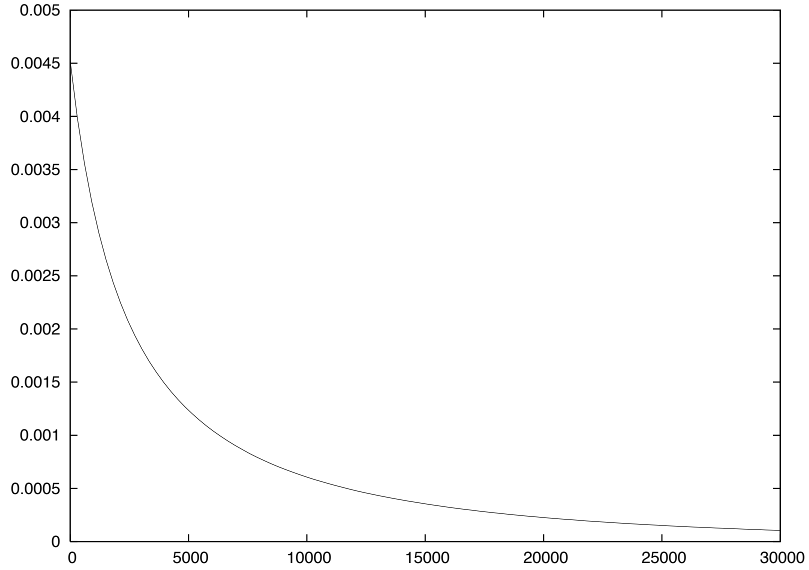

Figure 3: This plot shows the length spread of IBD segments that coalesce $0.2N$ generations ago. Comparing this to Figure 4, we see that it will be difficult to tell these segments apart from segments that coalesced $0.3N$ generations ago.

time are related so inexactly. We will see in Section 7, that precisely calculated IBS probabilities make it possible to use the distribution of shorter ROHs to estimate the effective population size at earlier points in history.

4 The probability of IBD given diploid IBS with uncertain haplotype phasing

In [31], Kong, et al. find IBS haplotypes by looking for diploid sequences L_1, L_2 with the property that $\mathrm{IBS}(L_1, L_2) \geq 1$ at every base in the sequence, i.e. that the alignment contains no locus for which L_1 and L_2 are homozygous for different alleles. However this condition does not guarantee that a haplotype of L_1 is

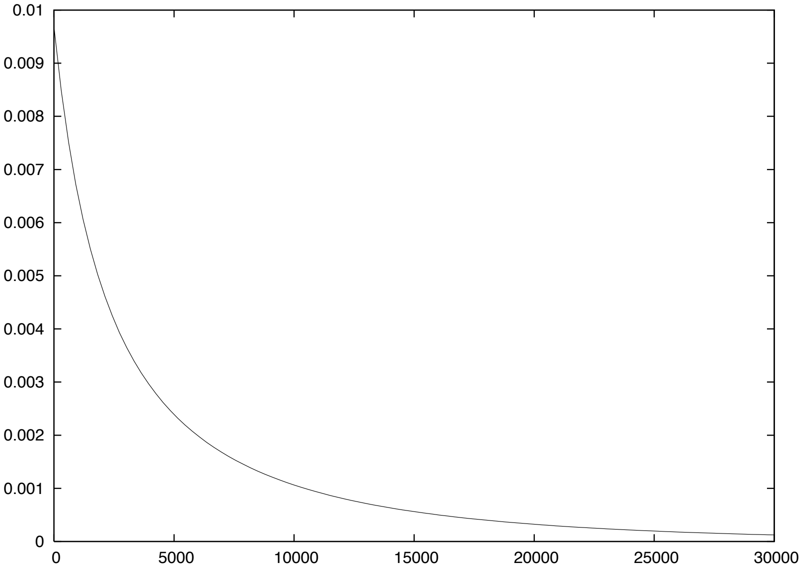

Figure 4: The length spread of IBD segments that coalesce $0.3N$ generations ago is different from the spread of segments that coalesce $0.2N$ generations ago (Figure 3), but overlaps enough that it would take a bit of work to learn about population history from a sum of density plots like these.

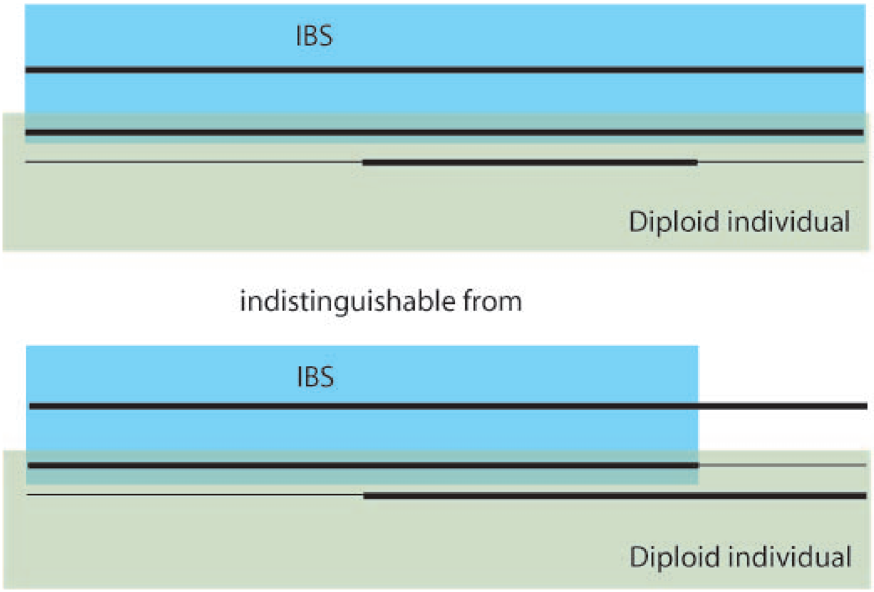

Figure 5: **IBS between phased haplotypes vs. IBS≥1** Here, the two chromosomes of a diploid individual are aligned to a reference haplotype. The diploid DNA is drawn in bold where it matches the reference haplotype IBS. Both the top and the bottom alignment have the property IBS$\geq$ 1, where at least one of the diploid sequences matches the reference at every base. However, only the diploid individual in the top alignment shares a haplotype IBS with the reference over the entire region.

IBS with a haplotype of L_2 as illustrated in Figure 5. The following question is motivated by this phasing issue, as well as the problem of reference panel imputation: Given an unphased diploid sequence d of length L aligned with a reference haplotype r, what is the probability that IBS$(r, d) \geq 1$ everywhere? A simpler problem is to find the probability that r and d are both IBD and IBS, IBS taken for the rest of this section to mean IBS$(r, d) \geq 1$, and again, both quantities are needed to find p(IBD|IBS)=p(IBD&IBS)/p(IBS).

In this section, it will be convenient to let $\mu = 2Nm$ and $\rho = 2Nr$ (which differs by a factor of two from in previous sections).

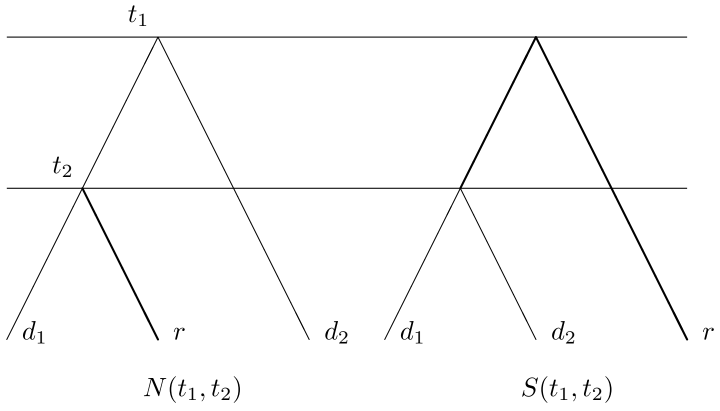

Figure 6: Natural and skew tree topologies. Boldface branches are drawn where mutations cannot occur without rendering the alignment non-IBS.

When describing the history of more than two sequences, it becomes necessary to discuss tree topology as well as coalescence time. In the case of comparing a haploid reference to the two chromosomes of a diploid sequence, we distinguish between *natural* and *skew* topologies: if L is sufficiently long and r is IBD with one of the haplotypes d_1, d_2, then it is unlikely, though not impossible, for the tree to have topology $(r(d_1d_2))$. We will refer to this as the *skew topology*, and to the other possible topologies as *natural topologies* (see Figure 6). Whatever the topology, we will let t_1 be the coalescence time of the root of the tree, and t_2 be the internal coalescence time. We refer to a particular skew tree as $S(t_1, t_2)$, and to both of the analogous natural trees as $N(t_1, t_2)$.

In order for the alignment to contain a pair of IBD haplotypes, there must be one coalescence time t that stays constant over the whole sequence. However, the coalescent history is free to vary from locus to locus over any trees of the form $N(t_1, t)$, $N(t, t_2)$, and $S(t, t_2)$. We will call transitions between such trees *allowed recombinations.*

In addition to these allowed recombinations, there is a set of allowed mutations that are compatible with the sequence containing a pair of IBS haplotypes. In the natural topology, mutations are allowed everywhere on the tree but on the branch joining r to its most recent common ancestor with one of d_1 and d_2, while in the skew topology, they are allowed *only* on the two branches joining d_1 and d_2 to their most recent common ancestor. Because $t_1 \ll t_2$ in general, long IBS alignments should statistically be dominated by the natural topology.

Given any coalescent tree on three leaves, the probability of t_2 being greater than t is e^{-3t}. Using this fact, we compute the probability $p_N(\text{IBD})$ that the alignment will coalesce IBD entirely in the natural topology:

$$\begin{aligned} P_N(\text{IBD\& IBS}) &= \int_{t_2=0}^{\infty}\int_{t_1=t_2}^{\infty} \frac{2}{3} \cdot e^{-t_2(\mu+3\rho)L} \cdot e^{-(t_1-t_2)} dt_1 \cdot 3e^{-3t_2} dt_2 \qquad (10) \\ &= \frac{2}{3+L(\mu+3\rho)}. \qquad (11) \end{aligned}$$

Here, $3e^{-3t_2}dt_2$ is the probability that the later coalescence will happen at exactly time t_2, while $2/3$ is the probability that it will be natural rather than skew. Given this event, $e^{-(t_1-t_2)}dt_1$ is the probability that the other coalescence happens exactly $t_1 - t_2$ time units earlier. $e^{-3L\rho t_2}$ is the probability that there will be no recombinations anywhere between t_2 and the present, at any locus on the alignment, and $e^{-L\mu t_2}$ is the probability that there will be no mutations on the thick branch joining s to the internal tree branch in Figure 6.

Similarly, we compute the probability $p_S(\text{IBD\& IBS})$ that the sequence coalesces IBD with the leftmost site in the skew topology:

$$\begin{aligned} P_S(\text{IBD\& IBS}) &= \int_{t_2=0}^{\infty}\int_{t_1=t_2}^{\infty} e^{-t_2(2-L(\rho+\mu))-t_1(1+2L(\rho+\mu))} dt_1 dt_2 \qquad (12) \\ &= \frac{1}{(1+2L(\rho+\mu))(3+L(\rho+\mu))} \qquad (13) \end{aligned}$$

In this last calculation, we neglect the fact that allowed recombinations can change the per-base mutation rate, decreasing the probability of no mutations from $e^{-t_1\mu}$ to as low as $e^{-2t_1\mu}$. However, these variations will affect $P_S(t_1, t_2)$ by at most a factor of 4. They do not change the fact that

$$\lim_{L\to\infty} \frac{P_S(\text{IBD\&IBS})}{P_N(\text{IBD\&IBS})} = 0.$$

As in Section 2, we calculate $P_L(\text{IBS})$ by induction, integrating $P_{L-1}(\text{IBS}, t_0)dt_0$ over a set of transition probabilities to find $P_L(\text{IBS}, t)$. We found the sequentially Markovian coalescent too complex to make this tractable, and it was necessary to make some simplifications, creating what we will call the forgetful SMC.

It is easiest to understand the difference between our forgetful SMC and the original SMC by analogy to the difference between the SMC and the full coalescent with recombination. As a point of reference, we reiterate that the SMC is a hidden Markov model where the hidden states are genealogies and the output of each genealogy is a locus in a sequence alignment [43]. The distribution of marginal genealogies at each site is the same as it would be under the full coalescent with recombination, but the transition probabilities between genealogies at neighboring sites are what differ between the two models. The genealogy distribution at base L, under the SMC, is completely determined by the distribution of genealogies at base $L-1$, while under the full coalescent it also depends on the distribution of genealogies at all previous bases in the sequence.

The distribution we wish to compute is not a full sampling distribution of sequence alignments, but simply the percentage of these alignments that are IBS$\geq$ 1. For our purposes, there are output two output states of the SMC is binary: each locus is IBS or non-IBS. The output distribution of a skew-topology genealogy depends only on the recent coalescence time, t_2, not on the older coalescence time t_1; t_2 affects the transition probabilities, but not the marginal outputs of the Markov chain. Motivated by this fact, we modify the SMC so that t_2 is forgotten after each transition event and the resampled before the next one. The precise construction is given in the following paragraph and illustrated in Figure 7 as an HMM flow diagram.

Instead of keeping track of a three-leaf coalescent tree at each site, we will only keep track of the time t at which the reference r coalesces with one of the haplotypes d_1, d_2. This is t_2 in the natural topology and t_1 in the skew topology. When we calculate the transition probability from t_0 to t, we will assume that the t_0 tree is in the natural topology and pick t_1 from its expected distribution, conditional on t. After a recombination, however, we allow the new tree to coalesce in either the natural or the skew topology. The small number of skew trees that are produced will be regarded as natural at the next recombination event. Our simulations suggest that this gives more accurate results than outlawing skew coalescences entirely, while keeping the computational complexity under control. The results agree closely with a $P(\text{IBD}|\text{IBS})$ curve that we constructed using MS simulations, conditioning on the full coalescent [25].

The following recursion summarizes the transition probabilities of the forgetful SMC. An explanation of each term will follow:

$$\begin{aligned}
P_L(\text{IBS}, t)dt &= p_{L-1}(\text{IBS}, t)dt \cdot e^{-t(\mu+2\rho)} \\
&+ \int_{t_0=0}^{t} \int_{t_r=0}^{t_0} p_{L-1}(\text{IBS}, t_0) \cdot e^{-\mu t} \Bigg(2\rho e^{-2\rho t_r} \cdot \frac{2}{3} \cdot 3e^{-3(t-t_r)} dt \\
&\quad +2\rho e^{-2\rho t_r} \int_{t_2=t_r}^{t} \frac{1}{3} \cdot 3e^{-3(t_2-t_r)} \cdot e^{-(t-t_2)} dt dt_2 \Bigg) dt_r dt_0 \\
&+ \int_{t_0=t}^{\infty} \int_{t_r=0}^{t} p_{L-1}(\text{IBS}, t_0) \cdot e^{-\mu t} \Bigg(\frac{2}{3} \cdot 3\rho e^{-3\rho t_r} \cdot \frac{1}{2} \cdot 2e^{-2(t-t_r)} + \\
&\quad + \frac{1}{3} \cdot 3\rho e^{-3\rho t_r} \cdot 2e^{-2(t-t_r)} dt \Bigg) dt_r dt_0 \\
&= p_{L-1}(\text{IBS}, t)dt \cdot e^{-t(\mu+2\rho)} \\
&+ \int_{t_0=0}^{t} \int_{t_r=0}^{t_0} p_{L-1}(\text{IBS}, t_0) \Big(3\rho e^{-t(3+\mu)+t_r(3-2\rho)} \\
&\quad + \rho e^{-t(1+\mu)+t_r(1-2\rho)} \Big) dt_r dt_0 dt \\
&+ \int_{t_0=t}^{\infty} \int_{t_r=0}^{t} p_{L-1}(\text{IBS}, t_0) \cdot 4\rho e^{-t(2+\mu)+t_r(3-2\rho)} dt_r dt_0 dt.
\end{aligned}$$

Since we are assuming that the initial $(L-1)$ bases of IBS end with a natural topology tree, we can let d_1 denote the haplotype that coalesces with r before the other haplotype does. The first integrand is the probability that an $(L-1)$-base alignment coalescences IBS, its rightmost site coalescing at time $t_0 < t$, along with one of the following events:

1. One of r and d_1 recombines at time t_r (probability $2\rho e^{-2\rho t_r} dt_r$). The first coalescence among r, d_1, d_2 occurs in the natural topology (probability $\frac{2}{3}$) at time t (probability $3e^{-3(t-t_r)}dt$).

2. One of r and d_1 recombines at time t_r (probability $2\rho e^{-2\rho t_r} dt_r$). The first

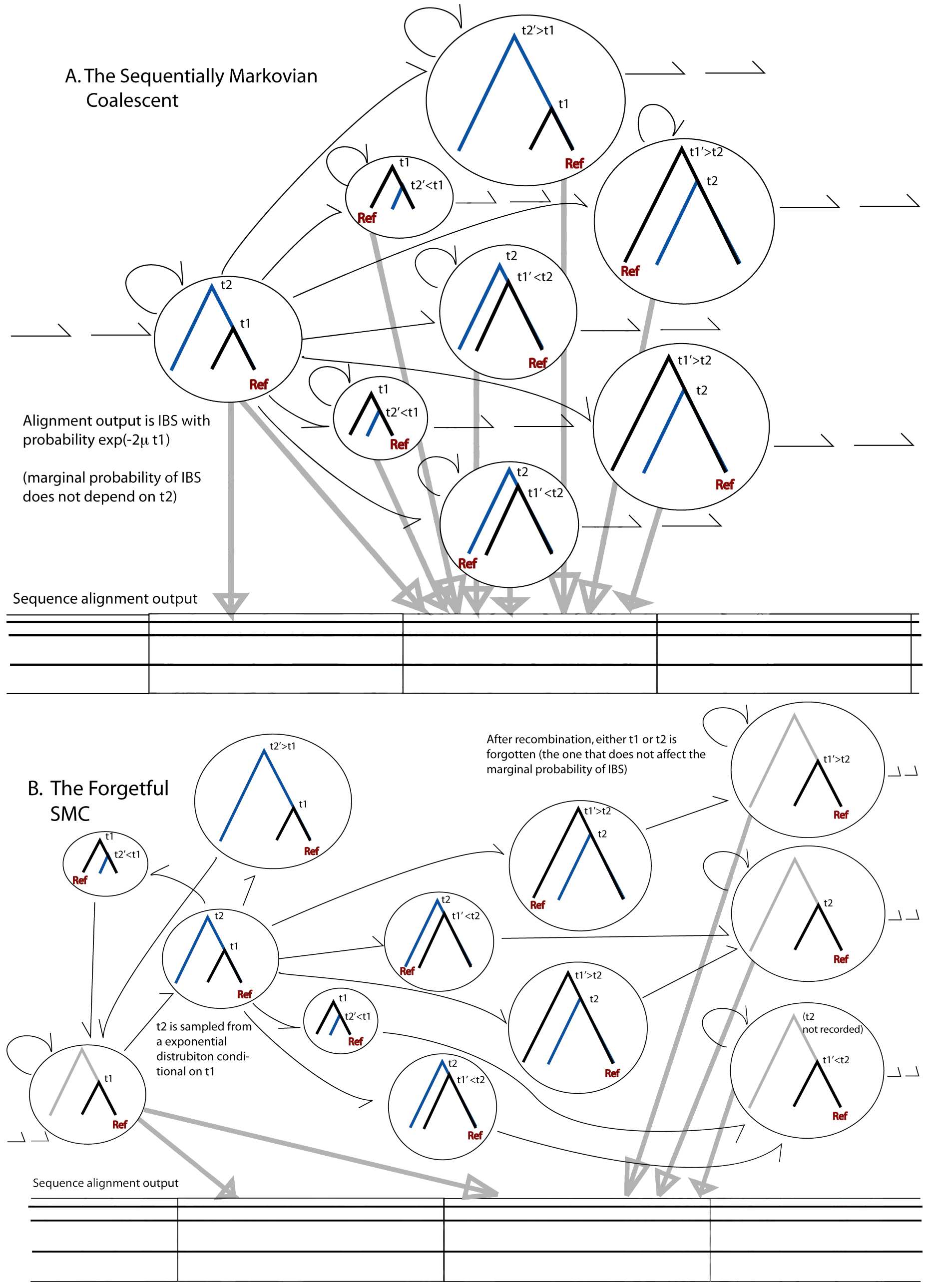
A. The Sequentially Markovian Coalescent
t2'>t1
t1
Ref
t1
t2'<t1
Ref
t1'>t2
t2
Ref
t2
t1'<t2
Ref
t2
t1
Ref
t1'>t2
t2
Ref
t1
t2'<t1
Ref
t2
t1'<t2
Ref
Alignment output is IBS with probability exp(-2μ t1)
(marginal probability of IBS does not depend on t2)
Sequence alignment output
B. The Forgetful SMC
t2'>t1
t1
Ref
t1
t2'<t1
Ref
After recombination, either t1 or t2 is forgotten (the one that does not affect the marginal probability of IBS)
t1'>t2
Ref
t1'>t2
t2
Ref
t2
Ref
t2
t1'<t2
Ref
t2
t1
Ref
t1'>t2
t2
Ref
t1
t2'<t1
Ref
(t2 not recorded)
t1'<t2
Ref
t1
Ref
t2 is sampled from a exponential distrubiton conditional on t1
t2
t1'<t2
Ref
Sequence alignment output

Figure 7: Hidden Markov model flow diagrams of the SMC and our forgetful approximation of the SMC. The position of the reference sequence is labeled to mark each genealogy as natural or skew. Each output extends the alignment by a triplet of bases (including one labeled reference base) that is either IBS ≥ 1 or not.

coalescence among r, d_1, d_2 occurs in the skew topology (probability $\frac{1}{3}$) at a time $t_2 < t$ (probability $3e^{-3(t_2-t_r)}dt_2$). The common ancestor of d_1 and d_2 coalesces with r at time t (probability $e^{-(t-t_2)}dt$).

The second integrand is similarly defined, with $t_0 > t$ and two possible coalescent scenarios. It is not possible for recombination to turn a natural-topology tree indexed by time t_0 into a skew-topology tree indexed by a time $t < t_0$.

1. The first recombination among r, d_1, d_2 occurs at time t_r (probability $3\rho e^{-3\rho t_r}dt_r$. It happens to d_1 or d_2 (probability $\frac{2}{3}$), and this sequence coalesces with r at time t (probability $e^{-2(t-t_r)}dt$). (It is certain that the second coalescence happens less recently than time t).

2. The first recombination among r, d_1, d_2 occurs at time t_r (probability $3\rho e^{-3\rho t_r}dt_r$. It happens to r (probability $\frac{1}{3}$). Sequence r coalesces with d_1 or d_2 at time t (probability $2e^{-2(t-t_r)}dt$).

As an aside, we will discuss the central difference between our model and the Sequentially Markovian Coalescent [43], the model that enabled our computation of two-haplotype IBS probabilities in Section 2. The SMC has the property that the history at position x depends only on the history at position $x-1$, but the history of three or more sequences is a hefty variable consisting of a topology and two interrelated coalescence times, and the SMC is not Markovian in either of those times on its own.

To illustrate, suppose that the alignment contains the topology sequence $((r, d_1), d_2), ((r, d_2), d_1), ((r, d_1), d_2)$. If (r, d_2) restricted to the middle section coalesces more recently than (r, d_1) in either outside section, then it is possible that r is IBD with d_2 throughout the composite alignment, a possiblity that our model does not capture. However, since $t_1 \gg t_2$ in general, it is unlikely for (r, d_2) to stay IBD over an interval where (r, d_1) are not IBD. Disallowing this fringe possibility makes our process Markovian in a single time variable, one that is much simpler to integrate over than a three-parameter history.

There exists a series of number pairs $\{(B, C_B)\}$ for which

$$P_L(\mathrm{IBS}, t) = \sum_B C_B(L) e^{-tB}; \tag{14}$$

performing the necessary integrals, we find that

$$\begin{aligned} P_{L+1}(\mathrm{IBS}, t) \quad = \quad & \sum_B C_B(L) \left(e^{-t(B+\mu+2\rho)} + \frac{3\rho}{B(B-3+2\rho)} \left(e^{-t(3+\mu)} - e^{-t(B+\mu+2\rho)} \right) \right. \\ & + \frac{\rho}{B(B-1+2\rho)} \left(e^{-t(1+\mu)} - e^{-t(B+\mu+2\rho)} \right) \\ & + \frac{3\rho}{B(3-2\rho)} \left(e^{-t(B+3+\mu)} - e^{-t(B+\mu+2\rho)} \right) \\ & + \frac{\rho}{B(1-2\rho)} \left(e^{-t(B+1+\mu)} - e^{-t(B+\mu+2\rho)} \right) \\ & \left. + \frac{4\rho}{B(2-3\rho)} \left(e^{-t(B+\mu+3\rho)} - e^{-t(B+2+\mu)} \right) \right). \end{aligned}$$

We can compute $P_L(\mathrm{IBS}, t)$ much more quickly, losing very little accuracy, by truncating the formula to

$$\begin{aligned} P_{L+1}(\mathrm{IBS}, t) \quad = \quad & \sum_B C_B(L) \left(e^{-t(B+\mu+2\rho)} + \frac{3\rho}{B(B-3+2\rho)} \left(e^{-t(3+\mu)} - e^{-t(B+\mu+2\rho)} \right) \right. \\ & \left. + \frac{\rho}{B(B-1+2\rho)} \left(e^{-t(1+\mu)} - e^{-t(B+\mu+2\rho)} \right) \right). \end{aligned}$$

In this way, we write

$$P_L(\mathrm{IBS}, t) = \sum_{i=0}^{L-1} C_i(L) e^{-t(1+\mu+i(\mu+2\rho))} + D_i(L) e^{-t(3+\mu+i(\mu+2\rho))}, \tag{15}$$

the coefficients satisfying the recursions

$$
\begin{aligned}
C_{i+1}(L+1) &= \left(1 - \frac{3\rho}{(1+\mu+i(2\rho+\mu))(\mu-2+2\rho+i(2\rho+\mu))}\right.\\
&\quad \left.-\frac{\rho}{(1+\mu+i(2\rho+\mu))(\mu+2\rho+i(2\rho+\mu))}\right) C_i(L)\\
D_{i+1}(L+1) &= \left(1 - \frac{3\rho}{(3+\mu+i(2\rho+\mu))(\mu+2\rho+i(2\rho+\mu))}\right.\\
&\quad \left.-\frac{\rho}{(3+\mu+i(2\rho+\mu))(2+\mu+2\rho+i(2\rho+\mu))}\right) D_i(L)\\
C_0(L+1) &= \sum_{i=0}^{L-1} \frac{3\rho}{(1+\mu+i(2\rho+\mu))(\mu-2+2\rho+i(2\rho+\mu))} C_i(L)\\
&\quad +\sum_{i=0}^{L-1} \frac{3\rho}{(3+\mu+i(2\rho+\mu))(\mu+2\rho+i(2\rho+\mu))} D_i(L)\\
D_0(L+1) &= \sum_{i=0}^{L-1} \frac{\rho}{(1+\mu+i(2\rho+\mu))(\mu+2\rho+i(2\rho+\mu))} C_i(L)\\
&\quad +\sum_{i=0}^{L-1} \frac{\rho}{(3+\mu+i(2\rho+\mu))(2+\mu+2\rho+i(2\rho+\mu))} D_i(L)
\end{aligned}
$$

with base case

$$
\begin{aligned}
P_1(\mathrm{IBS}, t) &= 2e^{-t(3+\mu)} + \int_{t_2=0}^{t} e^{-3t_2} \cdot e^{-(t-t_2)} \cdot e^{-t\mu} dt_2\\
&= \frac{1}{2} e^{-t(1+\mu)} + \frac{3}{2} e^{-t(3+\mu)}.
\end{aligned}
$$

For future reference, we will summarize this set of recursions in an operator $\mathcal{D}$ defined such that

$$
\mathcal{D}^{L-1}\left(P_1(\mathrm{IBS}|t)\right) = \mathcal{D}(P_{L-1}(\mathrm{IBS}|t)) = P_L(\mathrm{IBS}|t). \tag{16}
$$

As mentioned before, we performed MS coalescent simulations to check the results of the diploid computations [25], finding empirical probabilities of IBD and IBS based on 10^6 trial histories. Our formula underestimates $P_L(\mathrm{IBD}|\mathrm{IBS})$ for short sequences, predicting that $P_{10000}(\mathrm{IBD}|\mathrm{IBS}) = 0.676$ while the simulations say it should be 0.745. However, the discrepancy narrows quickly as L increases, with $P_{50000}(\mathrm{IBD}|\mathrm{IBS}) = 0.838$ and simulations showing it to be 0.848. Our underestimation of $P_L(\mathrm{IBD}|\mathrm{IBS})$ disappears less quickly when we simulate

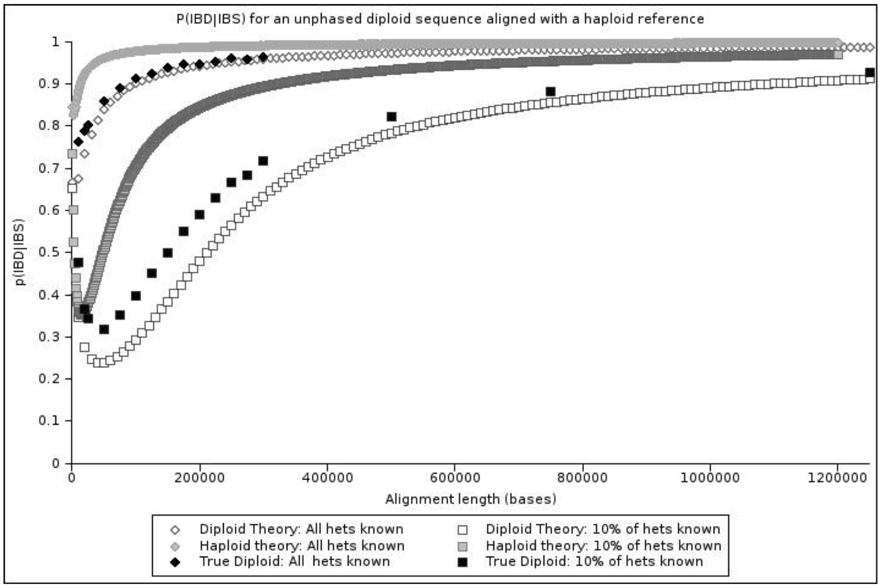

Figure 8: This plot compares our diploid results to the values that we obtained from MS simulations, which assume the full non-Markovian coalescent with recombination. Given a 10^5-base diploid sequence that is IBS ≥ 1 with a refernce, it is 90.1% likely to contain a haplotype that is IBD with the reference. This probability increases to 98.5% for an alignment 10^6 bases long. When we observe only 10% of all hets, the corresponding probabilities are 29.4% ($L = 10^5$) and 89.1% ($L = 10^6$).

the effect of thinned marker data, observing only 10% of all hets, but we still get within 1% of the true value for $L \geq 500,000$ (see Figure 8).

5 Using identity by state to phase and impute haplotypes

There are a number of questions in applied genetics research that require accurate identification of tracts of IBD. The oldest of these questions center around