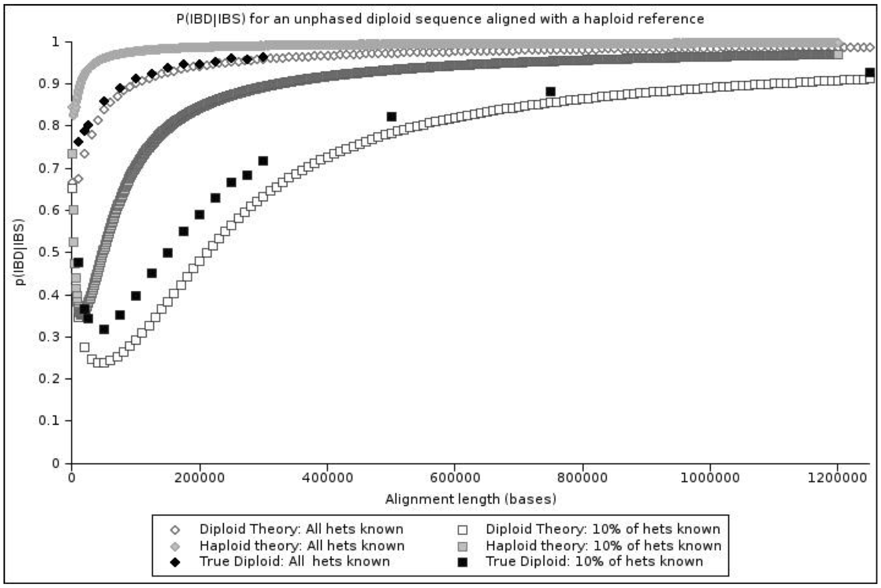

Figure 8: This plot compares our diploid results to the values that we obtained from MS simulations, which assume the full non-Markovian coalescent with recombination. Given a 10^5-base diploid sequence that is IBS ≥ 1 with a refernce, it is 90.1% likely to contain a haplotype that is IBD with the reference. 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 pedigree analysis, but we argue that our results are most applicable to the problem of phasing and imputation in unrelated individuals. Imputation accuracy is tied to the probability of diploid IBS as follows: if we have a thinned IBS alignment between a stretch of unphased genotypes and a haploid reference sequence and we compute that this alignment has a probability $P_L(\text{IBD}|\text{IBS})_{0.10}$ of containing an pair of haplotypes that are IBD and IBS, then with probability $P_L(\text{IBD}|\text{IBS})_{0.10}$, the unphased genome contains a perfect copy of the reference haplotype.

We can see in Figure 8 that the probability $P_L(\text{IBD}|\text{IBS})_{0.10}$ converges slowly to 1 as L gets very large. It reaches the value $P_L(\text{IBD}|\text{IBS})_{0.10} = 0.9$ when $L \approx 10^7$, and unfortunately it is rare to find such long IBS alignments between DNA from unrelated individuals. In shorter IBS alignments, however, we can be more certain of IBD near the alignment center than at its edges–even if the unphased genome is unlikely to contain a perfect copy of the entire reference haplotype, it is likely to contain a perfect copy of a subsequence of that haplotype. Given an $(L+2x)$-base thinned IBS aligment between a haploid reference and a diploid test sequence, we can compute the probability $I(L,x)_{0.10}$ that the middle L bases of the reference will be IBD with of the test haplotypes. If we are trying to impute a genotyped individual using a reference haplotype panel, then $I(L,x)_{0.10}$ can help us figure out how much sequence we can copy while keeping the expected number of errors per kilobase of imputed sequence below a specified threshold.

Before computing $I(L,x)_{0.10}$, we will address its relationship to the accuracy of the current state-of-the-art in imputation. While $I(L,x)_{0.10}$ predicts the accuracy of imputing the exact sequence of a genotyped individual, it is less common than to impute from full sequence data than from the densely genotyped panel of HapMap references. To accurately copy the states of HapMap SNPs from a reference haplotype to a test individual, it is perhaps overly conservative to ask for a high probability that the test individual contain a perfect copy of the reference; the program IMPUTE v2, for example, is consistently accurate at imputing sites with minor allele frequency $\geq 10\%$, but its accuracy at imputing rarer variants falls off at a rate that depends on the genotype chip and HapMap references being used [23, 39]. However, the authors of IMPUTE predict, in a review on imputation methods, that the 1000 Genomes Project will replace HapMap as the imputation reference of choice, and that one of the challenges associated with the switchover will be the fact that the 1000 Genomes references will contain more variants with frequencies in the 1%-5% range [39]. Using the 1000 Genomes data for imputation will confer both added power and added error, compared to using HapMap, and a way to estimate the extent of that added error would be to predict accuracy in terms of IBD, as we do here.

5.1 The probability of IBD in the central subset of an IBS alignment

In the last section, we derived an integration operator $\mathcal{D}_{0.10}$ for which

$$P_L(\mathrm{IBS}|t)_{0.10} = \mathcal{D}_{0.10}(P_{L-1}(\mathrm{IBS}|t)_{0.10}) = \mathcal{D}_{0.10}^{L-1}\left(P_1(\mathrm{IBS}|t)_{0.10}\right), \qquad (17)$$

making it possible to compute $p(\mathrm{IBD}|\mathrm{IBS})_{0.10}$ for unphased diploid alignments. It follows from the definition of $\mathcal{D}_{0.10}$ that

$$I(L,x)_{0.10} = \frac{1}{P_{L+2x}(\mathrm{IBS})_{0.10}} \int_{t=0}^{\infty} \mathcal{D}_{0.10}^{x}(e^{-tL(2\mu+2\rho)} \cdot P_x(\mathrm{IBS}|t)_{0.10})dt, \qquad (18)$$

where $e^{-tL(2\mu+2\rho)} = p_L(\mathrm{IBS\&IBD}|t)/e^{-t}$ is the probability that a base pair coalescing at time t is at the center of an L-base stretch that is IBS and IBD. Put another way, it is the Lth power of an operator for extending the test alignment by one IBD base, while $\mathcal{D}_{0.10}$ is an operator for extending the test alignment by one thinned IBS base.

Figure 9 plots $I(L,x)_{0.10}$ for $x = 10^4, 5\times10^4$, and 10^5, showing that removing the terminal 10^5 bases from each end of a thinned IBS alignment produces substantial gains in the likelihood of IBD.

Since $I(L,x)_{0.10}$ is the expected accuracy of imputing L bases from an $(L+2x)$-base alignment, it is possible to conduct imputation such that the L-base sequence calls should be e.g. 95% accurate. We need only find x for which

$I(L,x)_{0.10} > 0.95$ and not impute from any shorter IBS alignments. When such thresholds are set, however, a question of coverage arises: given n references, a large number of test sequences, and a minimum required accuracy $p < 1$, into how many sequences can we expect to impute a given L bases from the reference panel? As usual, the question is whether a test haplotype coalesces very early with one of the references, making it necessary to consider $(n+2)$-leaf coalescent trees.

The first coalescence between a test haplotype and a reference will be one of the $n+1$ coalescences that make up the nodes of an $(n+2)$-leaf tree; we must find formulas for when these events occur and also the likelihood that the kth of $n+1$ coalescences will be the particular event we are interested in.

It is proved in [18] that the following is a formula for the probability that n samples have exactly k ancestors at $t/(2N)$ generations before the present:

$$\begin{aligned} h_{n,k}(t) &= \sum_{i=k}^{n} e^{-\binom{i}{2}t} \frac{(2i-1)(-1)^{i-k}(k+i-2)!n!/(n-i)!}{k!(i-k)!(n+i-1)!/(n-1)!} \\ &= \sum_{i=k}^{n} e^{-\binom{i}{2}t} \frac{(2i-1)(-1)^{i-k}(k+i-2)!n!(n-1)!}{k!(i-k)!(n+i-1)!(n-i)!} \end{aligned}$$

Letting $P(T_k < t)$ be the probability that the kth of n coalescences happens before time t, it is easy to see that

$$h_{n,k}(t) = P(T_{n-k-1} < t)(1 - P(T_{n-k} - 1)),$$

and it is also true that

$$\lim_{n,k\to\infty} P(T_{n-k} = t) = \frac{h_{n,k}(t)dt}{\int_{t=0}^{\infty} h_{n,k}(t)dt}.$$

It is easy to see, combinatorially, that if the two test haplotypes haven't coalesced with each other yet, the coalescence from $k+1$ to k sequences will involve an ancestor of a test haplotype with probability

$$\frac{2k}{\binom{k+1}{2}} = \frac{4}{k+1}.$$

If the test haplotypes have coalesced with each other, the probability will instead be $2/(k+1)$; however, this is the fringe skew topology case. If we want the $(n+1-k)$th coalescence to involve an ancestor of a test haplotype with probability p, it must be true that

$$\left(1-\frac{4}{n+2}\right)\cdots\left(1-\frac{4}{k+1}\right) < 1-p,$$

meaning that

$$\frac{k(k-1)}{(n+2)(n+1)} < 1-p$$

and

$$k \approx n\sqrt{1-p}.$$

Therefore, the probability that the $(n-k)$th of n coalescences (with the first being closest to the present) is the earliest one to involve a test haplotype is

$$1-k^2/n^2-(1-(k+1)^2/n^2)=\frac{2k+1}{n^2}; \tag{19}$$

if $P_n(t)dt$ is the probability that t is the smallest time at which a test haplotype coalesces with a reference, then

$$P_n(t)=\sum_{k=1}^{n+1}\frac{2k+1}{n^2}P(T_k=t). \tag{20}$$

When n is large, it will be helpful to avoid summing over all possible values of k. Instead, we select a series of k values that correspond to fixed percentiles; i.e., k for which it is 90% likely that a reference coalesces with a test haplotype at or before the $(n-k)$th coalescence. We sum over k values corresponding to the 10th, 20th,..., 90th percentiles (indexed by m in the following sum), along with the 95th, 99th, and 99.9th percentiles:

$$\begin{aligned}P_n(t) < 0.1\sum_{m=1}^{9}P(T_{n-n\lfloor\sqrt{1-0.1m}\rfloor}=t)+0.05P(T_{n-n\lfloor\sqrt{0.05}\rfloor}=t)\\ +0.04P(T_{n-n\lfloor\sqrt{0.04}\rfloor}=t)+0.009P(T_{n-n\lfloor\sqrt{0.009}\rfloor}=t) := Q_n(t).\end{aligned}$$

The function $Q_n(t)$ has the property that

$$\int_{t=0}^{\infty}Q_n(t)dt=1; \tag{21}$$

it is approximately the distribution of coalescence times at the left endpoint of the longest thinned IBS alignment between a test haplotype and reference, not stipulating that the alignment be at least L bases long. In contrast, $\mathcal{D}_{0.10}^{L-1}(Q_n(t)e^{-0.10\mu t})dt$ is the probability that this endpoint will coalesce at time t and that, in addition, thinned IBS extends for at least L bases. We will take

$$\overline{Q}_n(L) = \int_{t=0}^{\infty} \mathcal{D}_{0.10}^{L-1}(Q_n(t)e^{-0.10\mu t})dt \qquad (22)$$

as our approximation for the probability that a test haplotype will be part of a thinned IBS alignment of length L with one of the references.

Figure 10 plots the probability that, given a panel of n references and a 1 kilobase region of a test sequence to be imputed, the region will be at the center of a $(2x + 1000)$-base thinned IBS alignment between the test sequence and one of the references. Figure 11 plots the accuracy distribution of the imputation calls made in this way. The function $I(1000, x)_{0,10}$ gives the accuracy of a call made from a $(2x+1000)$-base IBS alignment, while the probability of observing a $(2x + 1000)$-base IBS alignment from which to impute is $\overline{Q}_n(2x + 1000)$.

Since a constant effective population size of 10,000 is being assumed, the appearance of perfect power for a 1000-haplotype panel is overly optimistic. In outbred populations, exponential growth is likely to have broken up very long haplotypes, as reported by Hayes, et al [17]. Taken as a set of upper bounds, however, these plots show that a HapMap of 120 sequences gives far from perfect haplotype coverage, even in a moderately isolated population.

6 The effect of underestimating the linkage between markers when computing IBS probabilities

Although the aim of inferring IBD is to be confident of IBS at a dense set of markers, previous methods for inferring IBD tend to lose accuracy if the input set of markers is too dense, a fact that limits their precision. The problem with