}$ and $r_{a,c}$ denote the endpoints for a drawn to the left and right of the focal location at time u (lowercase denotes observation of random variables). B) We compute minimums of lengths drawn for c and a and c and b, respectively. We merge the sample haplotypes a and b in future calculations because the minimum lengths are both the recombination endpoints drawn from the common ancestor c. When comparing recombination endpoints at time $u+v$ with those of the haplotype d, we make one comparisons. The haplotypes remain longer than the detection threshold w Morgans.

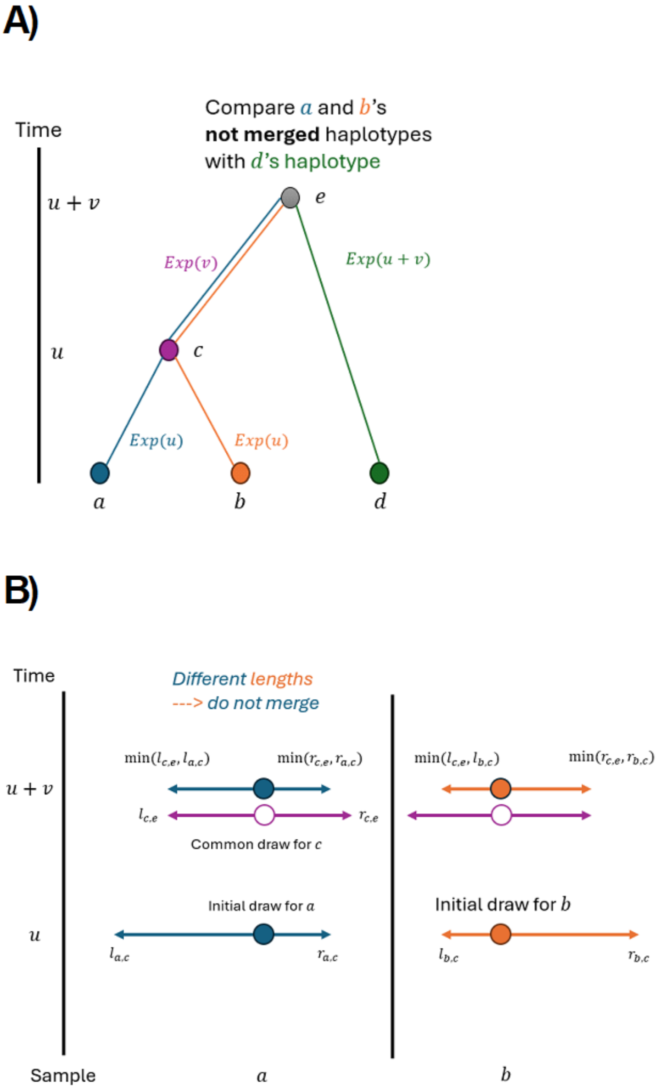

Fig. S3 Illustration of ***not*** merging haplotypes. A) We draw recombination endpoints to the left and right of the focal location from Exponential(u) for both sample haplotypes a and b at coalescent time u. We draw recombination endpoints to the left and right of the focal location from Exponential(v) for the common ancestor c of a and b at coalescent time $u + v$. Colors denote branch lengths and recombination endpoints corresponding to a given haplotype. $l_{a,c}$ and $r_{a,c}$ denote the endpoints for a drawn to the left and right of the focal location at time u (lowercase denotes observation of random variables). B) We compute minimums of lengths drawn for c and a and c and b, respectively. We ***do not*** merge the sample haplotypes a and b in future calculations because the minimum lengths are ***not*** both the recombination endpoints drawn from the common ancestor c. When comparing recombination endpoints at time $u+v$ with those of the haplotype d, we make ***two*** comparisons. The haplotypes remain longer than the detection threshold w Morgans.

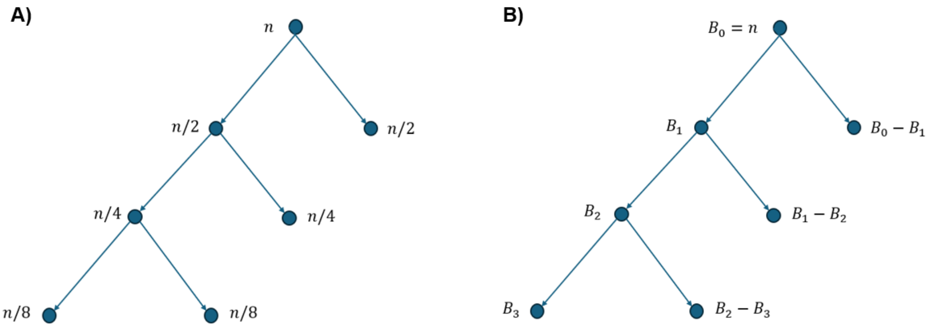

Fig. S4 Illustration of worst-case subtree sizes in random bifurcating tree. A) The worst-case subtree sizes of ancestors (dots) are when each bifurcation is an even split. B) The model $\{B_j\}$ of subtree sizes down one branching path is defined as random variables. The sample size is denoted as n.

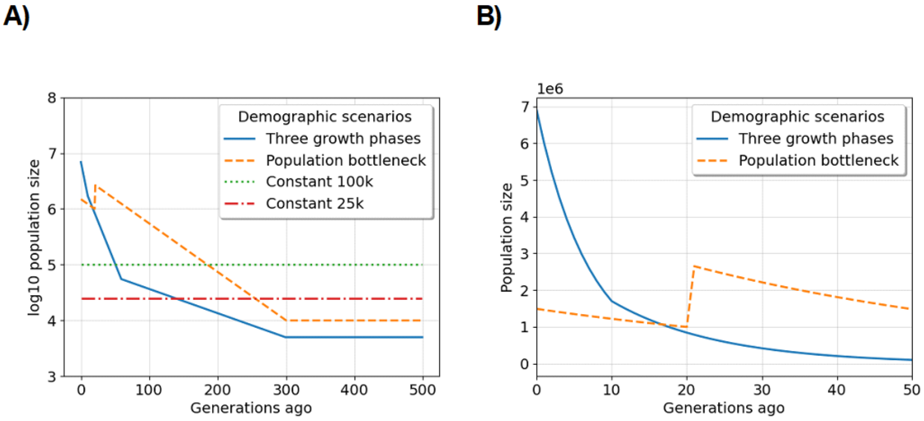

Fig. S5 Demographic scenarios we consider in simulation studies: A) coalescent time in generations ago by the log 10 population size, and B) the most recent fifty generations by population size for examples of exponential growth. The legends specify the color and line style for each scenario. As opposed to coalescent time used in the main text, we describe the scenarios moving forward in time here. The three phases of exponential growth model is as follows: a population of ancestral size five thousand diploids increases exponentially each generation at rates one, seven, and fifteen percent starting three hundred, sixty, and ten generations ago. This demographic model is similar to the "UK-like" model in Cai et al. (2023). The population bottleneck model is as follows: a population of ancestral size ten thousand diploids increases exponentially each generation at a rate of two percent starting three hundred generations ago, but twenty generations before the present day, the population experiences an instantaneous reduction in size to one million diploids. Otherwise, the demographic scenarios we explore here are populations of constant size twenty-five and one hundred thousand diploids.

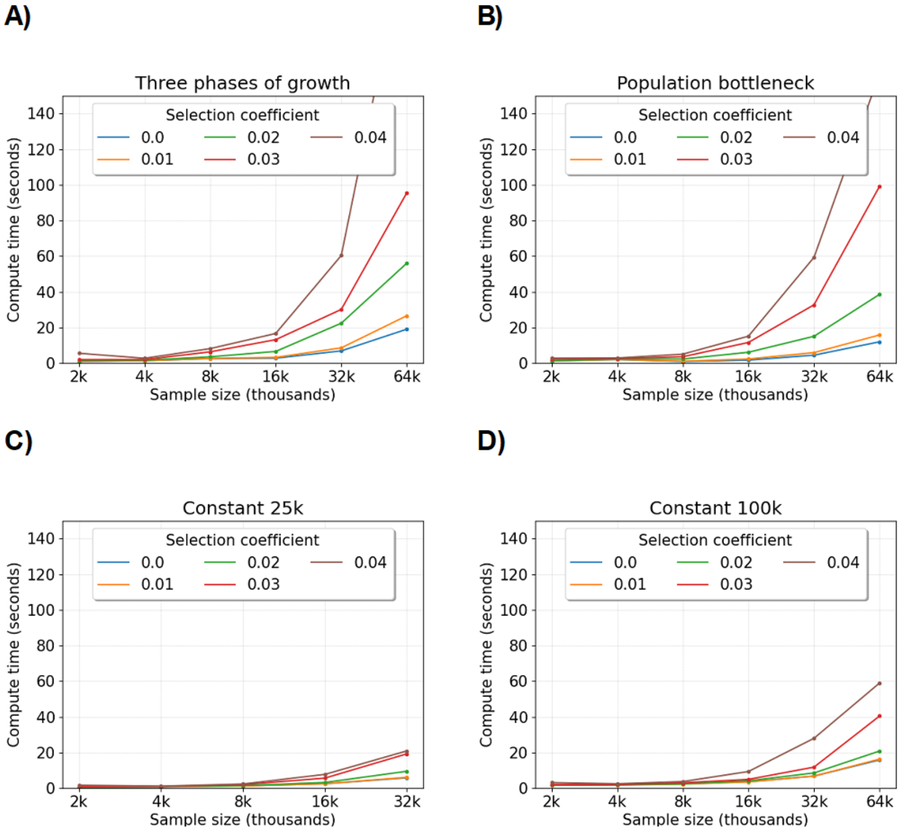

Fig. S6 Compute time to simulate IBD segment lengths around a locus depending on demography and selection. Compute time (y-axis) in seconds by sample size (x-axis) in thousands is averaged over five simulations. The legends denote colored line styles for different selection coefficients. A), B), C), and D) show results for demographic scenarios of three phases of exponential growth, a population bottleneck, and constant population sizes of twenty-five and one hundred thousand diploids, respectively. The Morgans length threshold is 0.01.

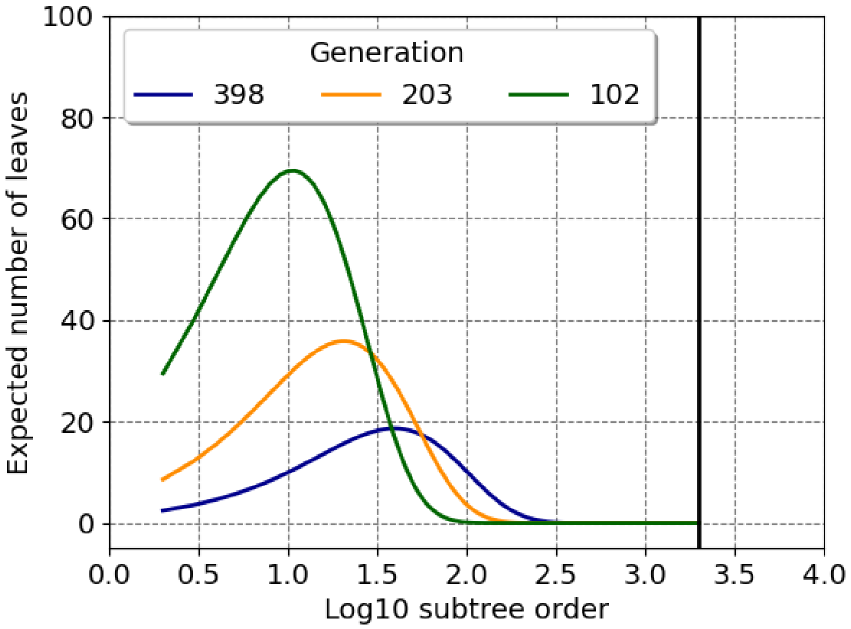

Fig. S7 The expected cardinality of subtree sizes at different coalescent times. Using Lemma 1 in Dahmer and Kersting (2015), we compute the expected number of subtrees containing r samples (x-axis) at the $(n-k)^{\mathrm{th}}$ coalescent event. We multiply these moments by r to get the expected number of leaves under such subtrees (y-axis). There are two thousand samples. Dark blue, orange, and green lines correspond to k =50, 95, and 180. We compute the expected time of the $(n-k)^{\mathrm{th}}$ coalescent event (Hein et al., 2005) and multiply by a population size of ten thousand to get generations (legend). The vertical line is logarithm 10 of sample size.