The relationship of identity by state to identity by descent and imputation accuracy in population sequencing data

Kelley Harris
Emmanuel College
University of Cambridge

Dissertation submitted for the degree of Master of Philosophy

March 29, 2011

PREFACE

This dissertation is my own work. The research was not done in collaboration except where marked explicitly in the text.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Richard Durbin, for suggesting these research questions and providing so much guidance through the process of answering them and writing up this report. He and Heng Li, who created the genome call sets used in Section 7, will be my coauthors on a manuscript based on the work described in this dissertation. Besides thanking Heng for providing the data I needed to complete the project, I would like to thank everyone in the Durbin group for welcoming me and teaching me so much over the past year. In particular, Kimmo Palin and Jared Simpson helped me learn to use the Sanger computer system, and Kimmo, as well as Aylwyn Scally, helped me edit previous versions of this writeup. In writing and researching, I also benefitted from the helpful input of my thesis committee: Carl Anderson, Vincent Plagnol, Chris Tyler-Smith, and Eleftheria Zeggini, and would like to thank Gilean McVean and Jeffrey Barrett for taking on the finished product to read. I am grateful for the financial support I received from the Harvard College Herchel Smith Postgraduate Fellowship program, and to the Wellcome Trust Sanger Institute for hosting me. Finally, I'd like to thank my friends for always supporting me and my parents, Glenn Harris and Anne Katten, for doing that and so much more.

Abstract

When two DNA sequences look *identical by state* (IBS) along a string of genotyped markers, the DNA between the markers is often *identical by descent* (IBD), meaning inherited from a recent common ancestor without recombination. This fact makes it possible to scan the whole genome for functional variants without typing every base directly, making use of information about unobserved bases that is provided by the states of observed bases. We attempt to quantify that information here, using coalescent theory to predict how strongly various degrees of IBS imply IBD, taking into account the density of genotyped markers and past effective population size.

In addition to calculating the probability of IBD between IBS haploid sequences, we consider the problem of matching an unphased diploid sequence to a reference haplotype panel. The results have bearing on the practices of haplotype phasing and genotype imputation, both of which become more reliable when the ends of an IBS alignment are not assumed to be IBD. To compute p(IBD|IBS) when phasing ambiguity is an issue, it was necessary to develop a new approximation to the neutral coalescent with recombination: a further simplification of the sequentially Markovian coalescent [43].

Computing p(IBD|IBS) by our method is closely related to predicting the length distribution of homozygous stretches in the genome. After accounting for sequencing errors, we predict this distribution correctly, as judged by data from eleven complete human genomes. We also predict the length distribution of segments that appear homozygous based on thinned marker data, noting that the probability of sequence IBS given "thinned IBS" is a natural measure of imputation accuracy.

The probability of IBS given thinned IBS varies with sequence length in a way that is very ethnically distinctive, as judged by data from five Africans, four Europeans, and two Asians. We are able to account for these differences in terms of past changes in effective population size: an out-of Africa bottleneck followed by a shallower, more recent Asian bottleneck. We predict that IBS implies IBD most strongly in historically

outbred populations, and that extra care should be taken when inferring IBD in bottlenecked populations.

Returning to the problem of imputation, we estimate the accuracy spread of the imputation calls that can be made from a panel of n reference haplotypes. For an idealized population of effective size $N = 10,000$, we find that the 120-reference HapMap should omit a significant amount of genetic variation; that given a 1-kilobase stretch of a genotyped individual's DNA, a 120-reference panel gives us only a 70% chance of imputing the sequence of that stretch with 99% accuracy. However, we find that a 1000-haplotype panel should enable near-perfect imputation in a population that has been isolated from recent exponential population growth, and such perfect imputation would allow for precise genetic mapping in groups much larger than extended families.

Contents

1 Introduction

Every child is born with a few *de novo* mutations, DNA sites where they differ from their parents and from most other humans. Most of the variants created this way die out within a few generations, but a minority of them spread to hundreds or thousands of the child's descendants and contribute to widespread human genetic variation [1, 27]. By mathematically modeling the emergence and spread of new alleles, population geneticists can make inferences about ancient periods of growth, decline, interbreeding, and the emergence of modern ethnic groups, as well as discover links between genetic and phenotypic variation.

Given DNA from one individual, it is much cheaper to genotype a few thousand genetic loci than to ascertain the entire genome sequence, so companies like Illumina and Affymetrix manufacture single nucleotide polymorphism (SNP) chips that can selectively ascertain the states of between 10,000 and 1,000,000 of the most variable sites in humans. By focusing on fewer genetic sites, one can afford to genotype those sites in more individuals, and this approach has been used since the invention of pedigree analysis to find many sites in the genome that correlate with disease risk or recent positive selection [36]. A problem with SNP chips, however, is that they omit sites where variant alleles arose too recently to spread to a significant fraction of the human population. Although none of the three billion sites that are omitted from a SNP chip is especially variable on its own, together they harbor a vast amount of additional genetic information [1, 26]. This hard-to-detect variation is a clear candidate harbor for "missing heritability" in disease genetics, where known genetic risk factors usually fail to account for the full heritability of complex diseases [29, 51].

One way to detect more low-frequency variants will be to gather more genotype and sequence data, working to make this process cheaper through improvements in biotechnology. Another approach, however, is to extract more information from available data sets by modeling a process known as *linkage disequilibrium* (LD). Even when site X does not appear on a chip being used to to gather data, it can still be possible to infer that two sequences match at site

X by looking for matching at sites close to X. DNA is passed from parents to children in continuous blocks between recombination sites; when two sequences share a rare allele, it is likely that the allele was inherited from a recent common ancestor along with a block of surrounding DNA containing some sites that appear on the SNP chip [10, 31].

Linkage disequilibrium affects the distribution of heterozygous sites (*hets*) in every diploid genome, even in outbred populations. If every site in the genome had an independent probability m of being a het, then the probability of an L-base region being devoid of hets, or *identical by state* (IBS) would be $(1-m)^L \approx e^{-L}$. The frequency of L-base regions of homozygosity (ROHs) is not observed to decline exponentially with L, however [41, 55], and the excess of long ROHs can be accounted for by modeling LD. If, for example, an individual's parents are ninth-degree cousins, there is only a one-in-2^{20} chance that both alleles at a given site in the child's DNA were inherited from the parents' most recent common ancestor, but given that both alleles *were* both inherited from that ancestor, the child is likely to be homozygous over 10 megabases of surrounding DNA [31]. Ten generations is not enough time for meiosis to break the DNA into smaller heritable pieces, and in general, the length of a homozygous stretch is inversely proportional to the age of the ancestor that the matching haplotypes derive from.

The key to understanding how hets are placed is understanding how coalescence time, or time to common ancestry, varies from site to site across the genome. We define *ancestral recombination sites* (ARs) to be loci where two neighboring allele pairs coalesce at different times, and say that an alignment is *identical by descent* (IBD) if it has no interior ARs (See Figure 1 for an illustration of IBD vs. IBS).

A consequence of coalescent theory is that hets are placed randomly within an IBD region, with their density proportional to the region's coalescence time t ($t = 0$ being the present and larger t's being more ancient). As we move from left to right across a region of IBD, each base has a constant probability of being a het and a constant probability of being an AR and ending the IBD stretch.

In human DNA, the het probability μt is about 2.5 times the AR probability ρt, such that each IBD region contains about 2.5 total hets. The length of the region will vary inversely with t, however, making the local density of hets very small when t is very small.

Commercial chips with at most 1,000,000 SNP sites can detect at most 10% of the hets in a diploid sequence. This suggests that, on average, an IBD region will contain 0.25 hets that are detectible with a 1,000,000 chip and that $e^{-0.25} > 0.77$ of all maximal IBD regions will appear IBS based on genotype data. In contrast, only $e^{-2.5} \approx 0.082$ of maximal IBD regions will appear IBS based on sequence data.

Although definitions of IBD differ widely in the literature, IBD between two sequences is usually taken to imply IBS at the level of genotype data, and we do not intend to create confusion by defining IBD such that it does not imply IBS. Rather, we note that sequence-level IBS will only be true of about $0.082/0.77 \approx 0.11$ of the regions that are inferred to be IBD by a program like BEAGLE, which used IBS at the genotype level to find segments of shared ancestry. In contrast, 77% of the 1 MB regions that we call IBD should also be identified as IBD by BEAGLE [10]. When looking for IBD in sequence data, it seems useful to drop the assumption that IBD implies IBS, just as it was necessary to change the definition of IBD when moving from pedigree analysis to the study of unrelated individuals.

The terms IBD and IBS were in fact both coined in the context of pedigree analysis, where a family with a history of a disease phenotype is scrutinized for genetic variants that might contribute to the appearance of that phenotype. Related individuals are genotyped at a sparse set of markers, and those markers are used, together with the family relationship pedigree, to find haplotypes that were often transmitted from diseased ancestors to diseased offspring [34, 37]. IBD sharing makes it likely that two individuals match at a long stretch of unobserved DNA, and inferring this matching is essential given that the variants causing the disease will almost certainly not be among the few directly genotyped marker sites.

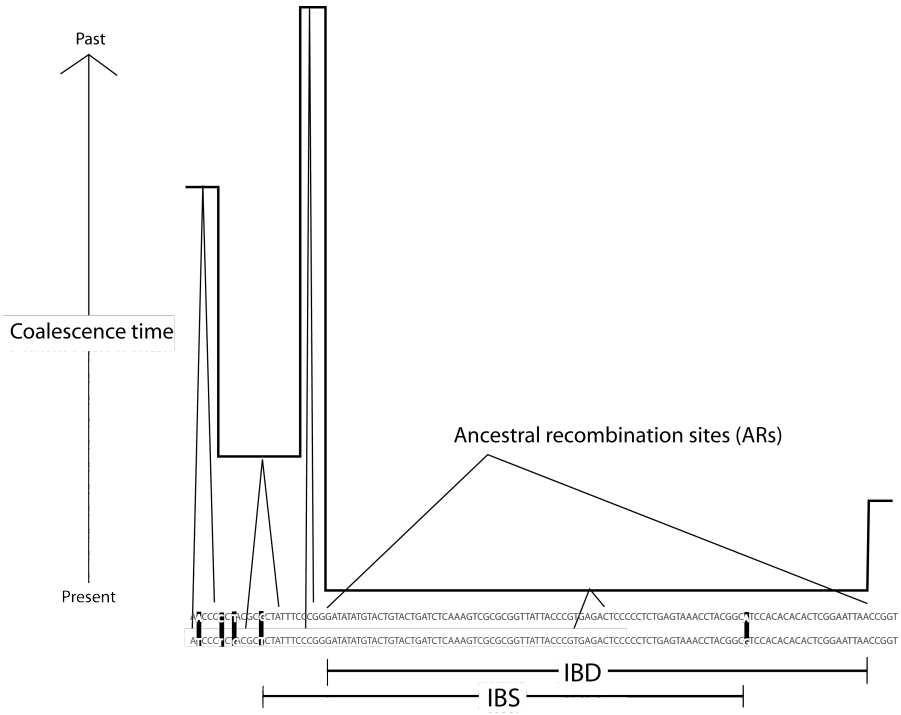

Figure 1: This picture illustrates the difference between IBS and IBD. IBS depends only on the observable differences between two sequences, while IBD depends on their hidden history: how long ago each site coalesces. Two sequences are IBD if each base coalesces at the same time, and IBS if each base matches by state (there are no internal hets). Long regions of IBS usually overlap with long regions of IBD, but as shown here, the regions rarely coincide exactly.

More data is available in a genome-wide association study (GWAS), where thousands of cases and controls are typed at hundreds of thousands of markers. However, it is impossible to know the family relationships among so many study individuals, making direct IBD inference more difficult than in a linkage study, and it is still likely that the causal variants will not be directly genotyped.

Rather than working to infer the genealogies of unobserved stretches of DNA, GWASs regard typed markers as one-to-one proxies for untyped markers, working to construct genotype sets for which each unobserved allele is usually coinherited with an observed allele. If the presence of allele A at observed locus x means that there is a 90% chance of observing allele B at locus y, then even if B is causal and A is not, it may be possible to observe a correlation between the presence of A and the disease. A strong pairwise association between A and B translates to a high correlation coefficient $r^2(A, B)$, which is calculated from the allele frequencies $f_{A(x)}$ and $f_{B(y)}$ along with the haplotype frequency $f_{A(x)B(y)}$:

$$r^2(A, B) = \frac{(f_{A(x)B(y)} - f_{A(x)}f_{B(y)})^2}{f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})} \tag{1}$$

A standard measure of a genetic tag set's efficacy is the percentage of untyped variable sites that are within $r^2 \geq 0.8$ of a typed SNP (see e.g. [6]).

By the definition given above, $r^2(A, B)$ is a statement about how often A and B occur together in extant individuals, not a statement about how much history the alleles have in common. McVean showed that $r^2(A(x), B(y))$ is related to the covariance between the coalescence times at sites x and y [42]; IBD histories are more probable when r^2 is close to 1, but knowing $r^2(A, B)$ is not sufficient to know the likelihood of IBD in the stretch between x and y. Similarly, Hayes, et al. showed that the mean r^2 for markers L bases apart is close to the frequency of L-base IBD stretches in the genome, but that their measure of IBD sharing has a lower variance than r^2 does, capturing strictly more information about the hidden history of the sequences [17].

Some have claimed that pairwise r^2 values behave badly when input into multivariate GWAS analyses, and that measures of IBD probability behave

much better. Terwilliger and Hiekkalinna argue that it is dangerous to assume that the correlation between a tag and a variant will be statistically independent of the correlation between the variant and the phenotype, and that this is a fatal flaw in the paradigm of using tag SNPs as one-to-one proxies for unobserved SNPs. In contrast, they argue that IBD sharing *should* be independent of whether any loci involved are functional [56], and that linkage studies may be inherently more powerful than GWAS as a result. Whether or not they are correct, the best of both worlds solution may be to conduct GWAS as much like linkage studies as possible, finding ways to look for IBD sharing, rather than simple IBS association, in genetic data sets that have no accompanying pedigree data.

Imputation can be viewed as a step toward making GWAS more like linkage studies, inferring IBD with the help of population genetics rather than pedigrees [38]. To avoid assuming that the effects of untyped variants will automatically show up by proxy association, these variants are imputed into test sequences and screened for association directly. Imputation is performed where IBD sharing is suspected between a sample and a reference haplotype, taking advantage of the good evidence for IBD that is provided by long IBS marker strings. We are able to compute precisely how long these marker strings must be for p(IBD|IBS) to be sufficiently close to 1, analytically predicting when imputation should be reliable.

Detecting IBD is especially important when causal variants are very rare or have very modest phenotypic effects. Several variants that affect the same condition may be clustered around an important protein or promoter, in which case it may be possible to pool their signals, i.e. regard the whole region as single locus where haplotypes are the alleles. The number of individuals with causal variants in the region should exceed the number with variants at any particular locus, and the pooled signal of these variants may just reach the threshold of detectability [9, 51, 53]. However, this approach depends on the ability to tell haplotypes apart based on marker IBS, and we will show that inferring a 1000-base haplotype with 99% accuracy requires imputing from nearly a megabase

of standard IBS marker data.

The probability p(IBD|IBS) depends on population history in a complex way; although long IBD tracts are most common in DNA from inbred groups, inbreeding actually increases the probability that a long IBS tract is not IBD [37, 52, 55]. False discoveries abound in linkage studies that do not adequately account for hidden founder relatedness, particularly with regard to long genetic loops that are seldom recorded in pedigrees [34, 37, 52]; however, the dependence of IBS sharing on population history can be useful as well as confounding, since the length distribution of shared IBS contains more information about population substructure than simpler measures like the coefficient of relatedness. Jakkula, et al., for example, found that the Finnish sub-populations have similar inbreeding coefficient distributions but differ significantly in their patterns of homozygosity and IBS sharing [28]. Similarly, Kong, et al. found long IBS sharing to be common in Iceland, though the average inbreeding coefficient (2.5×10^{-4}) was not especially high. In a collection of 35,528 Icelanders who were genotyped for a particular 10 Mb region, all but 1,995 shared that region IBS with another genotyped individual who was not closely related to them, enough to allow for long-range phasing within the population at large [31].

There exist several algorithms for estimating p(IBD|IBS), some conditioning on haplotype frequencies and some only on inheritance models. The data-dependent algorithms have the advantange of specificity, but they consistently underestimate p(IBD|IBS) because of the way they incorporate their test haplotype into their prior [9, 33, 46]. They can confirm that a medically interesting region is likely to be IBD, but are less useful for using IBD to study population history. p(IBD|IBS) has not been computed exactly with respect to the neutral coalescent, and we believe we are the first to compute it with respect to the sequentially Markovian coalescent [43].

Previous methods for estimating p(IBD|IBS) that do not condition on allele frequencies have begun to deduce the impact of history on genome-wide patterns of IBD [11, 17, 54, 55, 57]. However, most of them make assumptions that break down at certain segment lengths and marker densities, which prevents

them from making use of all available marker information. The PLINK hidden Markov model, for example, will only calculate $p(\text{IBD}|\text{IBS})$ between markers that are in linkage equilibrium with each other [33, 51]; their precision is limited by the sparseness of unlinked marker sets. A related assumption, which is implicitly made in all of the literature we found, is that the lengths of adjacent IBD segments are independently distributed [9, 17, 46, 51, 55], and we will show in Section 6 how this breaks down for large, dense data sets. Our method, in contrast, captures the dependence between the lengths of neighboring IBD segments, and can assume arbitrarily dense marker data without losing any accuracy. Given inputs of population size history, mutation rate, and recombination rate, we predict an ROH distribution that can be verified in genome data. After adjusting for the presence of sequencing errors, we are able to accurately predict the distribution of ROHs found in eleven complete human genome sequences.

Given that sequencing is much more costly than genotyping, we also adjust our method to predict IBS given a thinned-down set of markers. Our theory correctly predicts the distribution of segments that appear homozygous based on incomplete knowledge of the hets in the genome data, quantifying the correlation between IBS at the genotype level and IBS at the level of the complete sequence.

We also extend our theory to the case of unphased diploid sequences, deviating from the SMC slightly but checking the results against a full coalescent simulation. When phasing ambiguities are accounted for in this way, $p(\text{IBD}|\text{IBS})$ can be used to estimate the accuracy of an attempt at imputation and/or haplotype resolution. We conclude that both efforts become much more accurate if the ends of an IBS alignment are not considered likely to be IBD; when IBS is measured in a way that detects a het every 10,000 bases, it seems prudent to discard 10^5 bases from each end of an alignment, after which the probability of IBD is as great as if the full sequences were known. Finally, we estimate the accuracy spread of the imputation calls made from a panel of n reference haplotypes, showing that a thousand references should be sufficient in a population where recent exponential growth has not broken up moderately long stretches

of IBD sharing.

2 Computing the probability of identity by state

Since

$$p(\text{IBD}|\text{IBS}) = \frac{p(\text{IBD\&IBS})}{p(\text{IBS})}, \tag{2}$$

where p(IBD&IBS) is easy to compute (see equation (2.1)), the crux of our approach will be calculating p(IBS) given sequence length and the history of the effective population size. In section 2.1, we treat the case of constant effective population size, while section 2.3 describes how to condition on any locally constant population size history.

2.1 Constant effective population size

Let L be the length of an alignment between two haplotypes sampled at random from a diploid population of effective size N. Assume that the DNA undergoes m mutations per base per generation and r recombinations per base per generation, letting $\mu = 4Nm$ and $\rho = 4Nr$. We will hereafter measure time in units of $2N$ generations.

The alignment will coalesce at time t, both IBD and IBS, if and only if the following events coincide:

1. The leftmost locus coalesces at time t without mutating (probability $e^{-t(1+\mu)}dt$)
2. No other base in either sequence undergoes a mutation or a recombination between time zero and time t (probability $e^{-t(L-1)(\mu+\rho)}$)

From this observation, it follows that

$$p(\text{IBD\&IBS}) = \int_{t=0}^{\infty} e^{-t(L-1)(\mu+\rho)} \cdot e^{-t(1+\mu)}dt = \frac{1}{1 + L\mu + (L-1)\rho}. \tag{3}$$

In an analogous way, we will derive the probability $p_L(\text{IBS}|t)dt$ that the alignment coalesces IBS with its rightmost base coalescing at time t. We proceed

of IBD sharing.

2 Computing the probability of identity by state

Since

$$p(\text{IBD}|\text{IBS}) = \frac{p(\text{IBD\&IBS})}{p(\text{IBS})}, \tag{2}$$

where p(IBD&IBS) is easy to compute (see equation (2.1)), the crux of our approach will be calculating p(IBS) given sequence length and the history of the effective population size. In section 2.1, we treat the case of constant effective population size, while section 2.3 describes how to condition on any locally constant population size history.

2.1 Constant effective population size

Let L be the length of an alignment between two haplotypes sampled at random from a diploid population of effective size N. Assume that the DNA undergoes m mutations per base per generation and r recombinations per base per generation, letting $\mu = 4Nm$ and $\rho = 4Nr$. We will hereafter measure time in units of $2N$ generations.

The alignment will coalesce at time t, both IBD and IBS, if and only if the following events coincide:

1. The leftmost locus coalesces at time t without mutating (probability $e^{-t(1+\mu)}dt$)
2. No other base in either sequence undergoes a mutation or a recombination between time zero and time t (probability $e^{-t(L-1)(\mu+\rho)}$)

From this observation, it follows that

$$p(\text{IBD\&IBS}) = \int_{t=0}^{\infty} e^{-t(L-1)(\mu+\rho)} \cdot e^{-t(1+\mu)}dt = \frac{1}{1 + L\mu + (L-1)\rho}. \tag{3}$$

In an analogous way, we will derive the probability $p_L(\text{IBS}|t)dt$ that the alignment coalesces IBS with its rightmost base coalescing at time t. We proceed

by induction on the length variable L, claiming that

$$\begin{aligned} p_L(\text{IBS}|t)dt &= p_{L-1}(\text{IBS}|t)dt \cdot e^{-t(\mu+\rho)} \\ &+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS}|t_0)e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dtdt_rdt_0 \\ &+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS}|t_0)e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dtdt_rdt_0. \end{aligned}$$

The dummy variable t_0 is the coalescence time of the base next to the rightmost one. The first term is the probability that no recombination occurs between the rightmost base of the alignment and the base next to it, while the second term (the first integral) is the probability that a recombination occured at some time t_r, and that t is greater than t_0. The third term accounts for the remaining possibilities, integrating over times t_0 that are greater than t.

It will be convenient to write

$$p_L(\text{IBS}|t) = \sum_{i=1}^{L} A_i(L)e^{-t(1+i\mu+(i-1)\rho)}dt \tag{4}$$

and solve for the coefficients $A_1(L), \cdots, A_L(L)$, which will not depend on t.

Since

$$p_1(\text{IBS}|t) = e^{-t(1+\mu)}dt$$

and

$$\begin{aligned} & e^{-t_0(1+i\mu+(i-1)\rho)}\cdot e^{-t(\mu+\rho)} + \\ & \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} e^{-t_0(1+i\mu+(i-1)\rho)}e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dt_rdt_0 + \\ & \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} e^{-t_0(1+i\mu+(i-1)\rho)}e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dt_rdt_0 \\ = & \frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho)}e^{-t(\mu+1)} + \\ & \left(1-\frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho)}\right)e^{-t(1+(i+1)\mu+i\rho))}, \end{aligned}$$

we can let

$$C_i = \frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho))}$$

and conclude that

$$A_1(L) = \sum_{i=1}^{L-1} C_i A_i(L-1), \tag{5}$$

while

$$A_i(L) = (1 - C_{i-1}) A_{i-1}(L-1) \tag{6}$$

for $i > 1$.

Integrating equation (4) with respect to time, we find that

$$p_L(\text{IBS}) = \sum_{i=1}^{L} \frac{A_i(L)}{1 + i\mu + (i-1)\rho}. \tag{7}$$

Although it is time-intensive to compute $A_1(L), \ldots, A_L(L)$ for $L \gg 10^4$, the run time can be decreased by picking an appropriate constant c and substituting $(c\mu, c\rho, L/c)$ for (μ, ρ, L). This approximation reduces the run time c^2-fold, and Figure 2 records its modest effect on the computation accuracy.

The reader may prefer to think about $p_L(\text{IBS})$ using matrix algebra rather than recursion, seeing that

$$p_L(\text{IBS}) = \begin{pmatrix} \frac{1}{1+\mu} & \frac{1}{1+2\mu+\rho} & \cdots & \frac{1}{1+L\mu+(L-1)\rho} \end{pmatrix} \begin{pmatrix} C_1 & C_2 & \cdots & C_{L-1} & C_L \\ 1-C_1 & 0 & \cdots & 0 & 0 \\ 0 & 1-C_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-C_{L-1} & 0 \end{pmatrix}^L \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It is important to note that we have been talking about ROHs that are *at least* L bases long; when we compare our results to real genome data in Section 7, we will need to know the frequency of ROHs that are *exactly* L bases long. The following is the probability $p_{L\max}(\text{IBS})$ of observing an L-base IBS stretch ending with a het:

$$\begin{aligned} p_{L\max}(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \int_{t=0}^{\infty} e^{-t(1+i\mu+(i-1)\rho))}(1 - e^{-(\mu+\rho)t})dt \\ &= \sum_{i=1}^{L} \frac{A_i(L)(\mu+\rho)}{(1+i\mu+(i-1)\rho)(1+(i+1)\mu+i\rho)}. \end{aligned}$$

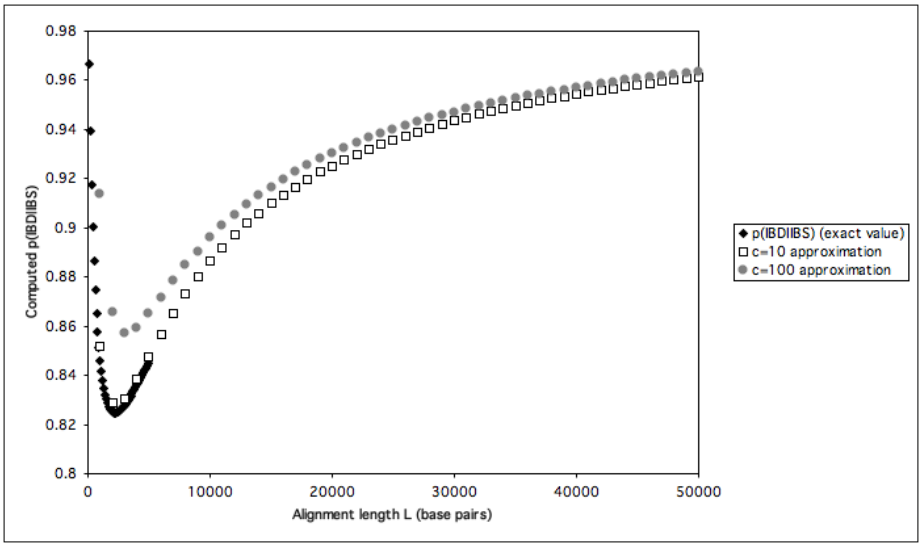

Figure 2: The parameter change $(\mu, \rho, L) \to (c\mu, c\rho, L/c)$ has its greatest effect when L is small. For $L = 1000$, the true value of $p(\text{IBD}|\text{IBS})$ is 0.8459; the calculated value increases to 0.8516 when we let $c = 10$, and increases to 0.9136 when we let $c = 100$. For $L = 50000$, the difference between the $c = 10$ value and the $c = 100$ value is only 0.0131, and taking $c = 100$ makes it practical to compute $p_L(\text{IBS})$ for L in the megabase range.

2.2 Non-uniform mutation and recombination

We have been assuming that μ and ρ are constant throughout the alignment to simplify the formulas as much as possible. However, it is easy to calculate p(IBD—IBS) exactly even when each locus i, $1 \leq i \leq n$, has a distinct mutation rate μ_i and recombination rate ρ_i. If we let $\vec{\mu}$ and $\vec{\rho}$ denote the vectors $(\mu_1, \ldots, \mu_n)$ and $(\rho_1, \ldots, \rho_n)$, it is easy to check (by generalizing the integrals in section 2.1) that

$$p_L(\text{IBS}|\vec{\mu}, \vec{\rho}) = \sum_{i=1}^{L} \frac{A_i(L, \vec{\mu}, \vec{\rho})}{((\mu_1 + \rho_1) + \cdots + (\mu_i + \rho_i))(1 + (\mu_1 + \rho_1) + \cdots + (\mu_{i-1} + \rho_{i-1}) + \mu_i)},$$

where

$$A_i(L, \vec{\mu}, \vec{\rho}) = (1 - C_{i-1}(\vec{\mu}, \vec{\rho}))A_{i-1}(L-1, \vec{\mu}, \vec{\rho})$$

and

$$A_1(L, \vec{\mu}, \vec{\rho}) = \sum_{i=1}^{L-1} C_i(\vec{\mu}, \vec{\rho})A_i(L-1, \vec{\mu}, \vec{\rho})$$

for

$$C_i(\vec{\mu}, \vec{\rho}) = \frac{\rho_i}{((\mu_1 + \rho_1) + \cdots + (\mu_i + \rho_i))(1 + (\mu_1 + \rho_1) + \cdots + (\mu_{i-1} + \rho_{i-1}) + \mu_i)}.$$

2.3 Correcting for changes in effective population size

Because most human populations have undergone growth and/or bottlenecking, we describe how to correct our model for historical changes in effective population size. We work through the example of a simple bottleneck, but the same method can accommodate any locally constant function $N(t)$.

We model a bottleneck following the convention in the coalescent theory reference [18], using a piecewise-constant time transform $t \to \tau(t)$. We suppose that the population began at size aN before the bottleneck, dipped to size fN during the time interval $[t_{B2}, t_{B1}]$, and has existed stably at size N from time t_{B2} to the present. The values t_{B1}, t_{B2}, and t_{B3} are measured in generations before the present, but we must map them to times $\tau(t)$ measured in units of $2N$ generations before the present:

$$\tau(t) = \begin{cases} (t - t_{B1})/(2Na) + (t_{B1} - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t > t_{B1} \\ (t - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t_{B1} < t < t_{B2} \\ t/(2N) & \text{if } t < t_{B2} \end{cases}$$

In addition to scaling t, we must scale μ and ρ, since each contains a factor of N.

When we make these modifications, equation (3) becomes

$$\begin{aligned} p_L(\text{IBD}) &= \int_{\tau=0}^{\tau(t_{B2})} e^{-\tau\cdot(1+L(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B2})}^{\tau(t_{B1})} e^{-\tau\cdot(1+Lf(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B1})}^{\infty} e^{-\tau\cdot(1+La(\mu+\rho))}d\tau \\ &= \frac{1 - e^{-\tau(t_{B2})(1+L(\mu+\rho))}}{1 + L(\mu+\rho)} + \frac{e^{-\tau(t_{B2})(1+Lf(\mu+\rho))} - e^{-\tau(t_{B1})(1+Lf(\mu+\rho))}}{1 + Lf(\mu+\rho)} \\ &\quad + \frac{e^{-\tau(t_{B1})(1+La(\mu+\rho))}}{1 + La(\mu+\rho)}. \end{aligned}$$

In the same way, we can correct $A_1(L), \ldots, A_L(L)$ for the bottleneck by replacing

$$C_i = \frac{\rho}{i(\mu+\rho)(1 + i(\mu+\rho))}$$

with

$$\begin{aligned} C_i &= \frac{\rho}{i(\mu+\rho)} \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

In terms of these corrected $A_i(L)$, we deduce that

$$\begin{aligned} p_L(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

3 The age distribution of maximal IBD segments

Our calculations, along with those in earlier papers, make it clear that IBD segment length is inversely related to age. In [17], Hayes, et al. go as far

$$\tau(t) = \begin{cases} (t - t_{B1})/(2Na) + (t_{B1} - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t > t_{B1} \\ (t - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t_{B1} < t < t_{B2} \\ t/(2N) & \text{if } t < t_{B2} \end{cases}$$

In addition to scaling t, we must scale μ and ρ, since each contains a factor of N.

When we make these modifications, equation (3) becomes

$$\begin{aligned} p_L(\text{IBD}) &= \int_{\tau=0}^{\tau(t_{B2})} e^{-\tau\cdot(1+L(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B2})}^{\tau(t_{B1})} e^{-\tau\cdot(1+Lf(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B1})}^{\infty} e^{-\tau\cdot(1+La(\mu+\rho))}d\tau \\ &= \frac{1 - e^{-\tau(t_{B2})(1+L(\mu+\rho))}}{1 + L(\mu+\rho)} + \frac{e^{-\tau(t_{B2})(1+Lf(\mu+\rho))} - e^{-\tau(t_{B1})(1+Lf(\mu+\rho))}}{1 + Lf(\mu+\rho)} \\ &\quad + \frac{e^{-\tau(t_{B1})(1+La(\mu+\rho))}}{1 + La(\mu+\rho)}. \end{aligned}$$

In the same way, we can correct $A_1(L), \ldots, A_L(L)$ for the bottleneck by replacing

$$C_i = \frac{\rho}{i(\mu+\rho)(1 + i(\mu+\rho))}$$

with

$$\begin{aligned} C_i &= \frac{\rho}{i(\mu+\rho)} \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

In terms of these corrected $A_i(L)$, we deduce that

$$\begin{aligned} p_L(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

3 The age distribution of maximal IBD segments

Our calculations, along with those in earlier papers, make it clear that IBD segment length is inversely related to age. In [17], Hayes, et al. go as far

as to draw a one-to-one correspondence between the abundance of maximal c-centimorgan IBD segments and the effective size of the population $1/(1+4c)$ generations ago. However, we show here that the mean coalescence time of an L-base IBD tract ($O(1/L)$) is much less than its standard deviation ($O(1/\sqrt{L})$), implying that the segments coalescing at time t have a significant length spread, particularly when t is very ancient. This complicates the effect of population size changes on the distribution of ROH length, particularly for shorter ROHs. While Hayes, et al. studied the distribution of ROHs that were 10^6 to 10^7 base pairs long and found their assumption useful at that length scale, we find that the relationship between effective population size and ROH length is more complicated for shorter ROHs, as we will see corroborated by data in Section 7 (Figures 14, 15, and 16).

As we saw in Section 2, the probability of an L-base ROH being IBD is

$$\int_{t=0}^{\infty} e^{-t(1+L\rho)}dt,$$

while the proability that it will be maximally IBD (i.e. not contained in a larger IBD segment) is

$$\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt.$$

We can use this to calculate a joint distribution between IBD segment length and coalescence time:

$$p_L(t|\text{IBD}) = \frac{e^{-t(1+L\rho)}(1-e^{-t\rho})^2}{\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt}. \quad (8)$$

We compute that

$$\begin{aligned}\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt &= \frac{1}{1+L\rho} - \frac{2}{1+(L+1)\rho} + \frac{1}{1+(L+2)\rho} \\ &= \frac{2\rho^2}{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)},\end{aligned}$$

such that

$$p_L(t|\text{IBD}) = \frac{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}{2\rho^2}e^{-t(1+L\rho)}(1-e^{-t\rho})^2. \quad (9)$$

Similarly, we can compute the expected t value $E_t(L)$, measured, as always, in units of $2N$ generations:

$$\begin{aligned} E_t(L) &= \int_{t=0}^{\infty} t p_L(t|\text{IBD}) dt \\ &= \frac{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}{\rho^2} \\ &\quad \cdot \left(\frac{1}{(1+L\rho)^2} - \frac{2}{1+(L+1)\rho)^2} + \frac{1}{(1+(L+2)\rho)^2} \right) \\ &= \frac{3L^2\rho^2 + 6L\rho^2 + 6L\rho + 2\rho^2 + 6\rho + 3}{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}. \end{aligned}$$

This differs from $1/(1+L\rho)$, the value given by Hayes, et al., because they don't distinguish between maximal and non-maximal IBD.

We go on to compute the variance

$$\begin{aligned} E_{t^2}(L) - E_t(L)^2 &= \int_{t=0}^{\infty} t^2 p_L(t|\text{IBD}) dt - \left(\int_{t=0}^{\infty} t p_L(t|\text{IBD}) dt \right)^2 \\ &= \frac{12(L^3 + \rho^2 L^2 + 2\rho L + \rho^2 + \rho + 1)(3\rho^2 L^2 + 6\rho^2 L + 10\rho^2 + 6\rho + 3)}{(1+\rho L)^2(1+\rho(L+1))^2(1+(L+2))^2} \\ &\quad - \frac{12(\rho L + 1)(\rho L + \rho + 1)}{(1+\rho(L+1))^2(1+\rho(L+2))^2} \\ &\quad - \frac{(3L^2\rho^2 + 6L\rho^2 + 6L\rho + 2\rho^2 + 6\rho + 3)^2}{(1+L\rho)^2(1+(L+1)\rho)^2(1+(L+2)\rho)^2}. \end{aligned}$$

Looking at the leading terms, we note that

$$E_t(L) \approx \frac{3}{\rho L} \ll \sqrt{E_{t^2}(L) - E_t(L)^2} \approx \frac{6}{\rho^2 \sqrt{L}},$$

meaning that the standard deviation of $E_t(L)$ is much greater than its mean.

Figure 3 shows the length distribution of IBS segments that coalesce $0.2N$ generations ago, while Figure 4 plots the length distribution of segments that coalesce $0.3N$ generations ago. Even if IBD were the same as IBS and segments coalesced at only these two times, it would not be straightforward to look at a sum of plots like this and quantify an excess of one type of segment. Hayes, et al. track recent population growth by assuming that a dearth of L-base IBD segments means a larger population at time $1/(1+L\rho)$, but it would seem that this approach must be modified for shorter L where the length and coalescence

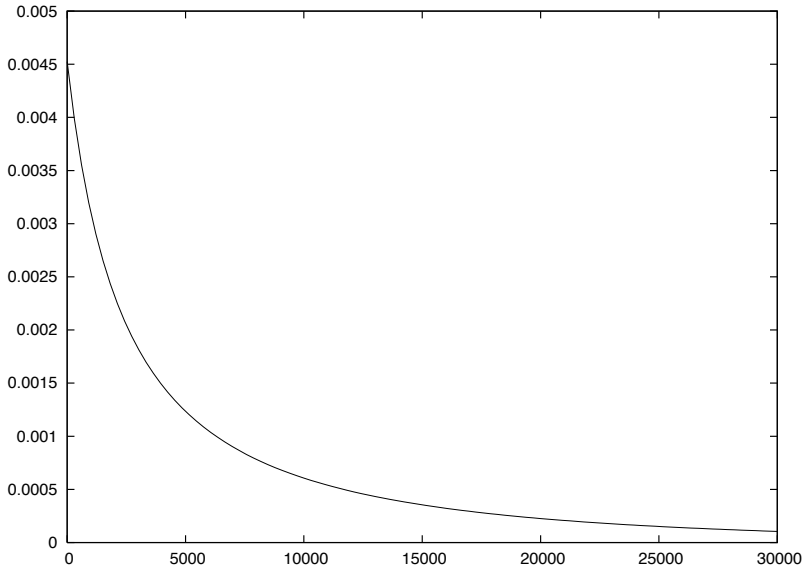

Figure 3: This plot shows the length spread of IBD segments that coalesce $0.2N$ generations ago. Comparing this to Figure 4, we see that it will be difficult to tell these segments apart from segments that coalesced $0.3N$ generations ago.

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 $\text{IBS}(L_1, L_2) \geq 1$ 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

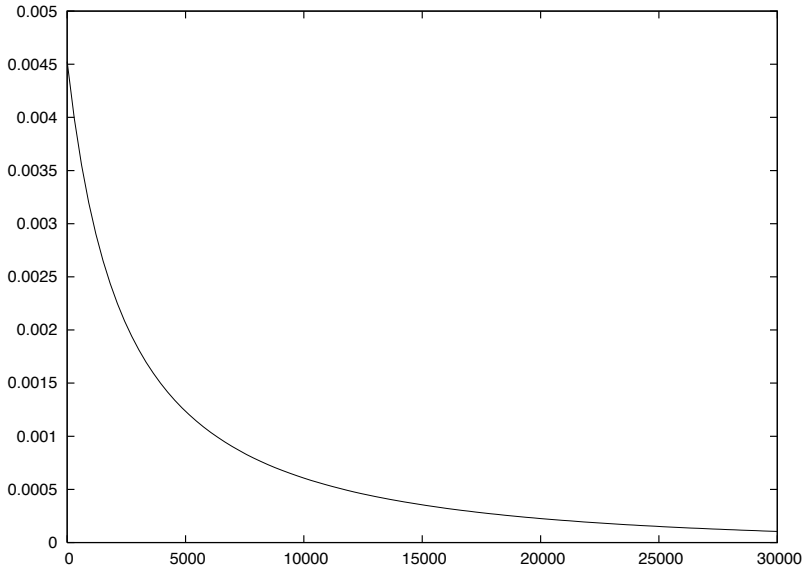

Figure 3: This plot shows the length spread of IBD segments that coalesce $0.2N$ generations ago. Comparing this to Figure 4, we see that it will be difficult to tell these segments apart from segments that coalesced $0.3N$ generations ago.

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 $\text{IBS}(L_1, L_2) \geq 1$ 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

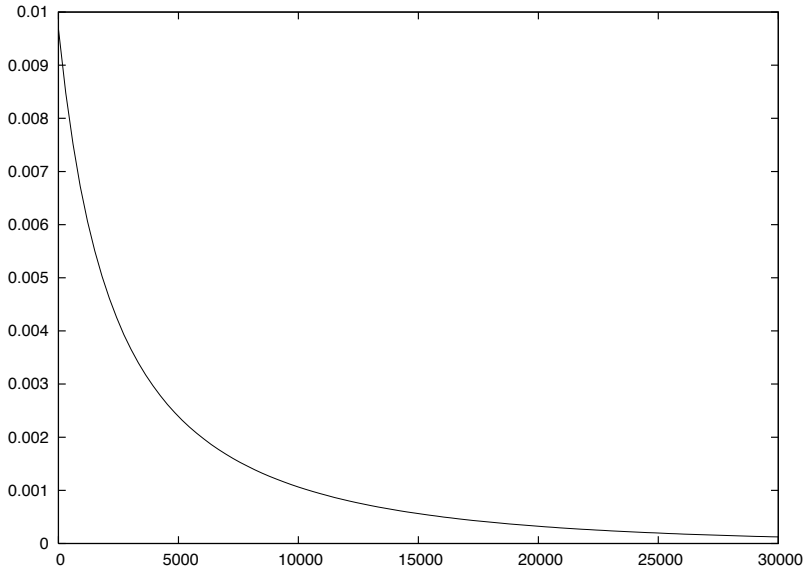

Figure 4: The length spread of IBD segments that coalesce $0.3N$ generations ago is different from the spread of segments that coalesce $0.2N$ generations ago (Figure 3), but overlaps enough that it would take a bit of work to learn about population history from a sum of density plots like these.

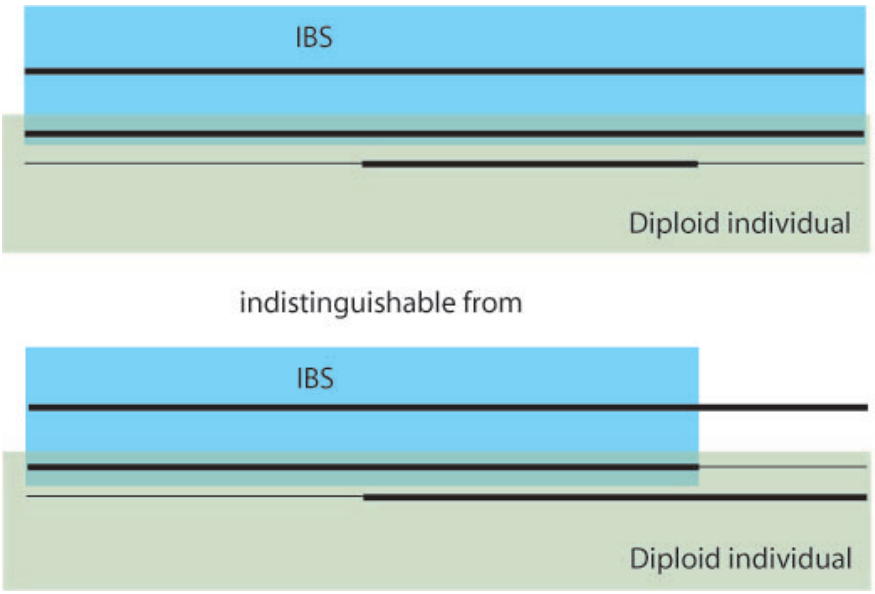

Figure 5: **IBS between phased haplotypes vs. IBS≥1** Here, the two chromosomes of a diploid individual are aligned to a reference haplotype. The diploid DNA is drawn in bold where it matches the reference haplotype IBS. Both the top and the bottom alignment have the property IBS≥ 1, where at least one 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 $\text{IBS}(r, d) \geq 1$ 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 $\text{IBS}(r, d) \geq 1$, 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).

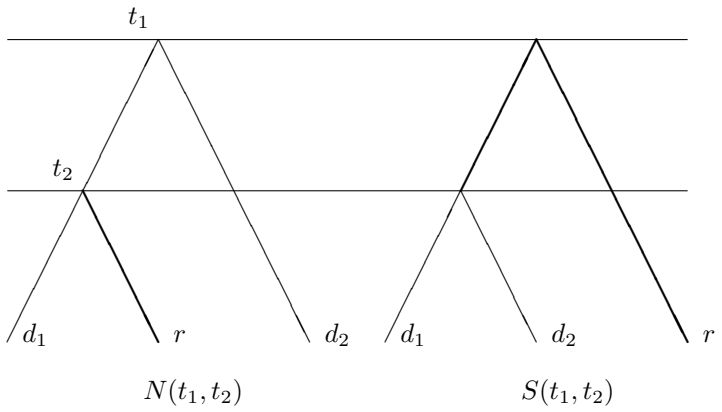

Figure 6: Natural and skew tree topologies. Boldface branches are drawn where mutations cannot occur without rendering the alignment non-IBS.

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 $(r(d_1 d_2))$. 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:

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

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:

$$\begin{aligned} P_S(\text{IBD\& IBS}) &= \int_{t_2=0}^{\infty}\int_{t_1=t_2}^{\infty} e^{-t_2(2-L(\rho+\mu))-t_1(1+2L(\rho+\mu))}dt_1dt_2 \quad (12)\\ &= \frac{1}{(1+2L(\rho+\mu))(3+L(\rho+\mu))} \quad (13)\end{aligned}$$

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:

$$\begin{aligned}
P_L(\text{IBS},t)dt &= p_{L-1}(\text{IBS},t)dt \cdot e^{-t(\mu+2\rho)} \\
&+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS},t_0)\cdot e^{-\mu t}\left(2\rho e^{-2\rho t_r}\cdot\frac{2}{3}\cdot 3e^{-3(t-t_r)}dt\right. \\
&\quad \left. +2\rho e^{-2\rho t_r}\int_{t_2=t_r}^{t}\frac{1}{3}\cdot 3e^{-3(t_2-t_r)}\cdot e^{-(t-t_2)}dtdt_2\right)dt_r dt_0 \\
&+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS},t_0)\cdot e^{-\mu t}\left(\frac{2}{3}\cdot 3\rho e^{-3\rho t_r}\cdot\frac{1}{2}\cdot 2e^{-2(t-t_r)}+\right. \\
&\quad \left. +\frac{1}{3}\cdot 3\rho e^{-3\rho t_r}\cdot 2e^{-2(t-t_r)}dt\right)dt_r dt_0 \\
&= p_{L-1}(\text{IBS},t)dt\cdot e^{-t(\mu+2\rho)} \\
&+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS},t_0)\left(3\rho e^{-t(3+\mu)+t_r(3-2\rho)}\right. \\
&\quad \left. +\rho e^{-t(1+\mu)+t_r(1-2\rho)}\right)dt_r dt_0 dt \\
&+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS},t_0)\cdot 4\rho e^{-t(2+\mu)+t_r(3-2\rho)}dt_r dt_0 dt.
\end{aligned}$$

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

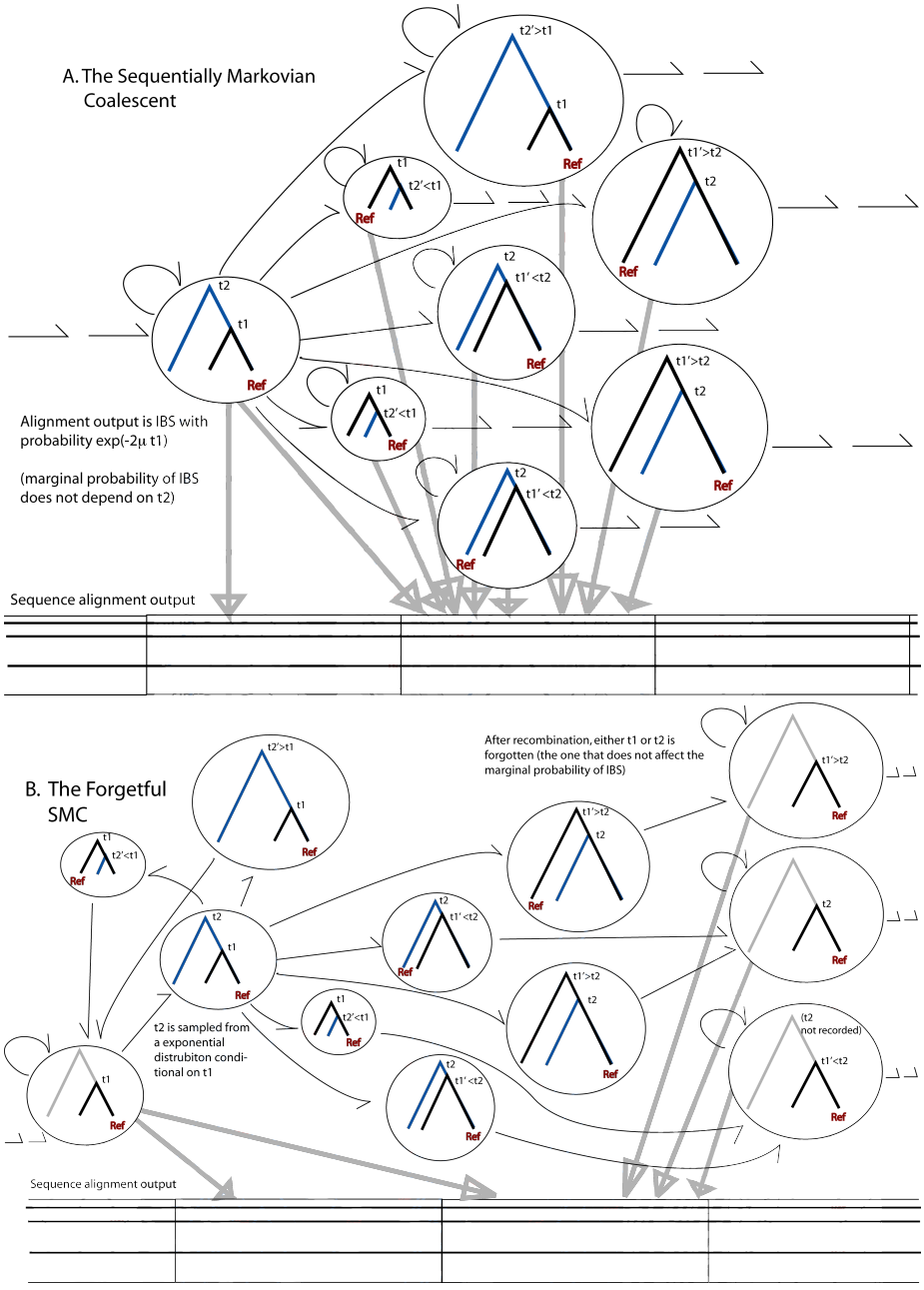

Figure 7: Hidden Markov model flow diagrams of the SMC and our forgetful approximation of the SMC. The position of the reference sequence is labeled to mark each genealogy as natural or skew. Each output extends the alignment by a triplet of bases (including one labeled reference base) that is either 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

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

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

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

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

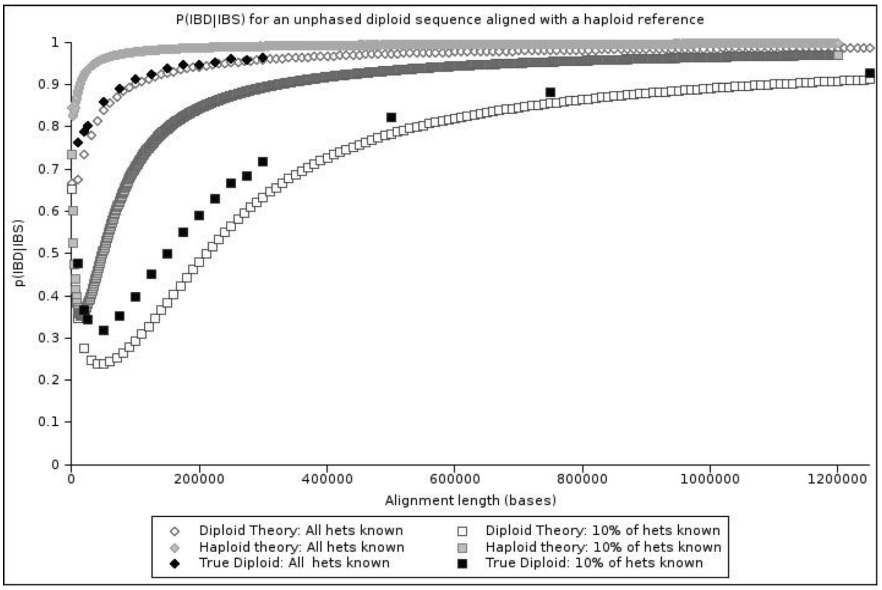

Figure 8: This plot compares our diploid results to the values that we obtained from MS simulations, which assume the full non-Markovian coalescent with recombination. Given a 10^5-base diploid sequence that is IBS ≥ 1 with a refernce, it is 90.1% likely to contain a haplotype that is IBD with the reference. 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

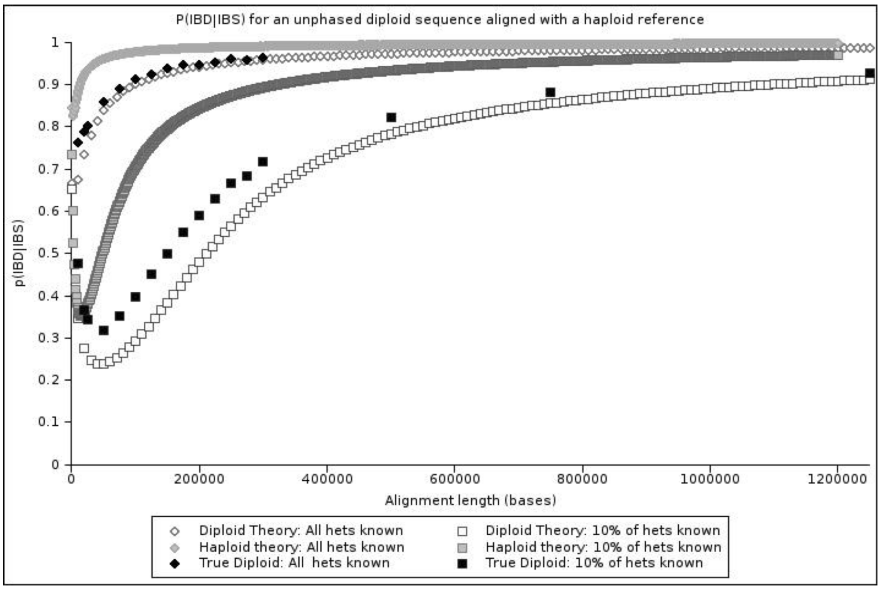

Figure 8: This plot compares our diploid results to the values that we obtained from MS simulations, which assume the full non-Markovian coalescent with recombination. Given a 10^5-base diploid sequence that is IBS ≥ 1 with a refernce, it is 90.1% likely to contain a haplotype that is IBD with the reference. 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

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

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

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

where $e^{-tL(2\mu+2\rho)} = p_L(\text{IBS\&IBD}|t)/e^{-t}$ 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:

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

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

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

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

The relationship of identity by state to identity by descent and imputation accuracy in population sequencing data

Kelley Harris
Emmanuel College
University of Cambridge

Dissertation submitted for the degree of Master of Philosophy

March 29, 2011

PREFACE

This dissertation is my own work. The research was not done in collaboration except where marked explicitly in the text.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Richard Durbin, for suggesting these research questions and providing so much guidance through the process of answering them and writing up this report. He and Heng Li, who created the genome call sets used in Section 7, will be my coauthors on a manuscript based on the work described in this dissertation. Besides thanking Heng for providing the data I needed to complete the project, I would like to thank everyone in the Durbin group for welcoming me and teaching me so much over the past year. In particular, Kimmo Palin and Jared Simpson helped me learn to use the Sanger computer system, and Kimmo, as well as Aylwyn Scally, helped me edit previous versions of this writeup. In writing and researching, I also benefitted from the helpful input of my thesis committee: Carl Anderson, Vincent Plagnol, Chris Tyler-Smith, and Eleftheria Zeggini, and would like to thank Gilean McVean and Jeffrey Barrett for taking on the finished product to read. I am grateful for the financial support I received from the Harvard College Herchel Smith Postgraduate Fellowship program, and to the Wellcome Trust Sanger Institute for hosting me. Finally, I'd like to thank my friends for always supporting me and my parents, Glenn Harris and Anne Katten, for doing that and so much more.

Abstract

When two DNA sequences look *identical by state* (IBS) along a string of genotyped markers, the DNA between the markers is often *identical by descent* (IBD), meaning inherited from a recent common ancestor without recombination. This fact makes it possible to scan the whole genome for functional variants without typing every base directly, making use of information about unobserved bases that is provided by the states of observed bases. We attempt to quantify that information here, using coalescent theory to predict how strongly various degrees of IBS imply IBD, taking into account the density of genotyped markers and past effective population size.

In addition to calculating the probability of IBD between IBS haploid sequences, we consider the problem of matching an unphased diploid sequence to a reference haplotype panel. The results have bearing on the practices of haplotype phasing and genotype imputation, both of which become more reliable when the ends of an IBS alignment are not assumed to be IBD. To compute p(IBD|IBS) when phasing ambiguity is an issue, it was necessary to develop a new approximation to the neutral coalescent with recombination: a further simplification of the sequentially Markovian coalescent [43].

Computing p(IBD|IBS) by our method is closely related to predicting the length distribution of homozygous stretches in the genome. After accounting for sequencing errors, we predict this distribution correctly, as judged by data from eleven complete human genomes. We also predict the length distribution of segments that appear homozygous based on thinned marker data, noting that the probability of sequence IBS given "thinned IBS" is a natural measure of imputation accuracy.

The probability of IBS given thinned IBS varies with sequence length in a way that is very ethnically distinctive, as judged by data from five Africans, four Europeans, and two Asians. We are able to account for these differences in terms of past changes in effective population size: an out-of Africa bottleneck followed by a shallower, more recent Asian bottleneck. We predict that IBS implies IBD most strongly in historically

outbred populations, and that extra care should be taken when inferring IBD in bottlenecked populations.

Returning to the problem of imputation, we estimate the accuracy spread of the imputation calls that can be made from a panel of n reference haplotypes. For an idealized population of effective size $N = 10,000$, we find that the 120-reference HapMap should omit a significant amount of genetic variation; that given a 1-kilobase stretch of a genotyped individual's DNA, a 120-reference panel gives us only a 70% chance of imputing the sequence of that stretch with 99% accuracy. However, we find that a 1000-haplotype panel should enable near-perfect imputation in a population that has been isolated from recent exponential population growth, and such perfect imputation would allow for precise genetic mapping in groups much larger than extended families.

Contents

1 Introduction

Every child is born with a few *de novo* mutations, DNA sites where they differ from their parents and from most other humans. Most of the variants created this way die out within a few generations, but a minority of them spread to hundreds or thousands of the child's descendants and contribute to widespread human genetic variation [1, 27]. By mathematically modeling the emergence and spread of new alleles, population geneticists can make inferences about ancient periods of growth, decline, interbreeding, and the emergence of modern ethnic groups, as well as discover links between genetic and phenotypic variation.

Given DNA from one individual, it is much cheaper to genotype a few thousand genetic loci than to ascertain the entire genome sequence, so companies like Illumina and Affymetrix manufacture single nucleotide polymorphism (SNP) chips that can selectively ascertain the states of between 10,000 and 1,000,000 of the most variable sites in humans. By focusing on fewer genetic sites, one can afford to genotype those sites in more individuals, and this approach has been used since the invention of pedigree analysis to find many sites in the genome that correlate with disease risk or recent positive selection [36]. A problem with SNP chips, however, is that they omit sites where variant alleles arose too recently to spread to a significant fraction of the human population. Although none of the three billion sites that are omitted from a SNP chip is especially variable on its own, together they harbor a vast amount of additional genetic information [1, 26]. This hard-to-detect variation is a clear candidate harbor for "missing heritability" in disease genetics, where known genetic risk factors usually fail to account for the full heritability of complex diseases [29, 51].

One way to detect more low-frequency variants will be to gather more genotype and sequence data, working to make this process cheaper through improvements in biotechnology. Another approach, however, is to extract more information from available data sets by modeling a process known as *linkage disequilibrium* (LD). Even when site X does not appear on a chip being used to to gather data, it can still be possible to infer that two sequences match at site

X by looking for matching at sites close to X. DNA is passed from parents to children in continuous blocks between recombination sites; when two sequences share a rare allele, it is likely that the allele was inherited from a recent common ancestor along with a block of surrounding DNA containing some sites that appear on the SNP chip [10, 31].

Linkage disequilibrium affects the distribution of heterozygous sites (*hets*) in every diploid genome, even in outbred populations. If every site in the genome had an independent probability m of being a het, then the probability of an L-base region being devoid of hets, or *identical by state* (IBS) would be $(1-m)^L \approx e^{-L}$. The frequency of L-base regions of homozygosity (ROHs) is not observed to decline exponentially with L, however [41, 55], and the excess of long ROHs can be accounted for by modeling LD. If, for example, an individual's parents are ninth-degree cousins, there is only a one-in-2^{20} chance that both alleles at a given site in the child's DNA were inherited from the parents' most recent common ancestor, but given that both alleles *were* both inherited from that ancestor, the child is likely to be homozygous over 10 megabases of surrounding DNA [31]. Ten generations is not enough time for meiosis to break the DNA into smaller heritable pieces, and in general, the length of a homozygous stretch is inversely proportional to the age of the ancestor that the matching haplotypes derive from.

The key to understanding how hets are placed is understanding how coalescence time, or time to common ancestry, varies from site to site across the genome. We define *ancestral recombination sites* (ARs) to be loci where two neighboring allele pairs coalesce at different times, and say that an alignment is *identical by descent* (IBD) if it has no interior ARs (See Figure 1 for an illustration of IBD vs. IBS).

A consequence of coalescent theory is that hets are placed randomly within an IBD region, with their density proportional to the region's coalescence time t ($t = 0$ being the present and larger t's being more ancient). As we move from left to right across a region of IBD, each base has a constant probability of being a het and a constant probability of being an AR and ending the IBD stretch.

In human DNA, the het probability μt is about 2.5 times the AR probability ρt, such that each IBD region contains about 2.5 total hets. The length of the region will vary inversely with t, however, making the local density of hets very small when t is very small.

Commercial chips with at most 1,000,000 SNP sites can detect at most 10% of the hets in a diploid sequence. This suggests that, on average, an IBD region will contain 0.25 hets that are detectible with a 1,000,000 chip and that $e^{-0.25} > 0.77$ of all maximal IBD regions will appear IBS based on genotype data. In contrast, only $e^{-2.5} \approx 0.082$ of maximal IBD regions will appear IBS based on sequence data.

Although definitions of IBD differ widely in the literature, IBD between two sequences is usually taken to imply IBS at the level of genotype data, and we do not intend to create confusion by defining IBD such that it does not imply IBS. Rather, we note that sequence-level IBS will only be true of about $0.082/0.77 \approx 0.11$ of the regions that are inferred to be IBD by a program like BEAGLE, which used IBS at the genotype level to find segments of shared ancestry. In contrast, 77% of the 1 MB regions that we call IBD should also be identified as IBD by BEAGLE [10]. When looking for IBD in sequence data, it seems useful to drop the assumption that IBD implies IBS, just as it was necessary to change the definition of IBD when moving from pedigree analysis to the study of unrelated individuals.

The terms IBD and IBS were in fact both coined in the context of pedigree analysis, where a family with a history of a disease phenotype is scrutinized for genetic variants that might contribute to the appearance of that phenotype. Related individuals are genotyped at a sparse set of markers, and those markers are used, together with the family relationship pedigree, to find haplotypes that were often transmitted from diseased ancestors to diseased offspring [34, 37]. IBD sharing makes it likely that two individuals match at a long stretch of unobserved DNA, and inferring this matching is essential given that the variants causing the disease will almost certainly not be among the few directly genotyped marker sites.

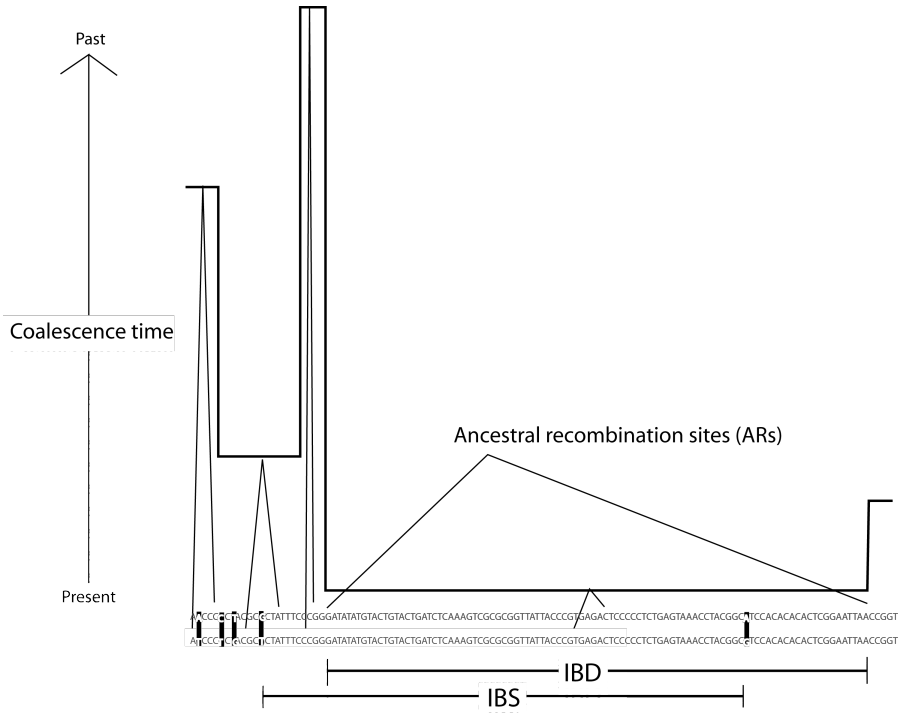

Figure 1: This picture illustrates the difference between IBS and IBD. IBS depends only on the observable differences between two sequences, while IBD depends on their hidden history: how long ago each site coalesces. Two sequences are IBD if each base coalesces at the same time, and IBS if each base matches by state (there are no internal hets). Long regions of IBS usually overlap with long regions of IBD, but as shown here, the regions rarely coincide exactly.

More data is available in a genome-wide association study (GWAS), where thousands of cases and controls are typed at hundreds of thousands of markers. However, it is impossible to know the family relationships among so many study individuals, making direct IBD inference more difficult than in a linkage study, and it is still likely that the causal variants will not be directly genotyped.

Rather than working to infer the genealogies of unobserved stretches of DNA, GWASs regard typed markers as one-to-one proxies for untyped markers, working to construct genotype sets for which each unobserved allele is usually coinherited with an observed allele. If the presence of allele A at observed locus x means that there is a 90% chance of observing allele B at locus y, then even if B is causal and A is not, it may be possible to observe a correlation between the presence of A and the disease. A strong pairwise association between A and B translates to a high correlation coefficient $r^2(A, B)$, which is calculated from the allele frequencies $f_{A(x)}$ and $f_{B(y)}$ along with the haplotype frequency $f_{A(x)B(y)}$:

$$r^2(A, B) = \frac{(f_{A(x)B(y)} - f_{A(x)}f_{B(y)})^2}{f_{A(x)}(1 - f_{A(x)})f_{B(y)}(1 - f_{B(y)})} \tag{1}$$

A standard measure of a genetic tag set's efficacy is the percentage of untyped variable sites that are within $r^2 \geq 0.8$ of a typed SNP (see e.g. [6]).

By the definition given above, $r^2(A, B)$ is a statement about how often A and B occur together in extant individuals, not a statement about how much history the alleles have in common. McVean showed that $r^2(A(x), B(y))$ is related to the covariance between the coalescence times at sites x and y [42]; IBD histories are more probable when r^2 is close to 1, but knowing $r^2(A, B)$ is not sufficient to know the likelihood of IBD in the stretch between x and y. Similarly, Hayes, et al. showed that the mean r^2 for markers L bases apart is close to the frequency of L-base IBD stretches in the genome, but that their measure of IBD sharing has a lower variance than r^2 does, capturing strictly more information about the hidden history of the sequences [17].

Some have claimed that pairwise r^2 values behave badly when input into multivariate GWAS analyses, and that measures of IBD probability behave

much better. Terwilliger and Hiekkalinna argue that it is dangerous to assume that the correlation between a tag and a variant will be statistically independent of the correlation between the variant and the phenotype, and that this is a fatal flaw in the paradigm of using tag SNPs as one-to-one proxies for unobserved SNPs. In contrast, they argue that IBD sharing *should* be independent of whether any loci involved are functional [56], and that linkage studies may be inherently more powerful than GWAS as a result. Whether or not they are correct, the best of both worlds solution may be to conduct GWAS as much like linkage studies as possible, finding ways to look for IBD sharing, rather than simple IBS association, in genetic data sets that have no accompanying pedigree data.

Imputation can be viewed as a step toward making GWAS more like linkage studies, inferring IBD with the help of population genetics rather than pedigrees [38]. To avoid assuming that the effects of untyped variants will automatically show up by proxy association, these variants are imputed into test sequences and screened for association directly. Imputation is performed where IBD sharing is suspected between a sample and a reference haplotype, taking advantage of the good evidence for IBD that is provided by long IBS marker strings. We are able to compute precisely how long these marker strings must be for p(IBD|IBS) to be sufficiently close to 1, analytically predicting when imputation should be reliable.

Detecting IBD is especially important when causal variants are very rare or have very modest phenotypic effects. Several variants that affect the same condition may be clustered around an important protein or promoter, in which case it may be possible to pool their signals, i.e. regard the whole region as single locus where haplotypes are the alleles. The number of individuals with causal variants in the region should exceed the number with variants at any particular locus, and the pooled signal of these variants may just reach the threshold of detectability [9, 51, 53]. However, this approach depends on the ability to tell haplotypes apart based on marker IBS, and we will show that inferring a 1000-base haplotype with 99% accuracy requires imputing from nearly a megabase

of standard IBS marker data.

The probability p(IBD|IBS) depends on population history in a complex way; although long IBD tracts are most common in DNA from inbred groups, inbreeding actually increases the probability that a long IBS tract is not IBD [37, 52, 55]. False discoveries abound in linkage studies that do not adequately account for hidden founder relatedness, particularly with regard to long genetic loops that are seldom recorded in pedigrees [34, 37, 52]; however, the dependence of IBS sharing on population history can be useful as well as confounding, since the length distribution of shared IBS contains more information about population substructure than simpler measures like the coefficient of relatedness. Jakkula, et al., for example, found that the Finnish sub-populations have similar inbreeding coefficient distributions but differ significantly in their patterns of homozygosity and IBS sharing [28]. Similarly, Kong, et al. found long IBS sharing to be common in Iceland, though the average inbreeding coefficient (2.5×10^{-4}) was not especially high. In a collection of 35,528 Icelanders who were genotyped for a particular 10 Mb region, all but 1,995 shared that region IBS with another genotyped individual who was not closely related to them, enough to allow for long-range phasing within the population at large [31].

There exist several algorithms for estimating p(IBD|IBS), some conditioning on haplotype frequencies and some only on inheritance models. The data-dependent algorithms have the advantange of specificity, but they consistently underestimate p(IBD|IBS) because of the way they incorporate their test haplotype into their prior [9, 33, 46]. They can confirm that a medically interesting region is likely to be IBD, but are less useful for using IBD to study population history. p(IBD|IBS) has not been computed exactly with respect to the neutral coalescent, and we believe we are the first to compute it with respect to the sequentially Markovian coalescent [43].

Previous methods for estimating p(IBD|IBS) that do not condition on allele frequencies have begun to deduce the impact of history on genome-wide patterns of IBD [11, 17, 54, 55, 57]. However, most of them make assumptions that break down at certain segment lengths and marker densities, which prevents

them from making use of all available marker information. The PLINK hidden Markov model, for example, will only calculate $p(\text{IBD}|\text{IBS})$ between markers that are in linkage equilibrium with each other [33, 51]; their precision is limited by the sparseness of unlinked marker sets. A related assumption, which is implicitly made in all of the literature we found, is that the lengths of adjacent IBD segments are independently distributed [9, 17, 46, 51, 55], and we will show in Section 6 how this breaks down for large, dense data sets. Our method, in contrast, captures the dependence between the lengths of neighboring IBD segments, and can assume arbitrarily dense marker data without losing any accuracy. Given inputs of population size history, mutation rate, and recombination rate, we predict an ROH distribution that can be verified in genome data. After adjusting for the presence of sequencing errors, we are able to accurately predict the distribution of ROHs found in eleven complete human genome sequences.

Given that sequencing is much more costly than genotyping, we also adjust our method to predict IBS given a thinned-down set of markers. Our theory correctly predicts the distribution of segments that appear homozygous based on incomplete knowledge of the hets in the genome data, quantifying the correlation between IBS at the genotype level and IBS at the level of the complete sequence.

We also extend our theory to the case of unphased diploid sequences, deviating from the SMC slightly but checking the results against a full coalescent simulation. When phasing ambiguities are accounted for in this way, $p(\text{IBD}|\text{IBS})$ can be used to estimate the accuracy of an attempt at imputation and/or haplotype resolution. We conclude that both efforts become much more accurate if the ends of an IBS alignment are not considered likely to be IBD; when IBS is measured in a way that detects a het every 10,000 bases, it seems prudent to discard 10^5 bases from each end of an alignment, after which the probability of IBD is as great as if the full sequences were known. Finally, we estimate the accuracy spread of the imputation calls made from a panel of n reference haplotypes, showing that a thousand references should be sufficient in a population where recent exponential growth has not broken up moderately long stretches

of IBD sharing.

2 Computing the probability of identity by state

Since

$$p(\text{IBD}|\text{IBS}) = \frac{p(\text{IBD\&IBS})}{p(\text{IBS})}, \tag{2}$$

where p(IBD&IBS) is easy to compute (see equation (2.1)), the crux of our approach will be calculating p(IBS) given sequence length and the history of the effective population size. In section 2.1, we treat the case of constant effective population size, while section 2.3 describes how to condition on any locally constant population size history.

2.1 Constant effective population size

Let L be the length of an alignment between two haplotypes sampled at random from a diploid population of effective size N. Assume that the DNA undergoes m mutations per base per generation and r recombinations per base per generation, letting $\mu = 4Nm$ and $\rho = 4Nr$. We will hereafter measure time in units of $2N$ generations.

The alignment will coalesce at time t, both IBD and IBS, if and only if the following events coincide:

1. The leftmost locus coalesces at time t without mutating (probability $e^{-t(1+\mu)}dt$)
2. No other base in either sequence undergoes a mutation or a recombination between time zero and time t (probability $e^{-t(L-1)(\mu+\rho)}$)

From this observation, it follows that

$$p(\text{IBD\&IBS}) = \int_{t=0}^{\infty} e^{-t(L-1)(\mu+\rho)} \cdot e^{-t(1+\mu)}dt = \frac{1}{1 + L\mu + (L-1)\rho}. \tag{3}$$

In an analogous way, we will derive the probability $p_L(\text{IBS}|t)dt$ that the alignment coalesces IBS with its rightmost base coalescing at time t. We proceed

by induction on the length variable L, claiming that

$$\begin{aligned} p_L(\text{IBS}|t)dt &= p_{L-1}(\text{IBS}|t)dt \cdot e^{-t(\mu+\rho)} \\ &+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS}|t_0)e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dtdt_rdt_0 \\ &+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS}|t_0)e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dtdt_rdt_0. \end{aligned}$$

The dummy variable t_0 is the coalescence time of the base next to the rightmost one. The first term is the probability that no recombination occurs between the rightmost base of the alignment and the base next to it, while the second term (the first integral) is the probability that a recombination occured at some time t_r, and that t is greater than t_0. The third term accounts for the remaining possibilities, integrating over times t_0 that are greater than t.

It will be convenient to write

$$p_L(\text{IBS}|t) = \sum_{i=1}^{L} A_i(L)e^{-t(1+i\mu+(i-1)\rho)}dt \tag{4}$$

and solve for the coefficients $A_1(L), \cdots, A_L(L)$, which will not depend on t.

Since

$$p_1(\text{IBS}|t) = e^{-t(1+\mu)}dt$$

and

$$\begin{aligned} & e^{-t_0(1+i\mu+(i-1)\rho)}\cdot e^{-t(\mu+\rho)} + \\ & \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} e^{-t_0(1+i\mu+(i-1)\rho)}e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dt_rdt_0 + \\ & \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} e^{-t_0(1+i\mu+(i-1)\rho)}e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dt_rdt_0 \\ = & \frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho)}e^{-t(\mu+1)} + \\ & \left(1-\frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho)}\right)e^{-t(1+(i+1)\mu+i\rho))}, \end{aligned}$$

we can let

$$C_i = \frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho))}$$

and conclude that

$$A_1(L) = \sum_{i=1}^{L-1} C_i A_i(L-1), \tag{5}$$

while

$$A_i(L) = (1 - C_{i-1}) A_{i-1}(L-1) \tag{6}$$

for $i > 1$.

Integrating equation (4) with respect to time, we find that

$$p_L(\text{IBS}) = \sum_{i=1}^{L} \frac{A_i(L)}{1 + i\mu + (i-1)\rho}. \tag{7}$$

Although it is time-intensive to compute $A_1(L), \ldots, A_L(L)$ for $L \gg 10^4$, the run time can be decreased by picking an appropriate constant c and substituting $(c\mu, c\rho, L/c)$ for (μ, ρ, L). This approximation reduces the run time c^2-fold, and Figure 2 records its modest effect on the computation accuracy.

The reader may prefer to think about $p_L(\text{IBS})$ using matrix algebra rather than recursion, seeing that

$$p_L(\text{IBS}) = \begin{pmatrix} \frac{1}{1+\mu} & \frac{1}{1+2\mu+\rho} & \cdots & \frac{1}{1+L\mu+(L-1)\rho} \end{pmatrix} \begin{pmatrix} C_1 & C_2 & \cdots & C_{L-1} & C_L \\ 1-C_1 & 0 & \cdots & 0 & 0 \\ 0 & 1-C_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-C_{L-1} & 0 \end{pmatrix}^L \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It is important to note that we have been talking about ROHs that are *at least* L bases long; when we compare our results to real genome data in Section 7, we will need to know the frequency of ROHs that are *exactly* L bases long. The following is the probability $p_{L\max}(\text{IBS})$ of observing an L-base IBS stretch ending with a het:

$$\begin{aligned} p_{L\max}(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \int_{t=0}^{\infty} e^{-t(1+i\mu+(i-1)\rho))}(1 - e^{-(\mu+\rho)t})dt \\ &= \sum_{i=1}^{L} \frac{A_i(L)(\mu+\rho)}{(1+i\mu+(i-1)\rho)(1+(i+1)\mu+i\rho)}. \end{aligned}$$

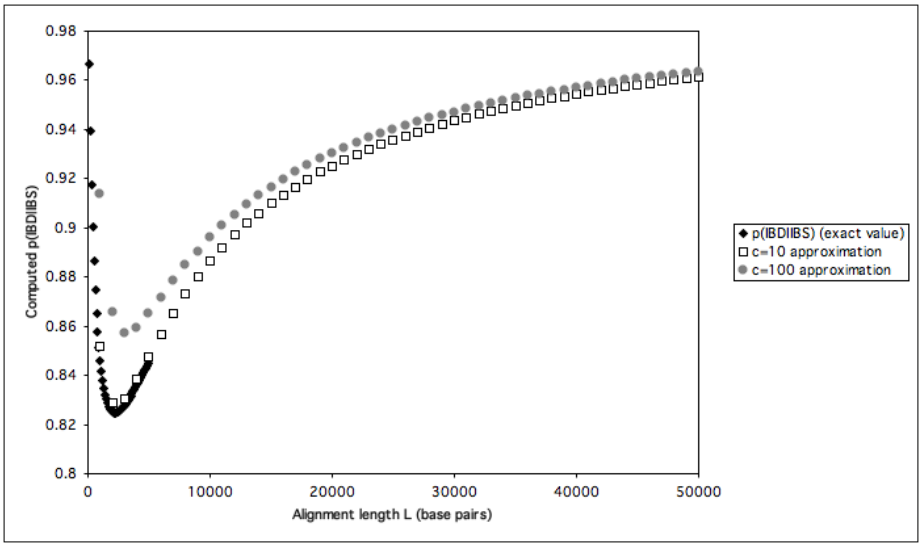

Figure 2: The parameter change $(\mu, \rho, L) \to (c\mu, c\rho, L/c)$ has its greatest effect when L is small. For $L = 1000$, the true value of $p(\text{IBD}|\text{IBS})$ is 0.8459; the calculated value increases to 0.8516 when we let $c = 10$, and increases to 0.9136 when we let $c = 100$. For $L = 50000$, the difference between the $c = 10$ value and the $c = 100$ value is only 0.0131, and taking $c = 100$ makes it practical to compute $p_L(\text{IBS})$ for L in the megabase range.

2.2 Non-uniform mutation and recombination

We have been assuming that μ and ρ are constant throughout the alignment to simplify the formulas as much as possible. However, it is easy to calculate p(IBD—IBS) exactly even when each locus i, $1 \leq i \leq n$, has a distinct mutation rate μ_i and recombination rate ρ_i. If we let $\vec{\mu}$ and $\vec{\rho}$ denote the vectors $(\mu_1, \ldots, \mu_n)$ and $(\rho_1, \ldots, \rho_n)$, it is easy to check (by generalizing the integrals in section 2.1) that

$$p_L(\text{IBS}|\vec{\mu}, \vec{\rho}) = \sum_{i=1}^{L} \frac{A_i(L, \vec{\mu}, \vec{\rho})}{((\mu_1 + \rho_1) + \cdots + (\mu_i + \rho_i))(1 + (\mu_1 + \rho_1) + \cdots + (\mu_{i-1} + \rho_{i-1}) + \mu_i)},$$

where

$$A_i(L, \vec{\mu}, \vec{\rho}) = (1 - C_{i-1}(\vec{\mu}, \vec{\rho}))A_{i-1}(L-1, \vec{\mu}, \vec{\rho})$$

and

$$A_1(L, \vec{\mu}, \vec{\rho}) = \sum_{i=1}^{L-1} C_i(\vec{\mu}, \vec{\rho})A_i(L-1, \vec{\mu}, \vec{\rho})$$

for

$$C_i(\vec{\mu}, \vec{\rho}) = \frac{\rho_i}{((\mu_1 + \rho_1) + \cdots + (\mu_i + \rho_i))(1 + (\mu_1 + \rho_1) + \cdots + (\mu_{i-1} + \rho_{i-1}) + \mu_i)}.$$

2.3 Correcting for changes in effective population size

Because most human populations have undergone growth and/or bottlenecking, we describe how to correct our model for historical changes in effective population size. We work through the example of a simple bottleneck, but the same method can accommodate any locally constant function $N(t)$.

We model a bottleneck following the convention in the coalescent theory reference [18], using a piecewise-constant time transform $t \to \tau(t)$. We suppose that the population began at size aN before the bottleneck, dipped to size fN during the time interval $[t_{B2}, t_{B1}]$, and has existed stably at size N from time t_{B2} to the present. The values t_{B1}, t_{B2}, and t_{B3} are measured in generations before the present, but we must map them to times $\tau(t)$ measured in units of $2N$ generations before the present:

$$\tau(t) = \begin{cases} (t - t_{B1})/(2Na) + (t_{B1} - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t > t_{B1} \\ (t - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t_{B1} < t < t_{B2} \\ t/(2N) & \text{if } t < t_{B2} \end{cases}$$

In addition to scaling t, we must scale μ and ρ, since each contains a factor of N.

When we make these modifications, equation (3) becomes

$$\begin{aligned} p_L(\text{IBD}) &= \int_{\tau=0}^{\tau(t_{B2})} e^{-\tau\cdot(1+L(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B2})}^{\tau(t_{B1})} e^{-\tau\cdot(1+Lf(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B1})}^{\infty} e^{-\tau\cdot(1+La(\mu+\rho))}d\tau \\ &= \frac{1 - e^{-\tau(t_{B2})(1+L(\mu+\rho))}}{1 + L(\mu+\rho)} + \frac{e^{-\tau(t_{B2})(1+Lf(\mu+\rho))} - e^{-\tau(t_{B1})(1+Lf(\mu+\rho))}}{1 + Lf(\mu+\rho)} \\ &\quad + \frac{e^{-\tau(t_{B1})(1+La(\mu+\rho))}}{1 + La(\mu+\rho)}. \end{aligned}$$

In the same way, we can correct $A_1(L), \ldots, A_L(L)$ for the bottleneck by replacing

$$C_i = \frac{\rho}{i(\mu+\rho)(1 + i(\mu+\rho))}$$

with

$$\begin{aligned} C_i &= \frac{\rho}{i(\mu+\rho)} \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

In terms of these corrected $A_i(L)$, we deduce that

$$\begin{aligned} p_L(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

3 The age distribution of maximal IBD segments

Our calculations, along with those in earlier papers, make it clear that IBD segment length is inversely related to age. In [17], Hayes, et al. go as far

as to draw a one-to-one correspondence between the abundance of maximal c-centimorgan IBD segments and the effective size of the population $1/(1+4c)$ generations ago. However, we show here that the mean coalescence time of an L-base IBD tract ($O(1/L)$) is much less than its standard deviation ($O(1/\sqrt{L})$), implying that the segments coalescing at time t have a significant length spread, particularly when t is very ancient. This complicates the effect of population size changes on the distribution of ROH length, particularly for shorter ROHs. While Hayes, et al. studied the distribution of ROHs that were 10^6 to 10^7 base pairs long and found their assumption useful at that length scale, we find that the relationship between effective population size and ROH length is more complicated for shorter ROHs, as we will see corroborated by data in Section 7 (Figures 14, 15, and 16).

As we saw in Section 2, the probability of an L-base ROH being IBD is

$$\int_{t=0}^{\infty} e^{-t(1+L\rho)}dt,$$

while the proability that it will be maximally IBD (i.e. not contained in a larger IBD segment) is

$$\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt.$$

We can use this to calculate a joint distribution between IBD segment length and coalescence time:

$$p_L(t|\text{IBD}) = \frac{e^{-t(1+L\rho)}(1-e^{-t\rho})^2}{\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt}. \quad (8)$$

We compute that

$$\begin{aligned}\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt &= \frac{1}{1+L\rho} - \frac{2}{1+(L+1)\rho} + \frac{1}{1+(L+2)\rho} \\ &= \frac{2\rho^2}{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)},\end{aligned}$$

such that

$$p_L(t|\text{IBD}) = \frac{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}{2\rho^2}e^{-t(1+L\rho)}(1-e^{-t\rho})^2. \quad (9)$$

Similarly, we can compute the expected t value $E_t(L)$, measured, as always, in units of $2N$ generations:

$$\begin{aligned} E_t(L) &= \int_{t=0}^{\infty} t p_L(t|\text{IBD}) dt \\ &= \frac{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}{\rho^2} \\ &\quad \cdot \left(\frac{1}{(1+L\rho)^2} - \frac{2}{1+(L+1)\rho)^2} + \frac{1}{(1+(L+2)\rho)^2} \right) \\ &= \frac{3L^2\rho^2 + 6L\rho^2 + 6L\rho + 2\rho^2 + 6\rho + 3}{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}. \end{aligned}$$

This differs from $1/(1+L\rho)$, the value given by Hayes, et al., because they don't distinguish between maximal and non-maximal IBD.

We go on to compute the variance

$$\begin{aligned} E_{t^2}(L) - E_t(L)^2 &= \int_{t=0}^{\infty} t^2 p_L(t|\text{IBD}) dt - \left(\int_{t=0}^{\infty} t p_L(t|\text{IBD}) dt \right)^2 \\ &= \frac{12(L^3 + \rho^2 L^2 + 2\rho L + \rho^2 + \rho + 1)(3\rho^2 L^2 + 6\rho^2 L + 10\rho^2 + 6\rho + 3)}{(1+\rho L)^2(1+\rho(L+1))^2(1+(L+2))^2} \\ &\quad - \frac{12(\rho L + 1)(\rho L + \rho + 1)}{(1+\rho(L+1))^2(1+\rho(L+2))^2} \\ &\quad - \frac{(3L^2\rho^2 + 6L\rho^2 + 6L\rho + 2\rho^2 + 6\rho + 3)^2}{(1+L\rho)^2(1+(L+1)\rho)^2(1+(L+2)\rho)^2}. \end{aligned}$$

Looking at the leading terms, we note that

$$E_t(L) \approx \frac{3}{\rho L} \ll \sqrt{E_{t^2}(L) - E_t(L)^2} \approx \frac{6}{\rho^2 \sqrt{L}},$$

meaning that the standard deviation of $E_t(L)$ is much greater than its mean.

Figure 3 shows the length distribution of IBS segments that coalesce $0.2N$ generations ago, while Figure 4 plots the length distribution of segments that coalesce $0.3N$ generations ago. Even if IBD were the same as IBS and segments coalesced at only these two times, it would not be straightforward to look at a sum of plots like this and quantify an excess of one type of segment. Hayes, et al. track recent population growth by assuming that a dearth of L-base IBD segments means a larger population at time $1/(1+L\rho)$, but it would seem that this approach must be modified for shorter L where the length and coalescence

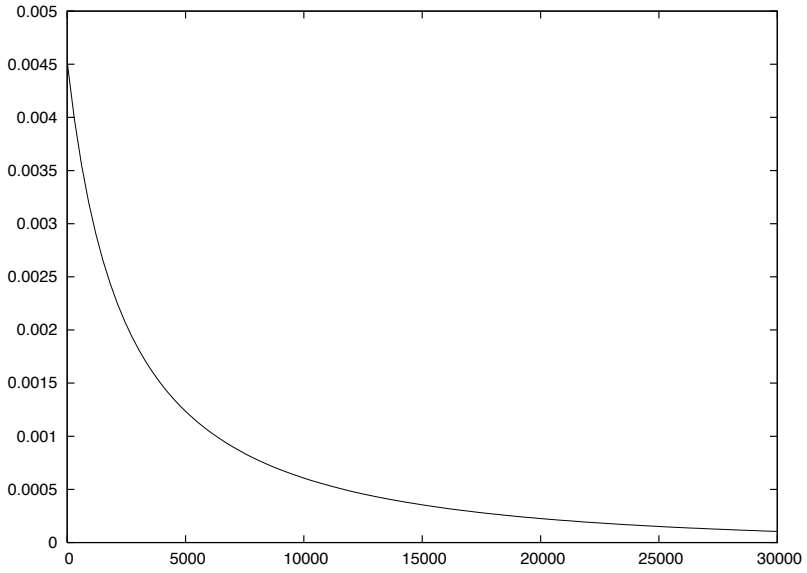

Figure 3: This plot shows the length spread of IBD segments that coalesce $0.2N$ generations ago. Comparing this to Figure 4, we see that it will be difficult to tell these segments apart from segments that coalesced $0.3N$ generations ago.

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 $\text{IBS}(L_1, L_2) \geq 1$ 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

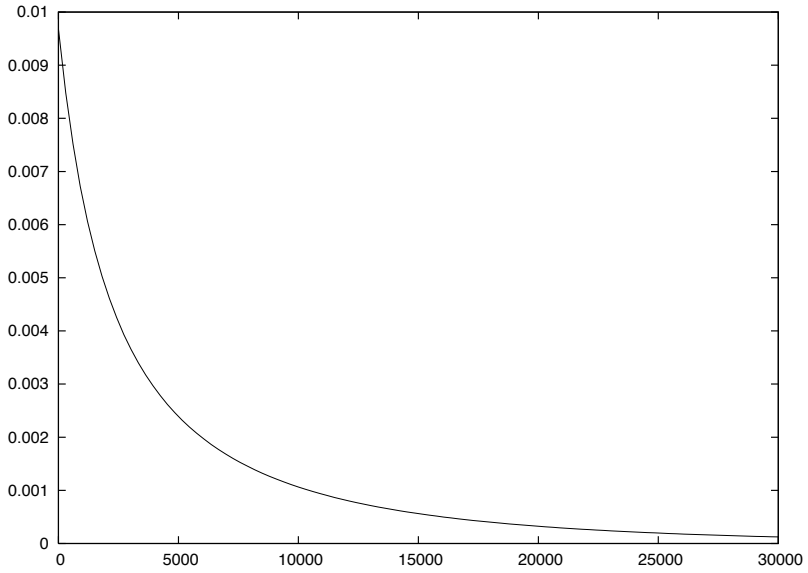

Figure 4: The length spread of IBD segments that coalesce $0.3N$ generations ago is different from the spread of segments that coalesce $0.2N$ generations ago (Figure 3), but overlaps enough that it would take a bit of work to learn about population history from a sum of density plots like these.

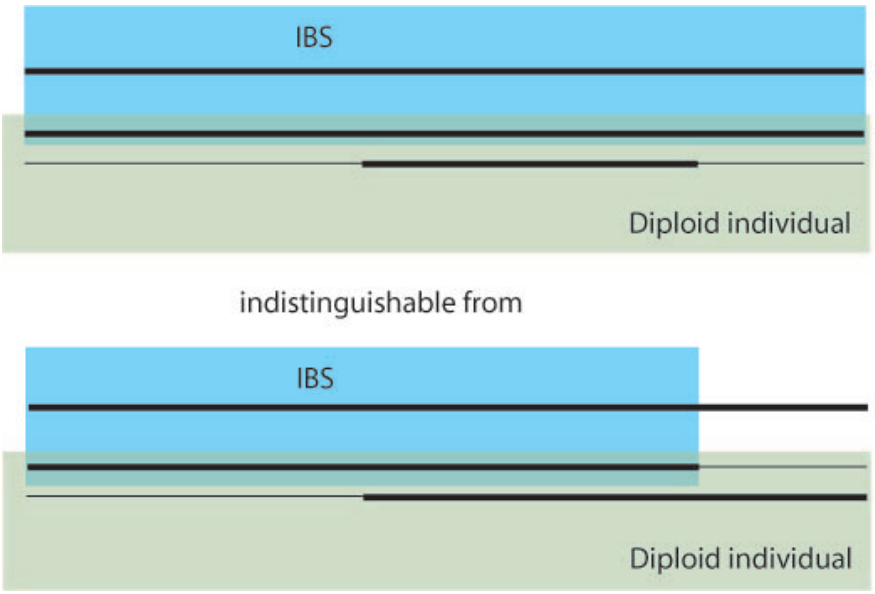

Figure 5: **IBS between phased haplotypes vs. IBS≥1** Here, the two chromosomes of a diploid individual are aligned to a reference haplotype. The diploid DNA is drawn in bold where it matches the reference haplotype IBS. Both the top and the bottom alignment have the property IBS≥ 1, where at least one 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 $\text{IBS}(r, d) \geq 1$ 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 $\text{IBS}(r, d) \geq 1$, 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).

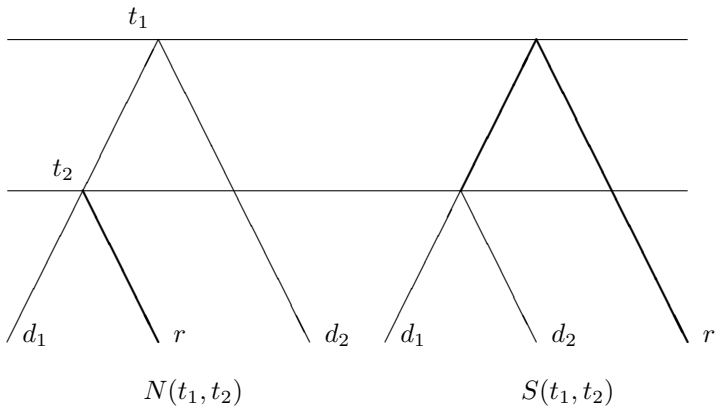

Figure 6: Natural and skew tree topologies. Boldface branches are drawn where mutations cannot occur without rendering the alignment non-IBS.

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 $(r(d_1 d_2))$. 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:

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

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:

$$\begin{aligned} P_S(\text{IBD\& IBS}) &= \int_{t_2=0}^{\infty}\int_{t_1=t_2}^{\infty} e^{-t_2(2-L(\rho+\mu))-t_1(1+2L(\rho+\mu))}dt_1dt_2 \quad (12)\\ &= \frac{1}{(1+2L(\rho+\mu))(3+L(\rho+\mu))} \quad (13)\end{aligned}$$

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:

$$\begin{aligned}
P_L(\text{IBS},t)dt &= p_{L-1}(\text{IBS},t)dt \cdot e^{-t(\mu+2\rho)} \\
&+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS},t_0)\cdot e^{-\mu t}\left(2\rho e^{-2\rho t_r}\cdot\frac{2}{3}\cdot 3e^{-3(t-t_r)}dt\right. \\
&\quad \left. +2\rho e^{-2\rho t_r}\int_{t_2=t_r}^{t}\frac{1}{3}\cdot 3e^{-3(t_2-t_r)}\cdot e^{-(t-t_2)}dtdt_2\right)dt_r dt_0 \\
&+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS},t_0)\cdot e^{-\mu t}\left(\frac{2}{3}\cdot 3\rho e^{-3\rho t_r}\cdot\frac{1}{2}\cdot 2e^{-2(t-t_r)}+\right. \\
&\quad \left. +\frac{1}{3}\cdot 3\rho e^{-3\rho t_r}\cdot 2e^{-2(t-t_r)}dt\right)dt_r dt_0 \\
&= p_{L-1}(\text{IBS},t)dt\cdot e^{-t(\mu+2\rho)} \\
&+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS},t_0)\left(3\rho e^{-t(3+\mu)+t_r(3-2\rho)}\right. \\
&\quad \left. +\rho e^{-t(1+\mu)+t_r(1-2\rho)}\right)dt_r dt_0 dt \\
&+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS},t_0)\cdot 4\rho e^{-t(2+\mu)+t_r(3-2\rho)}dt_r dt_0 dt.
\end{aligned}$$

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

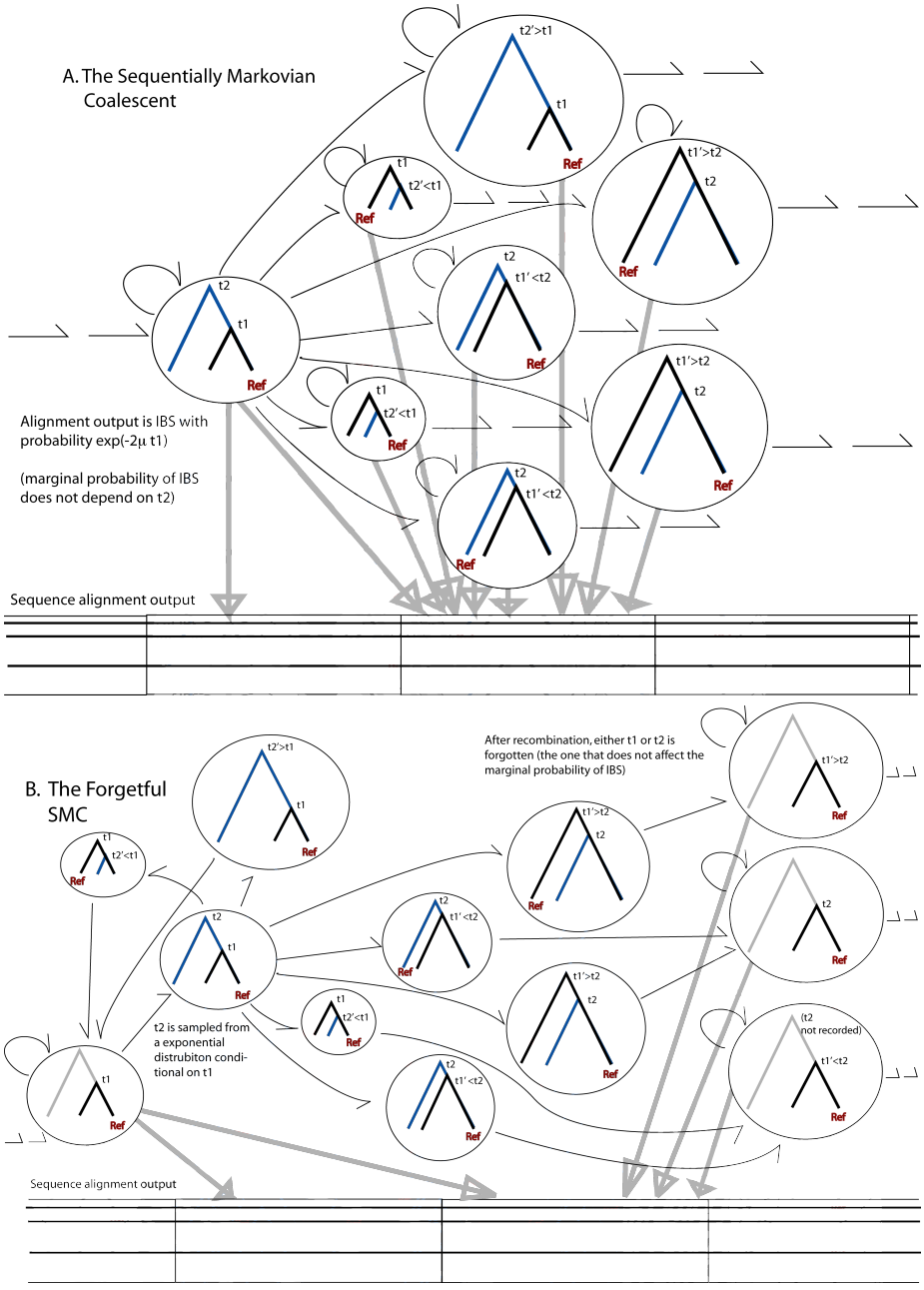

Figure 7: Hidden Markov model flow diagrams of the SMC and our forgetful approximation of the SMC. The position of the reference sequence is labeled to mark each genealogy as natural or skew. Each output extends the alignment by a triplet of bases (including one labeled reference base) that is either 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

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

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

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

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

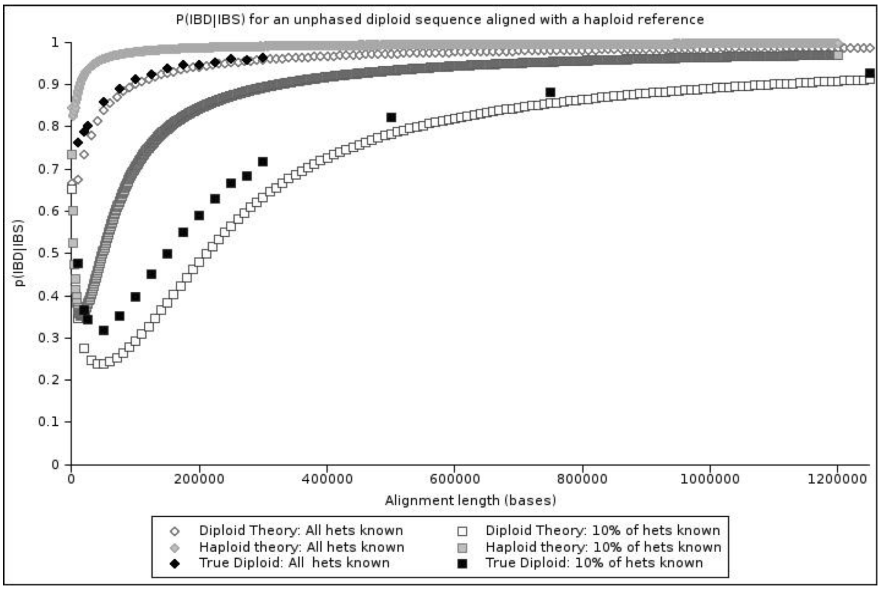

Figure 8: This plot compares our diploid results to the values that we obtained from MS simulations, which assume the full non-Markovian coalescent with recombination. Given a 10^5-base diploid sequence that is IBS ≥ 1 with a refernce, it is 90.1% likely to contain a haplotype that is IBD with the reference. 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

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

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

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

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

where $e^{-tL(2\mu+2\rho)} = p_L(\text{IBS\&IBD}|t)/e^{-t}$ 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:

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

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

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

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

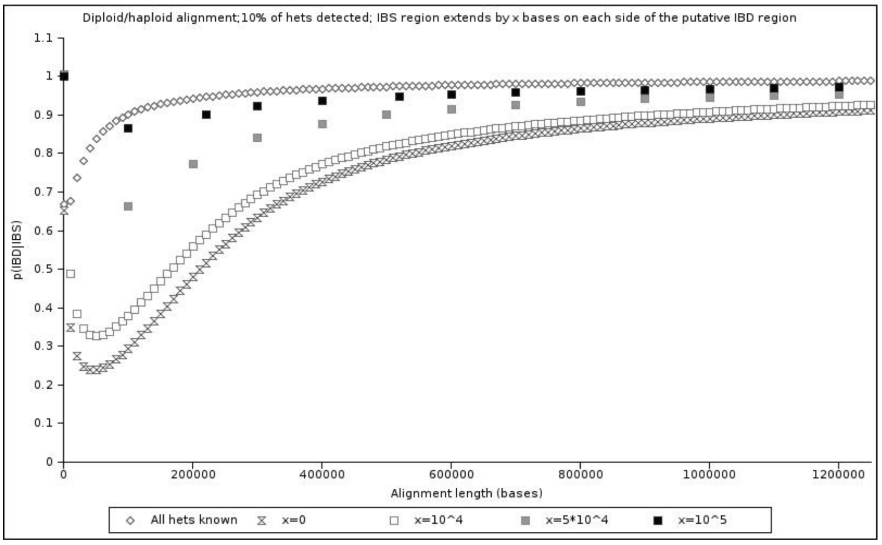

Figure 9: Not much imputation accuracy is gained by chopping 10^4 bases off each end of thinned IBS diploid alignment. However, a substantial amount of accuracy is gained by chopping off 10^5 bases; the resulting confidence of IBD is nearly as great as the confidence given complete sequence data.

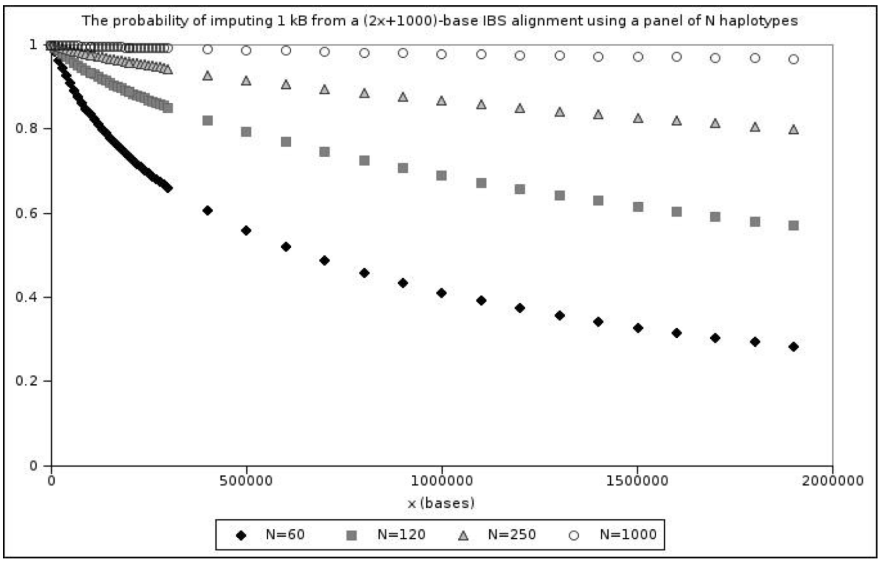

Figure 10: Here, we plot the expected size of the longest IBS alignment between a test sequence and a panel of N reference haplotypes that is centered at a small region to be imputed.

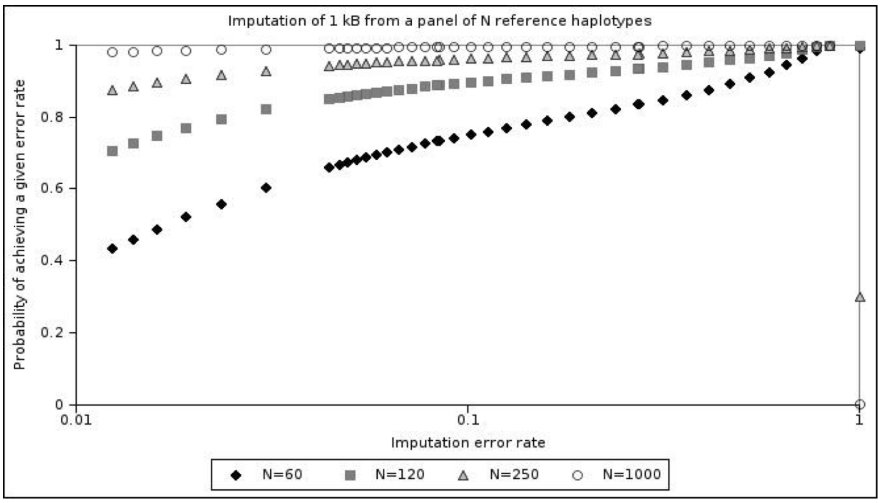

Figure 11: Using the function $I(L, x)_{0.10}$, we transform Figure 10 into a distribution of expected imputation accuracies. We plot the probability of being able to impute either one of the test haplotypes.

methods published prior to 2008 is that they assume the input markers are in linkage equilibrium with each other; Browning's method is the first to account for linkage disequilibrium by conditioning on the frequencies of the input haplotypes [9]. Although the assumption of linkage equilibrium is certainly a problem for dense marker data, we find another assumption in all previously published methods that could pose additional problems: they assume that linkage is an either/or phenomenon; that if two sites are not IBD, then their coalescence times are independently distributed [9, 17, 46, 51, 55].

Our method makes it straightforward to pinpoint where that assumption breaks down; to this end, we calculate $p_L(\text{IBS})$ given the assumption that neighboring IBD segments have uncorrelated coalescence times, calling this modified probability function $q_L(\text{IBS})$. We will see that $q_L(\text{IBS})$ agrees with $p_L(\text{IBS})$ for short alignments, but not for L values that are long enough that $p_{L+1}(\text{IBS}) > p_L(\text{IBS})$. Rather than depending on mutation rate or marker density, the threshold L value turns out to be a function of the number of markers between the ends of the alignment. For the standard parameters $N = 10000, \mu = 1.0 \times 10^{-3}$, and $\rho = 4.0 \times 10^{-4}$, the independent coalescence time assumption breaks down for $L \gg 1000$. The assumption stays valid for longer when only a tenth of the hets are observed, but still breaks down for $L \gg 10,000$ (see Figure 12).

As before, we can find coefficients $B_1(L), \ldots, B_L(L)$ for which

$$q_L(\text{IBS}|t) = \sum_{i=1}^{L} B_i(L) e^{-t(1+i(\mu+\rho))} dt$$

using inductive integration, defining $q_L(\text{IBS}|t)$ such that

$$q_L(\text{IBS}|t) = q_{L-1}(\text{IBS}|t) \cdot e^{-t\rho} + \int_{t_0=0}^{\infty} q_{L-1}(\text{IBS}|t_0)(1 - e^{-t_0\rho}) e^{-t(1+\mu)} dt_0.$$

We depart from the computation in Section 2 by eliminating the parameter t_r, the time at which the recombination occurs, which is the parameter that introduces dependence between the coalescence times of neighboring IBD segments.

Since

$$\int_{t_0=0}^{\infty} e^{-t_0(1+i(\mu+\rho))}(1-e^{-t_0\rho})e^{-t(1+\mu+\rho)}dt_0 = \frac{\rho e^{-t(1+\mu+\rho)}}{(1+i(\mu+\rho))(1+\rho+i(\mu+\rho))},$$

we can let

$$K_i = \frac{\rho}{(1+i(\mu+\rho))(1+\rho+i(\mu+\rho))}$$

and conclude that

$$B_i(L) = B_{i+1}(L+1)$$

for all $i > 1$, whereas

$$B_1(L) = \sum_{i=1}^{L-1} K_i B_i(L-1).$$

7 Empirical validation using genome sequence data

To measure the accuracy of our predicted $p_L(\text{IBS})$ values, we found the lengths of all maximal ROHs in the eleven human genome sequences referenced in Table 13. The bases were re-called in a consistent fashion with the intent to make the quality good enough for population genetic analysis; out of a total of 33,686,389,482 base pairs, 9,743,948,741 (28.9%) were marked unreliable due to unreliable read mapping, proximity to indels, or other other attributes that made them suspect (see Base Calling Methods appendix), and we deleted these bases before proceeding. Our call sets for all sequences are available for download at ftp://ftp.sanger.ac.uk/pub/rd/humanSequences.

In addition to counting the number $N_{\text{ROH}}(L)$ of ROHs in each genome that are between $(L-1000)$ and L bases long (for L divisibly by 1000), we counted the number $N_{\text{ROH}}(L)_{0.10}$ of L-base regions that appear homozygous when we detect a tenth of all hets. Specifically, we generated *thinned ROHs* whose endpoints are the mutations with positions congruent to zero mod ten relative to the 5' end of the chromosome, referring to these endpoints as *observed hets* as opposed

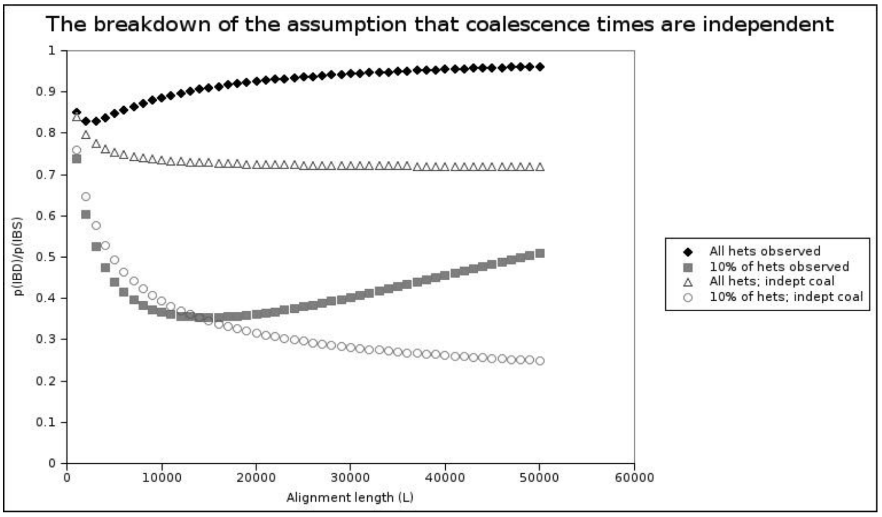

Figure 12: The two solid plots record the exact probability of IBD given IBS (assuming the SMC), the black plot with IBS required at every base in the alignment, and the grey plot where markers allow for only 10% of mutations to be detected. The empty triangles and empty circles record how that probability changes when we disregard linkage between non-IBD markers. The two curves almost never agree when complete sequences are used, in concordance with the fact that earlier methods do not claim to be accurate for such dense marker data.

Sequence Names	Origins
COLO-829-BL	Northern/Western European Ancestry [5]
NA12878, NA12891, NA12892	Northern/Western European Ancestry [1]
NA18507	Yoruba, Nigeria [7]
NA18506, NA18508, NA19239, NA19240	Yoruba, Nigeria (unpublished)
SJK	Korean [4]
YH	Chinese [60]

Figure 13: The eleven genomes used in our analysis

to *hidden hets*. We predict that

$$\frac{N_{\mathrm{ROH}}(L)}{N_{\mathrm{ROH}}(L)_{0.10}} = \frac{10 p_{L\max}(\mathrm{IBS})}{p_{L\max}(\mathrm{IBS})_{0.10}}, \tag{23}$$

adding the factor of 10 to account for the fact that there are ten times as many true ROHs as thinned ROHs (most of the excess ones being short).

Even though we take care to use genome data with a very low error rate, false positive hets (on the order of 1 per 10^5 bases) will present a significant problem for our analysis. We will be estimating the abundance of ROHs up to 10^7 bases long, and there is an overwhelming chance that their homozygosity will be broken up by false positives.

To correct for the breakup of ROHs by false positives, we estimate the false positive frequency f and multiply the measured value of $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ by $(1-f/10)^L/(1-f)^L$, reasoning that $(1-f)^L$ is the probability that an L-base ROHs will be broken up by a false positive het. We choose $f = 1.5 \times 10^{-5}$ because $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ tends toward $(1-f)^L/(1-f/10)^L$ in each genome as L gets large, while the ratio of thinned to true ROHs should tend toward 1.

The eleven plots of $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ versus L cluster clearly by ethnicity (see Figures 14, 15, 16), and we account for the differences by finding effective population size histories that fit $10p_{L\max}(\mathrm{IBS})/p_{L\max}(\mathrm{IBS})_{0.10}$ well in the data from each ethnic group. We also experiment with varying the mutation rate μ, motivated by the fact that the 1000 Genomes consortium recently

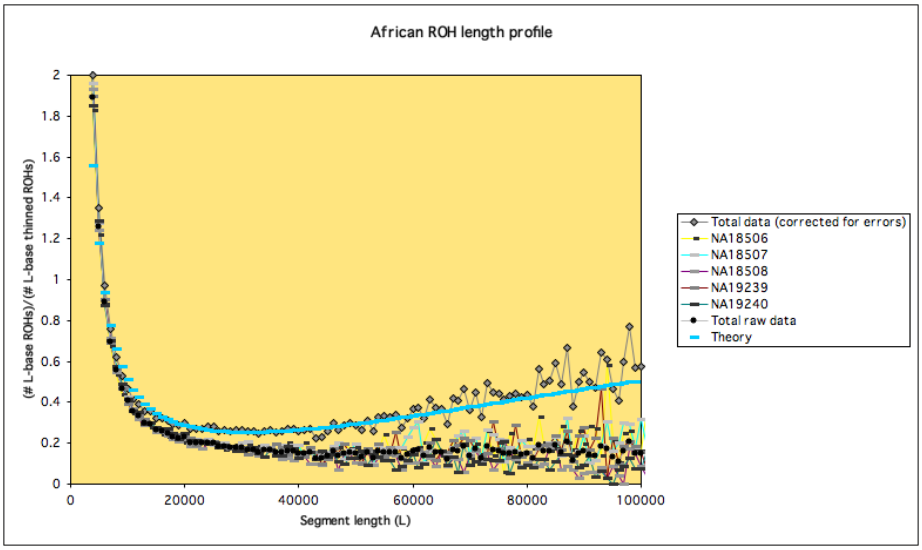

Figure 14: **A. Regions of homozygosity in African genome data** Here, we separately plot $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ for each of five African genomes, then average this function across the genomes and correct it for 1.5×10^{-5} false positives per base. In blue is the theory plot $\frac{10p_{L\max}(\mathrm{IBS})}{p_{L\max}(\mathrm{IBS})_{0.10}}$ for a population of constant effective size $N = 14,000$ and a mutation rate of $m = 1.6 \times 10^{-8}$ per base per generation (one of many histories that minimize the sum of square distances from the data points to the predicted curve).

estimated μ to be 1×10^{-8} per base per generation [1] rather than 2.5×10^{-8}.

The measured $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ ratios behave noisily for $L > 100,000$, likely because there are few such ROHs in the genome and each one is more likely than a short ROH to include recombination hotspots or other sites where the theory in this paper breaks down. Therefore, we define the best fit population history to be the one that minimizes the sum of squares distance from the predicted $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ values to the measured $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ values, the sum taken over L ranging from 10,000 to 100,000.

Let $T_{\mu,H}(L)$ denote the theory plot of $10p_{L\max}(\mathrm{IBS})/p_{L\max}(\mathrm{IBS})_{0.10}$ that is obtained a function of the mutation rate μ and the piecewise-constant popula-

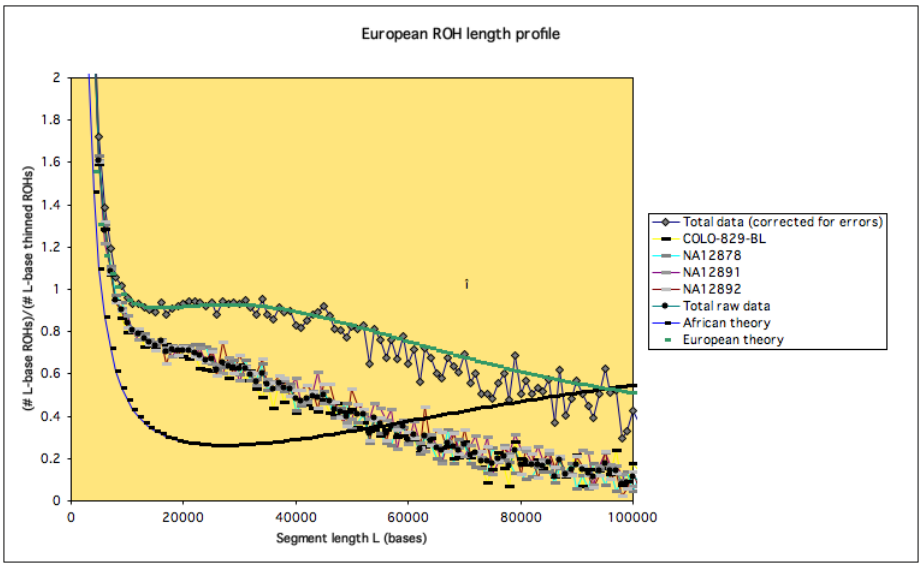

Figure 15: **B. Regions of homozygosity in European genome data** In green is the following single-bottleneck history, with mutation rate $m = 2.5 \times 10^{-8}$: $N = 11,900$, time ranging from 0 to 1240 generations ago (g.a.); $N = 4,530$, $1,240 - 1,770$ g.a.; $N = 15,000 \geq 1,770$ g.a. The African constant population size theory is included in black, for reference.

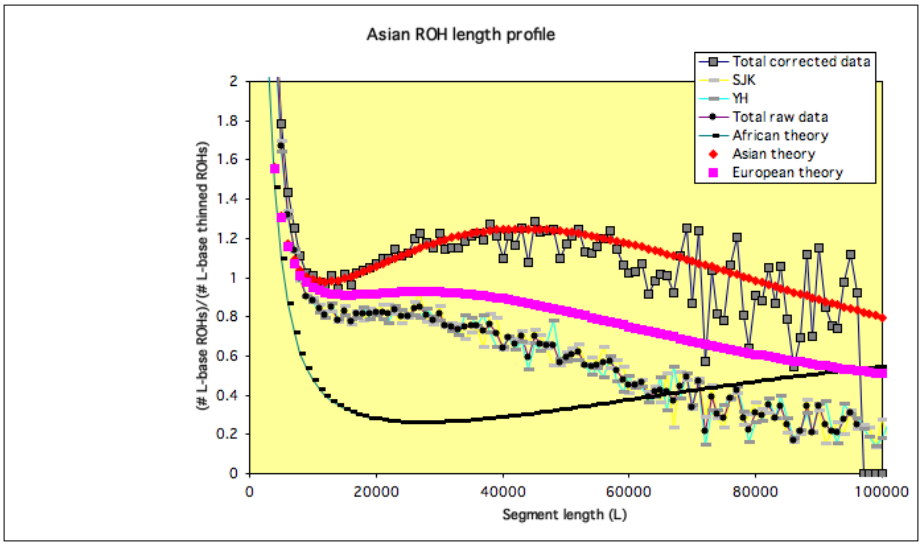

Figure 16: **C. Regions of homozygosity in Asian genome data** The single-bottleneck history shown in red, again assuming $m = 2.5 \times 10^{-8}$, is the following: $N = 8,670$, $0 - 1,380$ g.a.; $N = 1790$, $1,380 - 1,530$ g.a.; $N = 15,000 \geq 1,530$ g.a.. African and European theory plots are included for reference.

tion history H. Likewise, let $D_G(L)$ denote the set of data points $\{N_{\mathrm{ROH}}(10^3 \cdot i)/N_{\mathrm{ROH}}(10^3 \cdot i)_{0.10}\}_{i=1}^{10^3}$ that is obtained by counting all of the ROHs and thinned ROHs in some set of genomes G and correcting for 1.5×10^{-5} false positive hets per base. We measure the goodness of fit between the parameters (μ, H) and the data set G by calculating the sum of squares fit

$$SS(\mu, H, G, L) = \sum_{i=1}^{900} \left(T_{\mu,H}(10^3 \cdot (10+i)) - D_G(10^3 \cdot (10+i))\right)^2.$$

It remains to define a threshold for $SS(\mu, H, G, L)$ below which (μ, H) is deemed a good fit for G. Since $T_{\mu,H}(L)$ is not a straight line, we cannot perform a goodness-of-fit linear regression. We find it logical, instead, to define a threshold that depends on the noisiness of the curve $D_G(L)$, letting

$$N(G, L) = \sum_{i=1}^{(L-1)/1000} \left(D_G(10^3 \cdot (10+i+1)) - D_G(10^3 \cdot (10+i))\right)^2$$

denote the sum of squared distances between adjacent points of $D_G(L)$. If $T_{\mu,H}(L) = \frac{1}{2}(D_G(L) - D_G(L+1))$, making $T_{\mu,H}(L)$ a smoothed version of the data set $D_G(L)$, then

$$SS(\mu, H, G, L) = \frac{1}{4}N(G, L),$$

In each data plot $D_G(L)$, the left portion of the graph is much less noisy than the right portion and therefore provides more information about the mutation rate and population history. In the European genomes, for example, there is so little noise in the data set $D(G)|_{L<34000}$ that

$$\frac{1}{4}N(G_{\mathrm{European}}, 34000) < 0.0094,$$

while

$$\frac{1}{4}N(G_{\mathrm{European}}, 90000) > 0.26.$$

For each of the genome groups $G_{\mathrm{African}}, G_{\mathrm{European}}$, and G_{Asian}, we define L_{short} to be the largest L satisfying $\frac{1}{4}N(G, L-1) < 0.01$ and define L_{long} to be the longest $L \geq 1000$ satisfying $\frac{1}{4}N(G, L-1) < 0.5$ (specific values of L_{short} and L_{long} are recorded in Table 17). We then say that (μ, H) is a good fit for G if

$$SS(\mu, H, G, L_{\mathrm{short}}) < 0.01$$

Ethnicity	L_{short}	L_{long}
African	53000	90000
European	35000	89000
Asian	21000	60000

Figure 17: Thresholds for low noise ($N(G, L_{\text{short}} - 1) < 0.01$) and medium noise ($N(G, L_{\text{long}} - 1) < 0.2$) in the $D_G(L)$ ROH data sets. Our Asian data set, which contains half as many genomes as the others, appears commensurately noisier.

and

$$SS(\mu, H, G, L_{\text{long}}) < 0.5.$$

We searched for good parameter fits using a Monte Carlo Markov chain approach, beginning with a search of constant population size histories. As expected, the Africans are the only group for which we find good constant population size histories. Such histories fall within a narrow parameter space, namely $13,000 \leq N \leq 15,000$ and $1.55 \times 10^{-8} < m < 1.7 \times 10^{-8}$.

When we allow for a single population expansion or contraction, we find a large variety of histories that fit the African data well, though we still find no fits for the European or Asian data. These good African histories are all expansions when $m = 2.5 \times 10^{-8}$, all contractions when $m = 1 \times 10^{-8}$, and close to constant for $m = 1.75 \times 10^{-8}$, with a population size change in either the very recent past or the very distant past (see Figures 18, 19, and 20).

To find good theory fits for the European and Asian data, it was necessary to invoke a population bottleneck. To speed up our MCMC search, we fixed the mutation rate $m = 2.5 \times 10^{-8}$ and the ancestral population size $N_3 = 15,000$. This left two variable time parameters and two variable size parameters, enough to generate many optimal histories to fit both sets of non-African data (see Figures 21 and 22). In both sets of good histories, recent good-fit bottlenecks are shallower than ancient good-fit bottlenecks. The modern effective population size is lower, on average, in the Asian histories. This fits with the fact that the

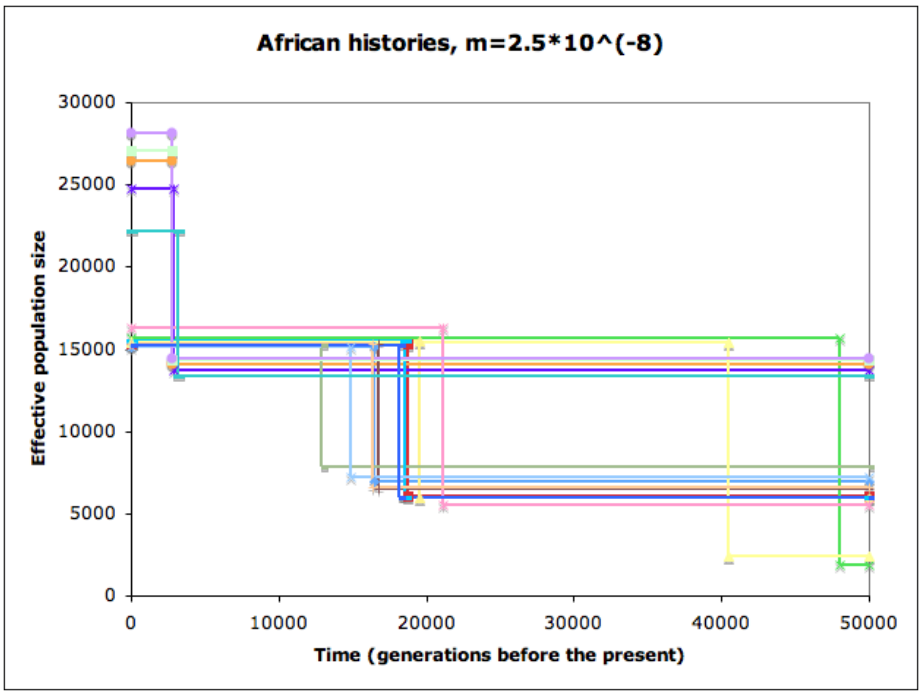

Figure 18: Some population histories satisfying our good fit criterion for African genome data assuming $m = 2.5 \times 10^{-8}$ mutations per base per generation

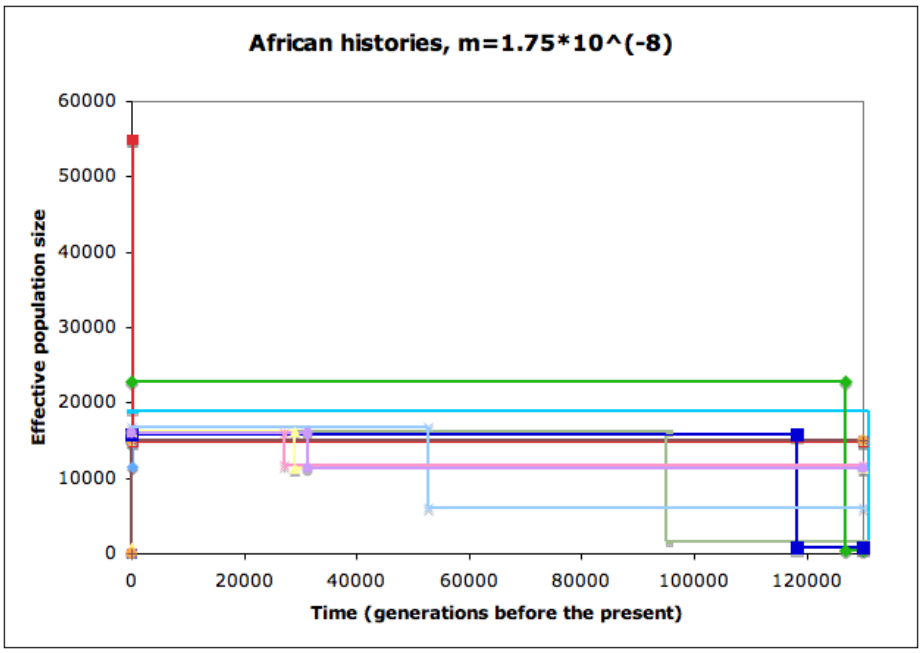

Figure 19: More good fit YRI histories; mutation rate $m = 1.75 \times 10^{-8}$ per base per generation

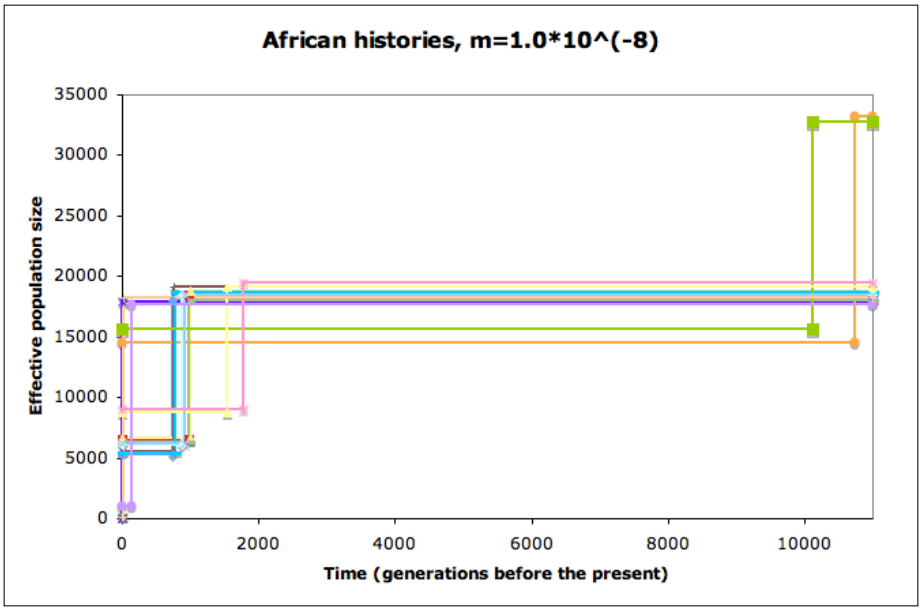

Figure 20: More good fit YRI histories, mutation rate $m = 1.0 \times 10^{-8}$ per base per generation

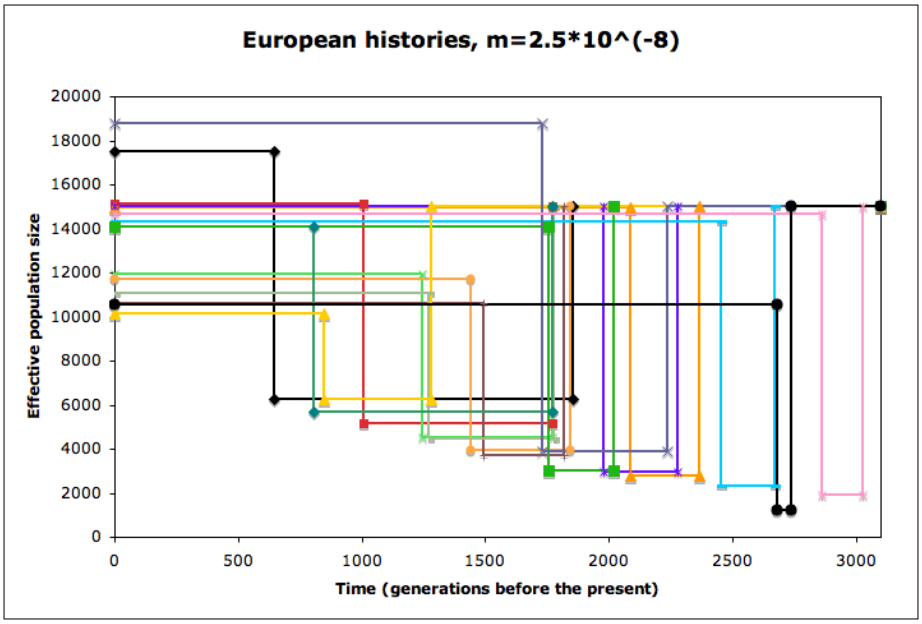

Figure 21: A population bottleneck is required to fit the CEU ROH data–either a shallow, recent bottleneck or a deeper, more ancient bottleneck.

Asian HapMap allele frequency spectrum shows more evidence of genetic drift than the European HapMap allele frequency spectrum [30].

It remains an open problem to mathematically describe the set of histories that fit the distribution of ROHs in each ethnic group. However, in showing that such histories exist, we achieve our aim of predicting the length distribution of ROHs in real genome data and validating the theory that we use to compute IBD probabilities.

8 Discussion

In this paper, we attempt to very precisely model patterns of linkage disequilibrium in genetic data, capturing its decay over long regions of the genome instead of assuming that certain blocks or pairs of loci assort independently. Rather than adding a new LD model to the myriad that exist already, we work

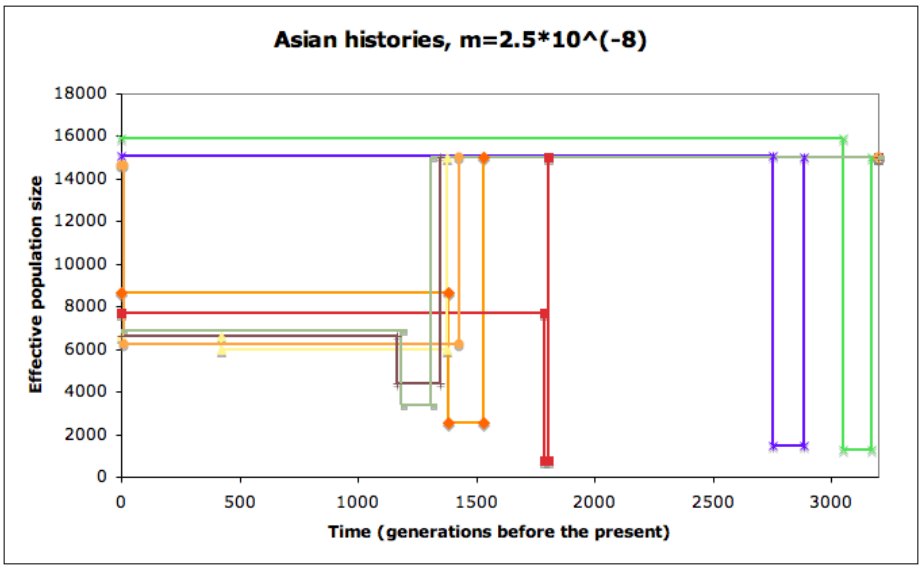

Figure 22: A bottleneck is required to fit Asian ROH data. These good fit histories have smaller recent effective population sizes than the histories that fit the CEU data (see Figure 21).

as closely as possible to the neutral coalescent with recombination, deriving results within a theoretical framework whose advantages and limits are well understood already. Our claims about patterns of homozygosity in the genome make no assumptions that weren't published as part of the sequentially Markovian coalescent (SMC) model [43]; while we do not account for recombination hotspots or other point irregularities, the results should be quite accurate across the genome as a whole, as seen empirically in Section 7.

While it was necessary to relax the SMC somewhat to model LD in diploid alignments with uncertain phasing, our model approaches the coalescent in the limit of high LD where phasing and imputation attain high accuracy. While it is possible to check our computations with a coalescent simulator such as MS, our method makes it much quicker to compute a length spread of IBS probabilities; without it, each data point on each of our plots would have to be obtained from a separate coalescent simulation with memory and storage requirements that get quite large for long IBS alignments. Efficiency was key when we had to compare many population histories to find a matches for the African, European, and Asian homozygosity data, and also when we considered the best alignment between a test sequence and some member of a reference panel.

The accuracy values that we compute for various imputation panels assume an idealized population of effective size 10,000, closer to the truth for Africans than for non-Africans. However, it is possible to condition instead on any piecewise-constant population size history, as is briefly explored in the last section of the paper, if e.g. the goal is to impute only Europeans. We concentrate less on evaluating the capabilities of an existing panel than on building a framework for making informed decisions about the design of future panels.

Our results do posit lower bounds on the amount of variation that should not be imputable using HapMap, as we estimate that a 120-reference panel has only a 70% chance of imputing a one kilobase mini-haplotype with 99% accuracy and an 80% chance of imputing it with 90% accuracy. Accuracy is likely higher when the aim is to impute only common SNPs, but it can only be higher insofar as those common SNPs fail to tag the rarer SNPs around them. Our rough

estimate of HapMap variation coverage is close to that obtained by Bhangale, et al., who resequenced 1.6 megabases in each of the HapMap individuals and reported that only 60-80% of the SNPs they found were within $r^2 > 0.8$ of a tag SNP [8].

Besides showing that it is feasible to study imputation with coalescent theory, we hope that the tricks and shortcuts we've developed might help make coalescent theory more applicable to other hard problems. Important as imputation is, it is not the only setting where understanding IBS could be useful. As seen in Section 7, real hets are distributed differently from false positive hets, such that hets scattered in regions of high homozygosity are very likely to be false positives. It is easy to compute the false positive probability of a het that is L bases away from the nearest het, and it might be useful to incorporate this result into base calling software. A long stretch of near-homozygosity provides a lot of evidence that the region is IBD (see Figure 8 for a plot of p(IBS) versus length), and there is less than a 1.5% chance that an IBD region will contain seven or more mismatches (the number of mismatches being Poisson-distributed with a mean of $\mu/\rho = 2.5$). When the sequencing error rate is 10^{-5} and a 10^7-base region contains about 100 scattered hets, the likely truth is that the region is IBD and most of those hets are errors.

A challenge for the future will be to address the effect of recent exponential population growth. Only the frequency of the longest IBS regions should be affected, but this will be enough to make a 1000-haplotype reference panel less perfect than it appears in figures 10 and 11; Ionita-Laza, et al. predict that more than 3,000 individuals will be needed to find all variants with frequency 0.1% based on allele frequency data from the outbred CEU and YRI sequence data [26]. One way around this issue would be to conduct GWASs in moderately isolated populations where exponential growth has been very recent and 1000 references do constitute a perfect imputation panel. Family-based linkage studies have been successful for some time at discovering functional variants that affect few people but shed valuable light on disease mechanisms, and our results suggest that IBD mapping need not be limited to groups as small as families. A

project the size of 1000 Genomes [1] should make it possible to essentially know the sequence of every individual in a small population like Iceland, which would, in principle, make it possible to test for functionality all across the genome, including at a large pool of rare and untaggable SNPs that are invisible even in GWASs conducted with the help of imputation from HapMap.

A Base-calling methods

The sequences used to validate our method were re-called with the hope of reducing the frequency of errors. The call sets are available for download at ftp://ftp.sanger.ac.uk/pub/rd/humanSequences.

Raw Illumina read data were obtained from NCBI's sequence read archive (see Table 13) and EMBL's European read archive. They were aligned by BWA (0.5.5), using human reference genome build 36, which masks pseudoautosomal regions on Y by including unassembled contigs and the Epstein-Barr virus genome (AC:NC_007605). Low quality bases were trimmed from the 3'-ends of Illumina short reads by applying BWA option '-q 15.' Because SJK base qualities were overestimated, they were recalibrated using the Genome Analysis Toolkit after discarding SNPs known from dbSNP-129. Default BWA-SW algorithms were used for capillary reads.

The 'pileup' command of the *samtools* software package was used to create each autosome's diploid consensus. The consensus was then filtered, the following kinds of bases being marked low-confidence calls:

1. Read depth is more than twice or less than half of the depth estimated at loci genotyped in HapMap3
2. Reads covering the locus have root mean square mapping quality below 25
3. There is a predicted short indel less than 10 base pairs away
4. Inferred consensus quality is below 20 (Illumina data) or 10 (capillary

data)

5. Out of the 35 reference sequence 35-mers that overlap with the site, fewer than 18 can be mapped elsewhere with at most one mismatch

References

[1] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature* (2010) 1061–1073.

[2] C.A. Anderson, D. Brocklebank, and A.P. Morris. "A comparison of reference panels for imputation in genome-wide association studies." *The European Journal of Human Genetics*, to appear.

[3] C.A. Anderson, F.H. Pettersson, J.C. Barrett, J.J. Zhuang, J. Ragoussis, L.R. Cardon, and A.P. Morris. "Evaluation the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms." *The American Journal of Human Genetics* (2008) 112–119.

[4] S.-M. Ahn, T.-H. Kim, S. Lee, et al. "The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group." *Genome Research* 19(9) (2009) 1622–1629.

[5] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dougan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. "THe COSMIC (Catalogue of Somatic Mutations in Cancer) database and website." *British Journal of Cancer* 91 (2004) 355–358.

[6] J.C. Barrett and L.R. Cardon. "Evaluating the coverage of genome-wide association studies." *Nature Genetics* 38(6) (2006) 659–662.

[7] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, et al. "Accurate whole genome sequencing using reversible terminator chemistry." *Nature* 456 (2008) 53–59.

[8] T.R. Bhangale, M.J. Rieder, and D.A. Nickerson. "Estimating coverage and power for genetic association studies." *Nature Genetics* 40(7) (2008) 841–843.

[9] S.R. Browning. "Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes." *Genetics* 187 (2008) 2123–2132.

[10] S.R. Browning and B.L. Browning. "High resolution detection of identity by descent in unrelated individuals." *American Journal of Human Genetics* 86 (2010) 1–14.

[11] N.H. Chapman and E.A. Thompson. "A model for the length of tracts of identity by descent in finite random mating populations." *Theoretical Population Biology* 64 (2003) 141–150.

[12] A.G. Clark. "The size distribution of homozygous segments in the human genome." *The American Journal of Human Genetics* 65 (6) (1999) 1489–1492.

[13] D. Curtis, A.E. Vine, and J. Knight. "Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations." *The American Journal of Human Genetics* 72 (2) (2008) 261–278.

[14] R.A. Fisher. "A fuller theory of junctions in inbreeding." *Heredity* 8 (9154) 187–197.

[15] J. Gibson, N.E. Morton, and A. Collins. "Extended tracts of homozygosity in outbred human populations." *Human Molecular Genetics* 15 (5) (2006) 789–795.

[16] A. Gusev, J.K. Lowe, M. Stoffel, M.J. Daly, D. Altshuler, J.L. Breslow, J.M. Friedman, and I. Pe'er. "Whole population, genome-wide mapping of hidden relatedness." *Genome Research* 19 (2009) 318–326.

[17] B.J. Hayes, P.M. Visscher, H.C McPartlan, and M.E. Goddard. "Novel multilocus measure of linkage disequilibrium to estimate past effective population size." *Genome Research* 13 (2003) 635–643.

[18] J. Hein, M.H. Schierup, and C. Wiuf. *Gene Genealogogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford: New York 2005.

[19] J. Hernandez-Sanchez, C.S. Haley, and J.A. Woolliams. "Prediction of IBD based on population history for fine gene mapping." *Genetics, Selection, and Evolution* 38 (2006) 231–252.

[20] W.G. Hill and J. Hernandez-Sanchez. "Prediction of multilocus identity-by-descent." *Genetics* 176 (2007) 2307–2315.

[21] W.G. Hill and A. Robertson. "Linkage disequilibrium in finite populations." *Theoretical and Applied Genetics* 38 (6) (1968) 1432–2242.

[22] A. Hodgkinson and A. Eyre-Walker. "Human triallelic sites: evidence for a new mutational mechanism?" *Genetics* 184 (2010) 233–241.

[23] B.N. Howie, P. Donnelly, and J. Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLOS Genetics* 5 (2009).

[24] L. Huang, Y. Li, A.B. Singleton, J.A. Hardy, G. Abecasis, N.A. Rosenberg, and P. Scheet. "Genotype-imputation accuracy across worldwide human populations." *The American Journal of Human Genetics* 84 (2009) 235–250.

[25] R.R. Hudson. "Generating samples under a Wright-Fisher neutral model of genetic variation." *Bioinformatics* 18 (2002) 337–338.

[26] I. Ionita-Laza, C. Lange, and N.M Laird. "Estimating the number of unseen variants in the human genome." *Proceedings of the National Academy of Sciences* 106(13) (2009) 5008–5013.

[27] The International HapMap Consortium. "A second generation human haplotype map of over 3.1 million SNPs." *Nature* 447 (2007) 661–678.

[28] E. Jakkula, K. Rehnstrom, T. Varilo, O.P.H. Pietilainen, T. Paunio, N.L. Pedersen, U. deFaire, M.-R. Jarvelin, J. Sahrinen, N. Freimer, S. Ripatti, S. Purcell, A. Collins, M.J. Daly, A. Palotie, and L. Peltonen. "The genome-wide patterns of variation expose significant substructure in a founder population." *The American Journal of Human Genetics* 83 (2008) 1–8.

[29] X. Ke, M.S. Taylor, and L.R. Cardon. "Singleton SNPs in the human genome and implications for genome-wide association studies." *European Journal of Human Genetics* 16 (2008) 506–515.

[30] A. Keinan, J.C. Mullikan, N. Patterson, and D. Reich. "Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans." *Nature Genetics* 39 (2007) 1251–1255.

[31] A. Kong, G. Masson, M.L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P.I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D.F. Gudbjartsson, H. Stefansson, and K. Stefansson. "Detection of sharing by descent, long-range phasing and haplotype imputation." *Nature Genetics* 40.9 (2008) 1068–1075.

[32] G. Laval, E. Patin, L.B. Barriero, and L. Quintana-Murci. "Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions." *PLOS ONE* (2010).

[33] A.-L. Leutnegger, B. Prum, E. Genin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E.A. Thompson. "Estimation of the inbreeding coefficient through use of genomic data." *American Journal of Human Genetics* 73 (2003) 516–523.

[34] A.-L. Leutnegger, A. Labalme, E. Genin, A. Toutain, E. Steichen, F. Clerget-Darpoux, and P. Edery. "Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to

Taybi-Linder syndrome." *American Journal of Human Genetics* 79 (2006) 62–66.

[35] Q. Li and R. Wu. "A multilocus model for constructing a linkage disequilibrium map in human populations." *Statistical Applications in Genetics and Molecular Biology* 8 (1) Article 18 (2009).

[36] Y. Li, C. Willer, S. Sanna, and G. Abecasis. "Genotype imputation." *Annual Reviews of Genomics and Human Genetics* 10 (2009)387–406.

[37] F. Liu, C.M. van Duijn, and Y.S. Aulchenko. "Ignoring distant genealogic loops leads to false-positives in homozygosity mapping." *Annals of Human Genetics* 70 (2006) 1–6.

[38] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. "A new multipoint method for genome-wide association studies by imputation of genotypes." *Nature Genetics* 39 (2007) 906–913.

[39] J. Marchini and B. Howie. "Genotype imputation for genome-wide association studies." *Nature Reviews: Genetics* 11 (2010) 499–511.

[40] G.T. Marth, E. Czabarka, J. Murvai, and S.T. Sherry. "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations." *Genetics* 166 (2004) 351–372.

[41] R. McQuillian, A.-L. Leutenegger, R. Abdel-Rahman, C.S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, A.K. MacLeod, S.M. Farraington, P. Rudan, C. Hayward, V. Vitart, I. Rudan, S.H. Wild, M.G. Dunlop, A.F. Wright, H. Campbell, and J.F. Wilson. "Runs of homozygosity in European populations." *American Journal of Human Genetics* 83 (3) (2008) 359–372.

[42] G.A.T. McVean. "A genealogical intepretation of linkage disequilibrium." *Genetics* 162 (2002) 987–991.

[43] G.A.T. McVean and N.J. Cardin. "Approximating the coalescent with recombination." *Philosophical Transactions of the Royal Society B* 360 (2005) 1387–1393.

[44] P. Mellars. "Going East: New genetic and archaeological perspectives on the modern human colonization of Eurasia." *Science* 11 (2006) 796–800.

[45] T.H.E Meuwissen and M.E. Goddard. "The use of marker haplotypes in animal breeding schemes." *Genetics, Selection, and Evolution* 28 (1996) 161–176.

[46] T.H.E. Meuwissen and M.E. Goddard. "Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci." *Genetics* 155 (2000) 421–430.

[47] T.H.E. Meuwissen and M.E. Goddard. "Prediction of identity by descent probabilities from marker-haplotypes." *Genetics, Selection, and Evolution* 33 (2001) 605–634.

[48] T.H.E. Meuwissen, B.J. Hayes, and M.E. Goddard. "Prediction of total genetic value using genome-wide dense marker maps." *Genetics* 157 (4) (2001) 1819–1829.

[49] T.H.E. Meuwissen and M.E. Goddard. "Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data." *Genetics, Selection, and Evolution* 36 (2004) 261–279.

[50] T. Ohta and M. Kimura. "Linkage disequilibrium between two segretation nucleotide sites under the steady flux of mutations in a finite population." *Genetics* 68 571–580.

[51] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, and M.A.R Ferreira. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *American Journal of Human Genetics* 81 (2007) 559–575.

[52] N.A. Sheehan and T. Egeland. "Adjusting for founder relatedness in a linkage analysis using prior information." *Human Heredity* 65 (2008) 221–231.

[53] C.C.A. Spencer, Z. Su, P. Donnelly, and J. Marchini. "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." *PLOS Genetics* 5(5) (2009) e1000477.

[54] P. Stam. "The distribution of the fraction of the genome identical by descent in finite random mating populations." *Genetical Research* 35 (1980) 131–155.

[55] J.A. Sved. "Linkage disequilibrium and homozygosity of chromosome segments in finite populations." *Theoretical Population Biology* 2 (2) (1971) 125–141.

[56] J.D. Terwilliger and T. Hiekkalinna. "An utter refutation of the fundamental theorem of the HapMap." *European Journal of Human Genetics* 14 (2006) 426–437.

[57] E.A. Thompson. "The IBD process along four chromosomes." *Theoretical Population Biology* 73 (3) (2008) 369–373.

[58] L.N. Trefethen. *Numerical Linear Algebra*. SIAM, Philadelphia: 1997.

[59] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company, Greenwood: 2009.

[60] J. Wang, W. Wang, R. Li, et al. "The diploid genome sequence of an Asian individual." *Nature* 456 (2008) 60–65.

[61] B.S. Weir and W.G. Hill. "Effect of mating structure on variation in linkage disequilibrium." *Genetics* 95 477–488.

Since

$$\int_{t_0=0}^{\infty} e^{-t_0(1+i(\mu+\rho))}(1-e^{-t_0\rho})e^{-t(1+\mu+\rho)}dt_0 = \frac{\rho e^{-t(1+\mu+\rho)}}{(1+i(\mu+\rho))(1+\rho+i(\mu+\rho))},$$

we can let

$$K_i = \frac{\rho}{(1+i(\mu+\rho))(1+\rho+i(\mu+\rho))}$$

and conclude that

$$B_i(L) = B_{i+1}(L+1)$$

for all $i > 1$, whereas

$$B_1(L) = \sum_{i=1}^{L-1} K_i B_i(L-1).$$

7 Empirical validation using genome sequence data

To measure the accuracy of our predicted $p_L(\text{IBS})$ values, we found the lengths of all maximal ROHs in the eleven human genome sequences referenced in Table 13. The bases were re-called in a consistent fashion with the intent to make the quality good enough for population genetic analysis; out of a total of 33,686,389,482 base pairs, 9,743,948,741 (28.9%) were marked unreliable due to unreliable read mapping, proximity to indels, or other other attributes that made them suspect (see Base Calling Methods appendix), and we deleted these bases before proceeding. Our call sets for all sequences are available for download at ftp://ftp.sanger.ac.uk/pub/rd/humanSequences.

In addition to counting the number $N_{\text{ROH}}(L)$ of ROHs in each genome that are between $(L-1000)$ and L bases long (for L divisibly by 1000), we counted the number $N_{\text{ROH}}(L)_{0.10}$ of L-base regions that appear homozygous when we detect a tenth of all hets. Specifically, we generated *thinned ROHs* whose endpoints are the mutations with positions congruent to zero mod ten relative to the 5' end of the chromosome, referring to these endpoints as *observed hets* as opposed

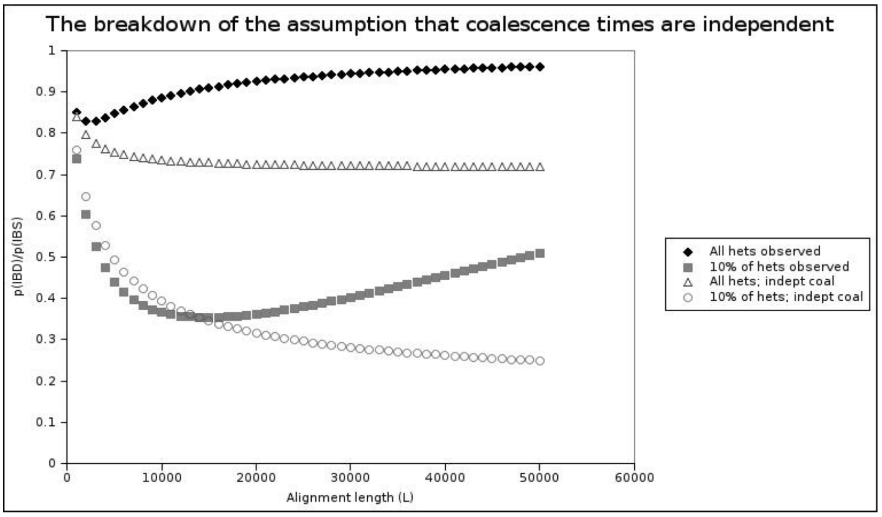

Figure 12: The two solid plots record the exact probability of IBD given IBS (assuming the SMC), the black plot with IBS required at every base in the alignment, and the grey plot where markers allow for only 10% of mutations to be detected. The empty triangles and empty circles record how that probability changes when we disregard linkage between non-IBD markers. The two curves almost never agree when complete sequences are used, in concordance with the fact that earlier methods do not claim to be accurate for such dense marker data.

Sequence Names	Origins
COLO-829-BL	Northern/Western European Ancestry [5]
NA12878, NA12891, NA12892	Northern/Western European Ancestry [1]
NA18507	Yoruba, Nigeria [7]
NA18506, NA18508, NA19239, NA19240	Yoruba, Nigeria (unpublished)
SJK	Korean [4]
YH	Chinese [60]

Figure 13: The eleven genomes used in our analysis

to *hidden hets*. We predict that

$$\frac{N_{\mathrm{ROH}}(L)}{N_{\mathrm{ROH}}(L)_{0.10}} = \frac{10 p_{L\max}(\mathrm{IBS})}{p_{L\max}(\mathrm{IBS})_{0.10}}, \tag{23}$$

adding the factor of 10 to account for the fact that there are ten times as many true ROHs as thinned ROHs (most of the excess ones being short).

Even though we take care to use genome data with a very low error rate, false positive hets (on the order of 1 per 10^5 bases) will present a significant problem for our analysis. We will be estimating the abundance of ROHs up to 10^7 bases long, and there is an overwhelming chance that their homozygosity will be broken up by false positives.

To correct for the breakup of ROHs by false positives, we estimate the false positive frequency f and multiply the measured value of $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ by $(1-f/10)^L/(1-f)^L$, reasoning that $(1-f)^L$ is the probability that an L-base ROHs will be broken up by a false positive het. We choose $f = 1.5 \times 10^{-5}$ because $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ tends toward $(1-f)^L/(1-f/10)^L$ in each genome as L gets large, while the ratio of thinned to true ROHs should tend toward 1.

The eleven plots of $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ versus L cluster clearly by ethnicity (see Figures 14, 15, 16), and we account for the differences by finding effective population size histories that fit $10p_{L\max}(\mathrm{IBS})/p_{L\max}(\mathrm{IBS})_{0.10}$ well in the data from each ethnic group. We also experiment with varying the mutation rate μ, motivated by the fact that the 1000 Genomes consortium recently

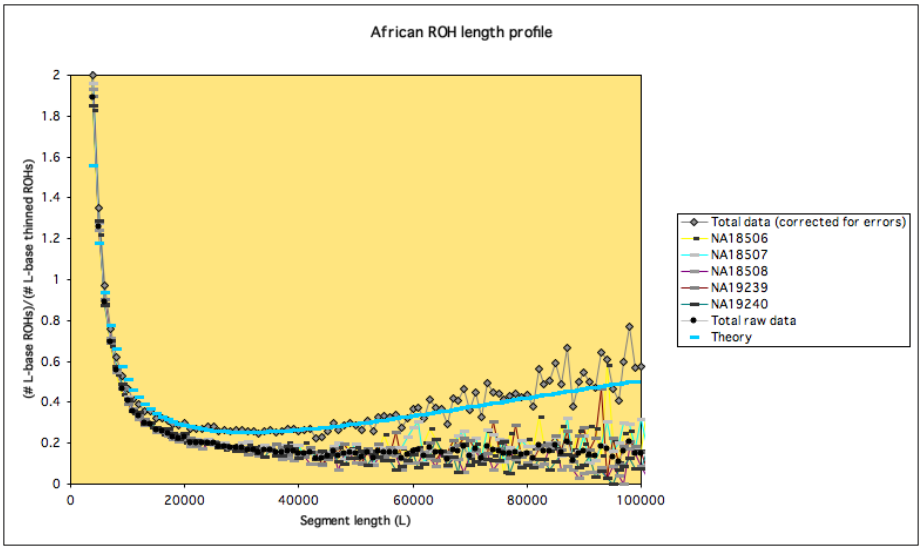

Figure 14: **A. Regions of homozygosity in African genome data** Here, we separately plot $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ for each of five African genomes, then average this function across the genomes and correct it for 1.5×10^{-5} false positives per base. In blue is the theory plot $\frac{10p_{L\max}(\mathrm{IBS})}{p_{L\max}(\mathrm{IBS})_{0.10}}$ for a population of constant effective size $N = 14,000$ and a mutation rate of $m = 1.6 \times 10^{-8}$ per base per generation (one of many histories that minimize the sum of square distances from the data points to the predicted curve).

estimated μ to be 1×10^{-8} per base per generation [1] rather than 2.5×10^{-8}.

The measured $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ ratios behave noisily for $L > 100,000$, likely because there are few such ROHs in the genome and each one is more likely than a short ROH to include recombination hotspots or other sites where the theory in this paper breaks down. Therefore, we define the best fit population history to be the one that minimizes the sum of squares distance from the predicted $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ values to the measured $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ values, the sum taken over L ranging from 10,000 to 100,000.

Let $T_{\mu,H}(L)$ denote the theory plot of $10p_{L\max}(\mathrm{IBS})/p_{L\max}(\mathrm{IBS})_{0.10}$ that is obtained a function of the mutation rate μ and the piecewise-constant popula-

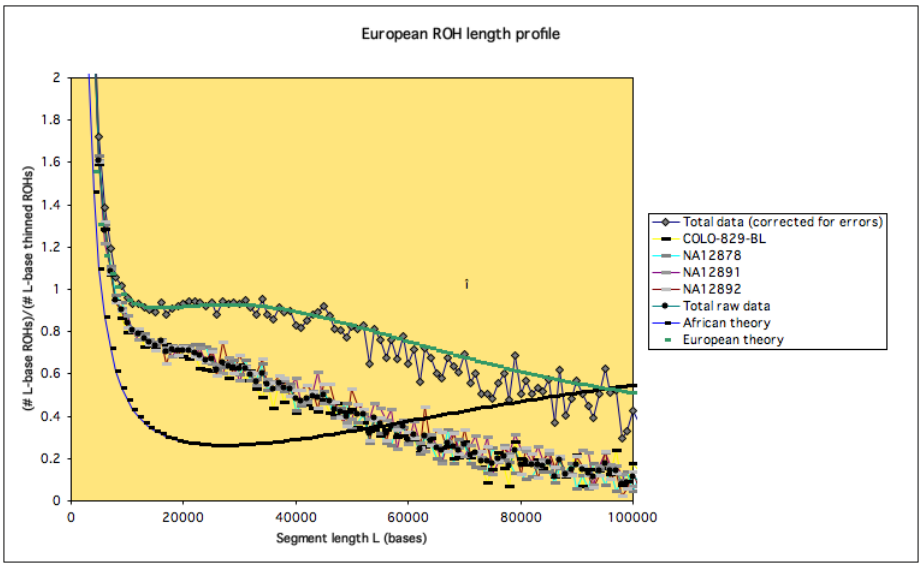

Figure 15: **B. Regions of homozygosity in European genome data** In green is the following single-bottleneck history, with mutation rate $m = 2.5 \times 10^{-8}$: $N = 11,900$, time ranging from 0 to 1240 generations ago (g.a.); $N = 4,530$, $1,240 - 1,770$ g.a.; $N = 15,000 \geq 1,770$ g.a. The African constant population size theory is included in black, for reference.

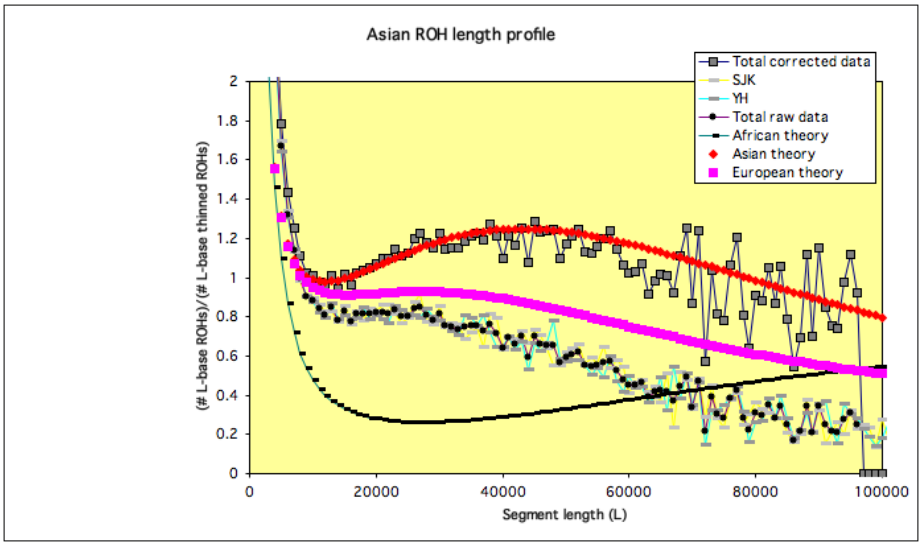

Figure 16: **C. Regions of homozygosity in Asian genome data** The single-bottleneck history shown in red, again assuming $m = 2.5 \times 10^{-8}$, is the following: $N = 8,670$, $0 - 1,380$ g.a.; $N = 1790$, $1,380 - 1,530$ g.a.; $N = 15,000 \geq 1,530$ g.a.. African and European theory plots are included for reference.

tion history H. Likewise, let $D_G(L)$ denote the set of data points $\{N_{\mathrm{ROH}}(10^3 \cdot i)/N_{\mathrm{ROH}}(10^3 \cdot i)_{0.10}\}_{i=1}^{10^3}$ that is obtained by counting all of the ROHs and thinned ROHs in some set of genomes G and correcting for 1.5×10^{-5} false positive hets per base. We measure the goodness of fit between the parameters (μ, H) and the data set G by calculating the sum of squares fit

$$SS(\mu, H, G, L) = \sum_{i=1}^{900} \left(T_{\mu,H}(10^3 \cdot (10+i)) - D_G(10^3 \cdot (10+i))\right)^2.$$

It remains to define a threshold for $SS(\mu, H, G, L)$ below which (μ, H) is deemed a good fit for G. Since $T_{\mu,H}(L)$ is not a straight line, we cannot perform a goodness-of-fit linear regression. We find it logical, instead, to define a threshold that depends on the noisiness of the curve $D_G(L)$, letting

$$N(G, L) = \sum_{i=1}^{(L-1)/1000} \left(D_G(10^3 \cdot (10+i+1)) - D_G(10^3 \cdot (10+i))\right)^2$$

denote the sum of squared distances between adjacent points of $D_G(L)$. If $T_{\mu,H}(L) = \frac{1}{2}(D_G(L) - D_G(L+1))$, making $T_{\mu,H}(L)$ a smoothed version of the data set $D_G(L)$, then

$$SS(\mu, H, G, L) = \frac{1}{4}N(G, L),$$

In each data plot $D_G(L)$, the left portion of the graph is much less noisy than the right portion and therefore provides more information about the mutation rate and population history. In the European genomes, for example, there is so little noise in the data set $D(G)|_{L<34000}$ that

$$\frac{1}{4}N(G_{\mathrm{European}}, 34000) < 0.0094,$$

while

$$\frac{1}{4}N(G_{\mathrm{European}}, 90000) > 0.26.$$

For each of the genome groups $G_{\mathrm{African}}, G_{\mathrm{European}}$, and G_{Asian}, we define L_{short} to be the largest L satisfying $\frac{1}{4}N(G, L-1) < 0.01$ and define L_{long} to be the longest $L \geq 1000$ satisfying $\frac{1}{4}N(G, L-1) < 0.5$ (specific values of L_{short} and L_{long} are recorded in Table 17). We then say that (μ, H) is a good fit for G if

$$SS(\mu, H, G, L_{\mathrm{short}}) < 0.01$$

Ethnicity	L_{short}	L_{long}
African	53000	90000
European	35000	89000
Asian	21000	60000

Figure 17: Thresholds for low noise ($N(G, L_{\text{short}} - 1) < 0.01$) and medium noise ($N(G, L_{\text{long}} - 1) < 0.2$) in the $D_G(L)$ ROH data sets. Our Asian data set, which contains half as many genomes as the others, appears commensurately noisier.

and

$$SS(\mu, H, G, L_{\text{long}}) < 0.5.$$

We searched for good parameter fits using a Monte Carlo Markov chain approach, beginning with a search of constant population size histories. As expected, the Africans are the only group for which we find good constant population size histories. Such histories fall within a narrow parameter space, namely $13,000 \leq N \leq 15,000$ and $1.55 \times 10^{-8} < m < 1.7 \times 10^{-8}$.

When we allow for a single population expansion or contraction, we find a large variety of histories that fit the African data well, though we still find no fits for the European or Asian data. These good African histories are all expansions when $m = 2.5 \times 10^{-8}$, all contractions when $m = 1 \times 10^{-8}$, and close to constant for $m = 1.75 \times 10^{-8}$, with a population size change in either the very recent past or the very distant past (see Figures 18, 19, and 20).

To find good theory fits for the European and Asian data, it was necessary to invoke a population bottleneck. To speed up our MCMC search, we fixed the mutation rate $m = 2.5 \times 10^{-8}$ and the ancestral population size $N_3 = 15,000$. This left two variable time parameters and two variable size parameters, enough to generate many optimal histories to fit both sets of non-African data (see Figures 21 and 22). In both sets of good histories, recent good-fit bottlenecks are shallower than ancient good-fit bottlenecks. The modern effective population size is lower, on average, in the Asian histories. This fits with the fact that the

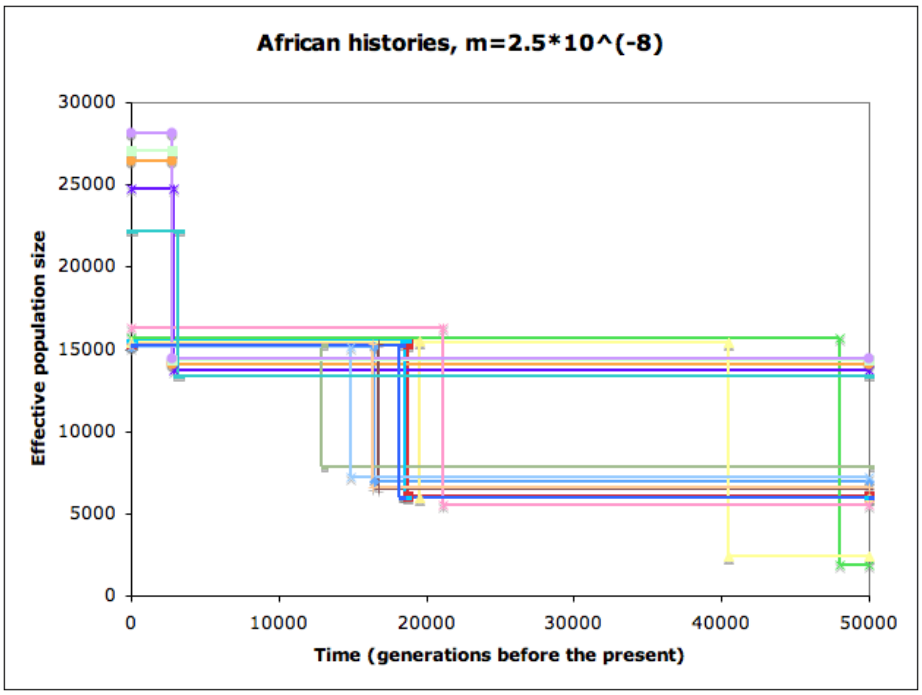

Figure 18: Some population histories satisfying our good fit criterion for African genome data assuming $m = 2.5 \times 10^{-8}$ mutations per base per generation

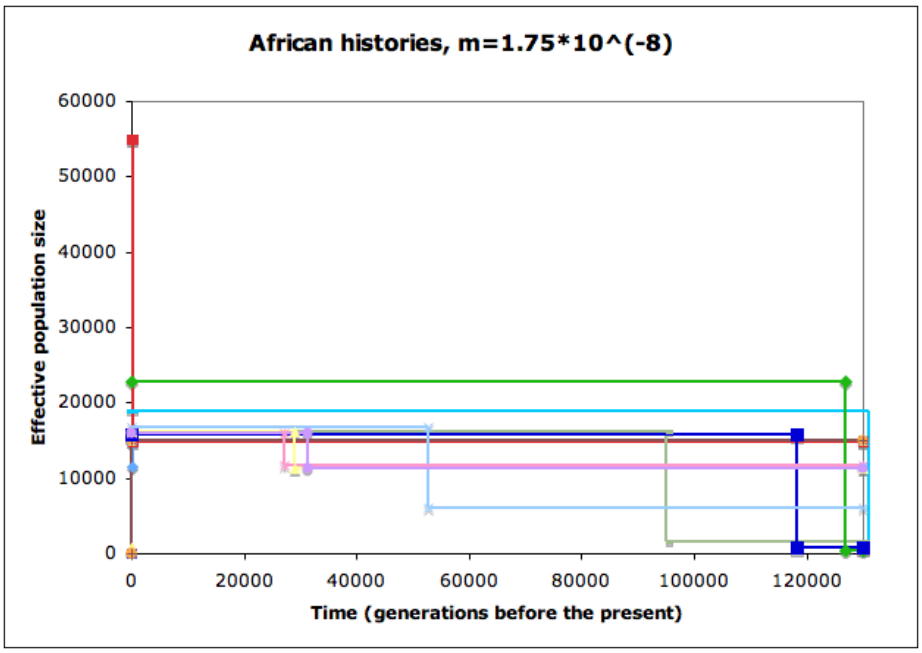

Figure 19: More good fit YRI histories; mutation rate $m = 1.75 \times 10^{-8}$ per base per generation

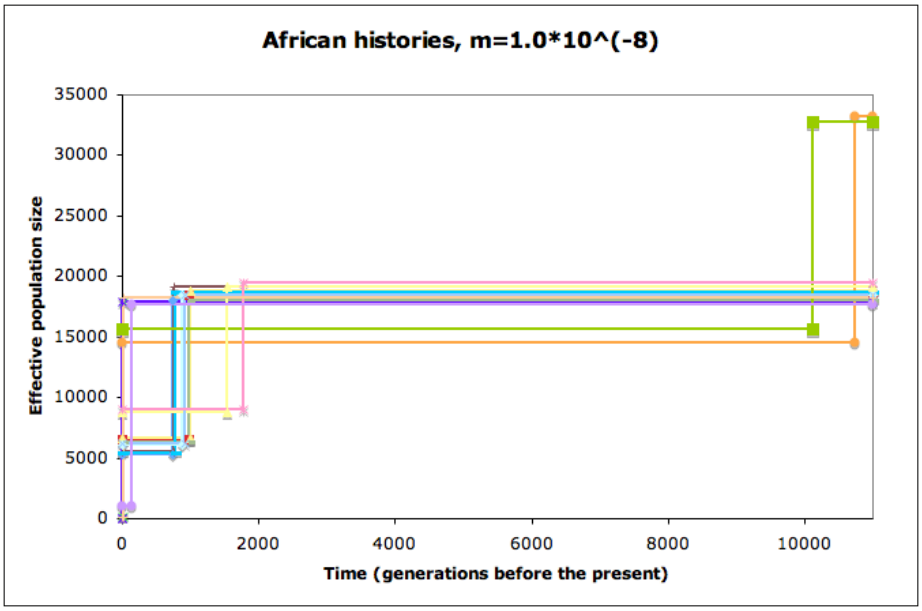

Figure 20: More good fit YRI histories, mutation rate $m = 1.0 \times 10^{-8}$ per base per generation

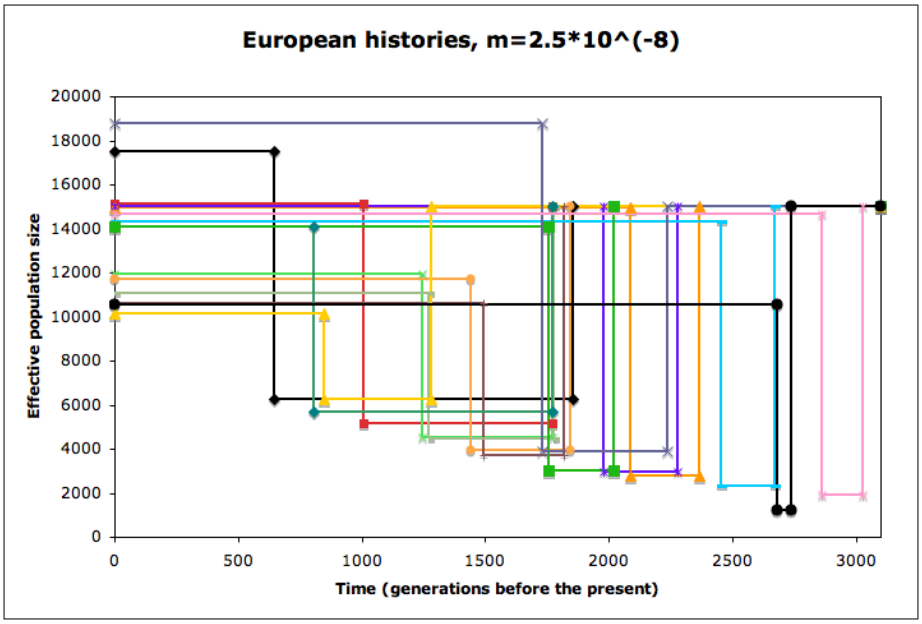

Figure 21: A population bottleneck is required to fit the CEU ROH data–either a shallow, recent bottleneck or a deeper, more ancient bottleneck.

Asian HapMap allele frequency spectrum shows more evidence of genetic drift than the European HapMap allele frequency spectrum [30].

It remains an open problem to mathematically describe the set of histories that fit the distribution of ROHs in each ethnic group. However, in showing that such histories exist, we achieve our aim of predicting the length distribution of ROHs in real genome data and validating the theory that we use to compute IBD probabilities.

8 Discussion

In this paper, we attempt to very precisely model patterns of linkage disequilibrium in genetic data, capturing its decay over long regions of the genome instead of assuming that certain blocks or pairs of loci assort independently. Rather than adding a new LD model to the myriad that exist already, we work

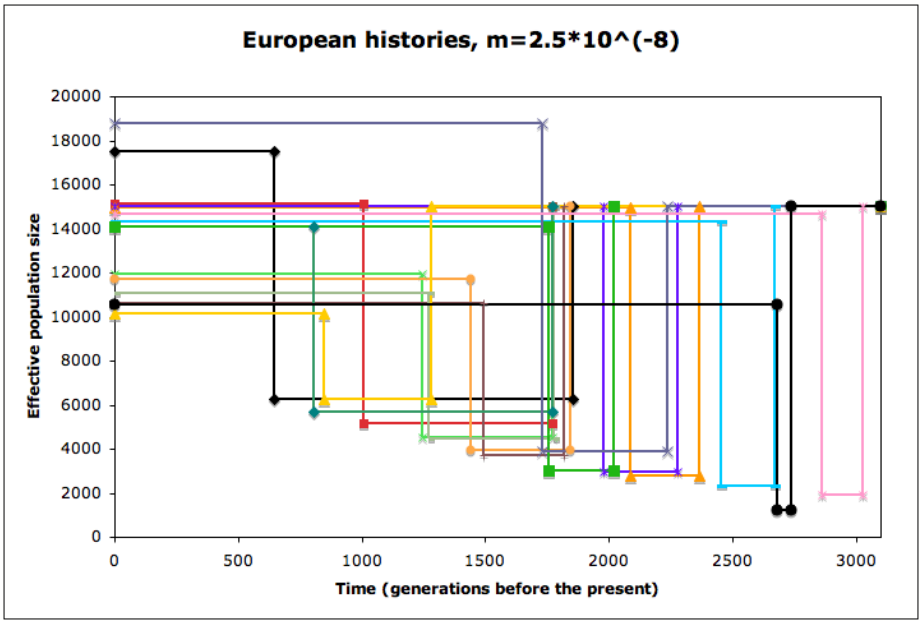

Figure 21: A population bottleneck is required to fit the CEU ROH data–either a shallow, recent bottleneck or a deeper, more ancient bottleneck.

Asian HapMap allele frequency spectrum shows more evidence of genetic drift than the European HapMap allele frequency spectrum [30].

It remains an open problem to mathematically describe the set of histories that fit the distribution of ROHs in each ethnic group. However, in showing that such histories exist, we achieve our aim of predicting the length distribution of ROHs in real genome data and validating the theory that we use to compute IBD probabilities.

8 Discussion

In this paper, we attempt to very precisely model patterns of linkage disequilibrium in genetic data, capturing its decay over long regions of the genome instead of assuming that certain blocks or pairs of loci assort independently. Rather than adding a new LD model to the myriad that exist already, we work

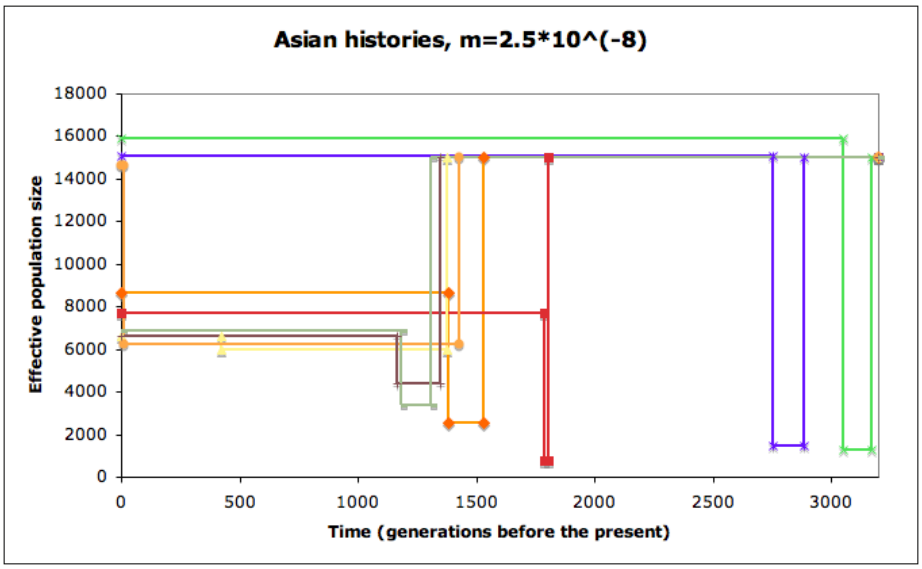

Figure 22: A bottleneck is required to fit Asian ROH data. These good fit histories have smaller recent effective population sizes than the histories that fit the CEU data (see Figure 21).

as closely as possible to the neutral coalescent with recombination, deriving results within a theoretical framework whose advantages and limits are well understood already. Our claims about patterns of homozygosity in the genome make no assumptions that weren't published as part of the sequentially Markovian coalescent (SMC) model [43]; while we do not account for recombination hotspots or other point irregularities, the results should be quite accurate across the genome as a whole, as seen empirically in Section 7.

While it was necessary to relax the SMC somewhat to model LD in diploid alignments with uncertain phasing, our model approaches the coalescent in the limit of high LD where phasing and imputation attain high accuracy. While it is possible to check our computations with a coalescent simulator such as MS, our method makes it much quicker to compute a length spread of IBS probabilities; without it, each data point on each of our plots would have to be obtained from a separate coalescent simulation with memory and storage requirements that get quite large for long IBS alignments. Efficiency was key when we had to compare many population histories to find a matches for the African, European, and Asian homozygosity data, and also when we considered the best alignment between a test sequence and some member of a reference panel.

The accuracy values that we compute for various imputation panels assume an idealized population of effective size 10,000, closer to the truth for Africans than for non-Africans. However, it is possible to condition instead on any piecewise-constant population size history, as is briefly explored in the last section of the paper, if e.g. the goal is to impute only Europeans. We concentrate less on evaluating the capabilities of an existing panel than on building a framework for making informed decisions about the design of future panels.

Our results do posit lower bounds on the amount of variation that should not be imputable using HapMap, as we estimate that a 120-reference panel has only a 70% chance of imputing a one kilobase mini-haplotype with 99% accuracy and an 80% chance of imputing it with 90% accuracy. Accuracy is likely higher when the aim is to impute only common SNPs, but it can only be higher insofar as those common SNPs fail to tag the rarer SNPs around them. Our rough

estimate of HapMap variation coverage is close to that obtained by Bhangale, et al., who resequenced 1.6 megabases in each of the HapMap individuals and reported that only 60-80% of the SNPs they found were within $r^2 > 0.8$ of a tag SNP [8].

Besides showing that it is feasible to study imputation with coalescent theory, we hope that the tricks and shortcuts we've developed might help make coalescent theory more applicable to other hard problems. Important as imputation is, it is not the only setting where understanding IBS could be useful. As seen in Section 7, real hets are distributed differently from false positive hets, such that hets scattered in regions of high homozygosity are very likely to be false positives. It is easy to compute the false positive probability of a het that is L bases away from the nearest het, and it might be useful to incorporate this result into base calling software. A long stretch of near-homozygosity provides a lot of evidence that the region is IBD (see Figure 8 for a plot of p(IBS) versus length), and there is less than a 1.5% chance that an IBD region will contain seven or more mismatches (the number of mismatches being Poisson-distributed with a mean of $\mu/\rho = 2.5$). When the sequencing error rate is 10^{-5} and a 10^7-base region contains about 100 scattered hets, the likely truth is that the region is IBD and most of those hets are errors.

A challenge for the future will be to address the effect of recent exponential population growth. Only the frequency of the longest IBS regions should be affected, but this will be enough to make a 1000-haplotype reference panel less perfect than it appears in figures 10 and 11; Ionita-Laza, et al. predict that more than 3,000 individuals will be needed to find all variants with frequency 0.1% based on allele frequency data from the outbred CEU and YRI sequence data [26]. One way around this issue would be to conduct GWASs in moderately isolated populations where exponential growth has been very recent and 1000 references do constitute a perfect imputation panel. Family-based linkage studies have been successful for some time at discovering functional variants that affect few people but shed valuable light on disease mechanisms, and our results suggest that IBD mapping need not be limited to groups as small as families. A

project the size of 1000 Genomes [1] should make it possible to essentially know the sequence of every individual in a small population like Iceland, which would, in principle, make it possible to test for functionality all across the genome, including at a large pool of rare and untaggable SNPs that are invisible even in GWASs conducted with the help of imputation from HapMap.

A Base-calling methods

The sequences used to validate our method were re-called with the hope of reducing the frequency of errors. The call sets are available for download at ftp://ftp.sanger.ac.uk/pub/rd/humanSequences.

Raw Illumina read data were obtained from NCBI's sequence read archive (see Table 13) and EMBL's European read archive. They were aligned by BWA (0.5.5), using human reference genome build 36, which masks pseudoautosomal regions on Y by including unassembled contigs and the Epstein-Barr virus genome (AC:NC_007605). Low quality bases were trimmed from the 3'-ends of Illumina short reads by applying BWA option '-q 15.' Because SJK base qualities were overestimated, they were recalibrated using the Genome Analysis Toolkit after discarding SNPs known from dbSNP-129. Default BWA-SW algorithms were used for capillary reads.

The 'pileup' command of the *samtools* software package was used to create each autosome's diploid consensus. The consensus was then filtered, the following kinds of bases being marked low-confidence calls:

1. Read depth is more than twice or less than half of the depth estimated at loci genotyped in HapMap3
2. Reads covering the locus have root mean square mapping quality below 25
3. There is a predicted short indel less than 10 base pairs away
4. Inferred consensus quality is below 20 (Illumina data) or 10 (capillary

data)

5. Out of the 35 reference sequence 35-mers that overlap with the site, fewer than 18 can be mapped elsewhere with at most one mismatch

References

[1] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature* (2010) 1061–1073.

[2] C.A. Anderson, D. Brocklebank, and A.P. Morris. "A comparison of reference panels for imputation in genome-wide association studies." *The European Journal of Human Genetics*, to appear.

[3] C.A. Anderson, F.H. Pettersson, J.C. Barrett, J.J. Zhuang, J. Ragoussis, L.R. Cardon, and A.P. Morris. "Evaluation the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms." *The American Journal of Human Genetics* (2008) 112–119.

[4] S.-M. Ahn, T.-H. Kim, S. Lee, et al. "The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group." *Genome Research* 19(9) (2009) 1622–1629.

[5] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dougan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. "THe COSMIC (Catalogue of Somatic Mutations in Cancer) database and website." *British Journal of Cancer* 91 (2004) 355–358.

[6] J.C. Barrett and L.R. Cardon. "Evaluating the coverage of genome-wide association studies." *Nature Genetics* 38(6) (2006) 659–662.

[7] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, et al. "Accurate whole genome sequencing using reversible terminator chemistry." *Nature* 456 (2008) 53–59.

data)

5. Out of the 35 reference sequence 35-mers that overlap with the site, fewer than 18 can be mapped elsewhere with at most one mismatch

References

[1] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature* (2010) 1061–1073.

[2] C.A. Anderson, D. Brocklebank, and A.P. Morris. "A comparison of reference panels for imputation in genome-wide association studies." *The European Journal of Human Genetics*, to appear.

[3] C.A. Anderson, F.H. Pettersson, J.C. Barrett, J.J. Zhuang, J. Ragoussis, L.R. Cardon, and A.P. Morris. "Evaluation the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms." *The American Journal of Human Genetics* (2008) 112–119.

[4] S.-M. Ahn, T.-H. Kim, S. Lee, et al. "The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group." *Genome Research* 19(9) (2009) 1622–1629.

[5] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dougan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. "THe COSMIC (Catalogue of Somatic Mutations in Cancer) database and website." *British Journal of Cancer* 91 (2004) 355–358.

[6] J.C. Barrett and L.R. Cardon. "Evaluating the coverage of genome-wide association studies." *Nature Genetics* 38(6) (2006) 659–662.

[7] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, et al. "Accurate whole genome sequencing using reversible terminator chemistry." *Nature* 456 (2008) 53–59.

[8] T.R. Bhangale, M.J. Rieder, and D.A. Nickerson. "Estimating coverage and power for genetic association studies." *Nature Genetics* 40(7) (2008) 841–843.

[9] S.R. Browning. "Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes." *Genetics* 187 (2008) 2123–2132.

[10] S.R. Browning and B.L. Browning. "High resolution detection of identity by descent in unrelated individuals." *American Journal of Human Genetics* 86 (2010) 1–14.

[11] N.H. Chapman and E.A. Thompson. "A model for the length of tracts of identity by descent in finite random mating populations." *Theoretical Population Biology* 64 (2003) 141–150.

[12] A.G. Clark. "The size distribution of homozygous segments in the human genome." *The American Journal of Human Genetics* 65 (6) (1999) 1489–1492.

[13] D. Curtis, A.E. Vine, and J. Knight. "Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations." *The American Journal of Human Genetics* 72 (2) (2008) 261–278.

[14] R.A. Fisher. "A fuller theory of junctions in inbreeding." *Heredity* 8 (9154) 187–197.

[15] J. Gibson, N.E. Morton, and A. Collins. "Extended tracts of homozygosity in outbred human populations." *Human Molecular Genetics* 15 (5) (2006) 789–795.

[16] A. Gusev, J.K. Lowe, M. Stoffel, M.J. Daly, D. Altshuler, J.L. Breslow, J.M. Friedman, and I. Pe'er. "Whole population, genome-wide mapping of hidden relatedness." *Genome Research* 19 (2009) 318–326.

[17] B.J. Hayes, P.M. Visscher, H.C McPartlan, and M.E. Goddard. "Novel multilocus measure of linkage disequilibrium to estimate past effective population size." *Genome Research* 13 (2003) 635–643.

[18] J. Hein, M.H. Schierup, and C. Wiuf. *Gene Genealogogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford: New York 2005.

[19] J. Hernandez-Sanchez, C.S. Haley, and J.A. Woolliams. "Prediction of IBD based on population history for fine gene mapping." *Genetics, Selection, and Evolution* 38 (2006) 231–252.

[20] W.G. Hill and J. Hernandez-Sanchez. "Prediction of multilocus identity-by-descent." *Genetics* 176 (2007) 2307–2315.

[21] W.G. Hill and A. Robertson. "Linkage disequilibrium in finite populations." *Theoretical and Applied Genetics* 38 (6) (1968) 1432–2242.

[22] A. Hodgkinson and A. Eyre-Walker. "Human triallelic sites: evidence for a new mutational mechanism?" *Genetics* 184 (2010) 233–241.

[23] B.N. Howie, P. Donnelly, and J. Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLOS Genetics* 5 (2009).

[24] L. Huang, Y. Li, A.B. Singleton, J.A. Hardy, G. Abecasis, N.A. Rosenberg, and P. Scheet. "Genotype-imputation accuracy across worldwide human populations." *The American Journal of Human Genetics* 84 (2009) 235–250.

[25] R.R. Hudson. "Generating samples under a Wright-Fisher neutral model of genetic variation." *Bioinformatics* 18 (2002) 337–338.

[26] I. Ionita-Laza, C. Lange, and N.M Laird. "Estimating the number of unseen variants in the human genome." *Proceedings of the National Academy of Sciences* 106(13) (2009) 5008–5013.

[27] The International HapMap Consortium. "A second generation human haplotype map of over 3.1 million SNPs." *Nature* 447 (2007) 661–678.

[28] E. Jakkula, K. Rehnstrom, T. Varilo, O.P.H. Pietilainen, T. Paunio, N.L. Pedersen, U. deFaire, M.-R. Jarvelin, J. Sahrinen, N. Freimer, S. Ripatti, S. Purcell, A. Collins, M.J. Daly, A. Palotie, and L. Peltonen. "The genome-wide patterns of variation expose significant substructure in a founder population." *The American Journal of Human Genetics* 83 (2008) 1–8.

[29] X. Ke, M.S. Taylor, and L.R. Cardon. "Singleton SNPs in the human genome and implications for genome-wide association studies." *European Journal of Human Genetics* 16 (2008) 506–515.

[30] A. Keinan, J.C. Mullikan, N. Patterson, and D. Reich. "Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans." *Nature Genetics* 39 (2007) 1251–1255.

[31] A. Kong, G. Masson, M.L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P.I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D.F. Gudbjartsson, H. Stefansson, and K. Stefansson. "Detection of sharing by descent, long-range phasing and haplotype imputation." *Nature Genetics* 40.9 (2008) 1068–1075.

[32] G. Laval, E. Patin, L.B. Barriero, and L. Quintana-Murci. "Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions." *PLOS ONE* (2010).

[33] A.-L. Leutnegger, B. Prum, E. Genin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E.A. Thompson. "Estimation of the inbreeding coefficient through use of genomic data." *American Journal of Human Genetics* 73 (2003) 516–523.

[34] A.-L. Leutnegger, A. Labalme, E. Genin, A. Toutain, E. Steichen, F. Clerget-Darpoux, and P. Edery. "Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to

Taybi-Linder syndrome." *American Journal of Human Genetics* 79 (2006) 62–66.

[35] Q. Li and R. Wu. "A multilocus model for constructing a linkage disequilibrium map in human populations." *Statistical Applications in Genetics and Molecular Biology* 8 (1) Article 18 (2009).

[36] Y. Li, C. Willer, S. Sanna, and G. Abecasis. "Genotype imputation." *Annual Reviews of Genomics and Human Genetics* 10 (2009)387–406.

[37] F. Liu, C.M. van Duijn, and Y.S. Aulchenko. "Ignoring distant genealogic loops leads to false-positives in homozygosity mapping." *Annals of Human Genetics* 70 (2006) 1–6.

[38] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. "A new multipoint method for genome-wide association studies by imputation of genotypes." *Nature Genetics* 39 (2007) 906–913.

[39] J. Marchini and B. Howie. "Genotype imputation for genome-wide association studies." *Nature Reviews: Genetics* 11 (2010) 499–511.

[40] G.T. Marth, E. Czabarka, J. Murvai, and S.T. Sherry. "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations." *Genetics* 166 (2004) 351–372.

[41] R. McQuillian, A.-L. Leutenegger, R. Abdel-Rahman, C.S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, A.K. MacLeod, S.M. Farraington, P. Rudan, C. Hayward, V. Vitart, I. Rudan, S.H. Wild, M.G. Dunlop, A.F. Wright, H. Campbell, and J.F. Wilson. "Runs of homozygosity in European populations." *American Journal of Human Genetics* 83 (3) (2008) 359–372.

[42] G.A.T. McVean. "A genealogical intepretation of linkage disequilibrium." *Genetics* 162 (2002) 987–991.

[43] G.A.T. McVean and N.J. Cardin. "Approximating the coalescent with recombination." *Philosophical Transactions of the Royal Society B* 360 (2005) 1387–1393.

[44] P. Mellars. "Going East: New genetic and archaeological perspectives on the modern human colonization of Eurasia." *Science* 11 (2006) 796–800.

[45] T.H.E Meuwissen and M.E. Goddard. "The use of marker haplotypes in animal breeding schemes." *Genetics, Selection, and Evolution* 28 (1996) 161–176.

[46] T.H.E. Meuwissen and M.E. Goddard. "Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci." *Genetics* 155 (2000) 421–430.

[47] T.H.E. Meuwissen and M.E. Goddard. "Prediction of identity by descent probabilities from marker-haplotypes." *Genetics, Selection, and Evolution* 33 (2001) 605–634.

[48] T.H.E. Meuwissen, B.J. Hayes, and M.E. Goddard. "Prediction of total genetic value using genome-wide dense marker maps." *Genetics* 157 (4) (2001) 1819–1829.

[49] T.H.E. Meuwissen and M.E. Goddard. "Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data." *Genetics, Selection, and Evolution* 36 (2004) 261–279.

[50] T. Ohta and M. Kimura. "Linkage disequilibrium between two segretation nucleotide sites under the steady flux of mutations in a finite population." *Genetics* 68 571–580.

[51] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, and M.A.R Ferreira. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *American Journal of Human Genetics* 81 (2007) 559–575.

[52] N.A. Sheehan and T. Egeland. "Adjusting for founder relatedness in a linkage analysis using prior information." *Human Heredity* 65 (2008) 221–231.

[53] C.C.A. Spencer, Z. Su, P. Donnelly, and J. Marchini. "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." *PLOS Genetics* 5(5) (2009) e1000477.

[54] P. Stam. "The distribution of the fraction of the genome identical by descent in finite random mating populations." *Genetical Research* 35 (1980) 131–155.

[55] J.A. Sved. "Linkage disequilibrium and homozygosity of chromosome segments in finite populations." *Theoretical Population Biology* 2 (2) (1971) 125–141.

[56] J.D. Terwilliger and T. Hiekkalinna. "An utter refutation of the fundamental theorem of the HapMap." *European Journal of Human Genetics* 14 (2006) 426–437.

[57] E.A. Thompson. "The IBD process along four chromosomes." *Theoretical Population Biology* 73 (3) (2008) 369–373.

[58] L.N. Trefethen. *Numerical Linear Algebra*. SIAM, Philadelphia: 1997.

[59] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company, Greenwood: 2009.

[60] J. Wang, W. Wang, R. Li, et al. "The diploid genome sequence of an Asian individual." *Nature* 456 (2008) 60–65.

[61] B.S. Weir and W.G. Hill. "Effect of mating structure on variation in linkage disequilibrium." *Genetics* 95 477–488.

List of *M.Phil.* Thesis Corrections

Kelley Harris

March 28, 2011

1. The first three paragraphs (pages 7-8) are new (the more gradual introduction requested as correction 4). Paragraph 3 of the intro (the first complete paragraph on page 2) is partly new and partly old.

2. The first two complete paragraphs on page 9 were added to address the non-standard use of "identical by descent" (see correction 2)

3. Figure 1 on page 10 was converted to a higher quality .png graphic (correction 8)

4. In the first complete paragraph of page 13, two long lists of references were pruned down. Two more long reference lists were pruned in the paragraph that starts on page 13 and finishes on page 14 (correction 1)

5. Figure 2 on page 18 was converted to a higher quality .png graphic (correction 8)

6. Subsection 2.2 on page 19 was added to address the issue of non-uniform recombination (correction 5)

7. New figures were added on pages 25 and 30 in improve the clarity of the diploid theory section. In addition, page 28 consists mostly of new exposition of my coalescent approximation.

8. Figure 8 on page 34 was converted to a higher quality .png graphic (correction 8)

9. The paragraph beginning on page 35 and ending on page 36 was added to address the relationship of this theory to imputation as performed today (correction 7)

10. Figures 9-11 on page 40-42 were converted to higher quality .png graphics (correction 8)

11. Spelling of "independently" corrected on page 43, paragraph 1 (correction 6)

12. Long list of references pruned on page 43, paragraph 1, (correction 1)

13. In order to address correction 9, it was necessary to substantially rewrite section 7 (pages 44-55). Instead of an ad hoc procedure for estimating demographic parameters, I used a Monte Carlo Markov Chain method to find them.