Minimal-assumption inference from population-genomic data

Daniel B. Weissman[1,2], Oskar Hallatschek[2]
[1]Dept. of Physics, Emory University, Atlanta, GA 30322
[2]Depts. of Physics and Integrative Biology, University of California, Berkeley, CA 94720

October 30, 2016

Abstract

Samples of multiple complete genome sequences contain vast amounts of information about the evolutionary history of populations, much of it in the associations among polymorphisms at different loci. Current methods that take advantage of this linkage information rely on models of recombination and coalescence, limiting the sample sizes and populations that they can analyze. We introduce a method, Minimal-Assumption Genomic Inference of Coalescence (MAGIC), that reconstructs key features of the evolutionary history, including the distribution of coalescence times, by integrating information across genomic length scales without using an explicit model of recombination, demography or selection. Using simulated data, we show that MAGIC's performance is comparable to PSMC' on single diploid samples generated with standard coalescent and recombination models. More importantly, MAGIC can also analyze arbitrarily large samples and is robust to changes in the coalescent and recombination processes. Using MAGIC, we show that the inferred coalescence time histories of samples of multiple human genomes exhibit inconsistencies with a description in terms of an effective population size based on single-genome data.

Introduction

The continuing progress in genetic sequencing technology is enabling the collection of vast amounts of data on the genomic diversity of populations. This data is by far our richest source of information on evolutionary history. The challenge now is to figure out how to extract this information – how to learn as much as possible about the history of populations from modern data sets of many densely-sequenced individuals.

Perhaps the best-established approach to historical inference from genetic data is to fit demographic models to the site frequency spectrum (SFS) (e.g., Gutenkunst et al., 2009; Excoffier et al., 2013). The SFS is easy to calculate, even from very large samples, and demographic models can be fit to it without a specific model of recombination, but it neglects all information about how diversity is distributed across the genome, treating each site independently. This is a natural approximation for sparse sequencing data, where polymorphic sites are generally only very weakly linked, but in large samples sequenced at high coverage much of the information is contained in associations among different polymorphisms. Because SFS-based approaches cannot use this information, they can only reliably determine models with a small number of parameters (Myers, Fefferman & Patterson, 2008; Bhaskar & Song, 2014).

Recently, an alternative approach has been developed in which a hidden Markov model is used to explicitly model recombination along the genome (the "sequential Markovian coalescent", SMC or SMC', McVean & Cardin, 2005; Marjoram & Wall, 2006; Paul, Steinrücken & Song, 2011), vastly increasing the amount of information that can be gleaned from samples of a small number of individuals (Hobolth et al., 2007; Li & Durbin, 2011; Harris & Nielsen, 2013; Sheehan, Harris & Song, 2013; Schiffels & Durbin, 2014; Steinrücken, Kamm & Song, 2015). But this requires modeling coalescence and recombination throughout the analysis, and as a result becomes computationally intractable for large samples. Additionally, for an increasing number of populations, we have multiple genomic sequences but know almost nothing about their natural histories, including plausible historical demographies and patterns of recombination and selection; this is true even for some model organisms (see Alfred & Baldwin, 2015 and other articles in series). It is unclear how accurate these model-based methods are on populations which violate their underlying assumptions.

Here we present a method for Minimal-Assumption Genomic Inference of Coalescence (MAGIC) that infers statistics of the ancestry and history of recombination of an arbitrarily large sample of genomes while making only minimal, generic assumptions about recombination, selection, and demography. MAGIC finds approximate distributions of times to different common ancestors of the sample. These distributions can then be used to fit and test potential models for the history of the population, including the simplest model of a single time-dependent "effective population size", $N_e(t)$. MAGIC strikes a balance between the SFS- and SMC-based approaches, using the distribution of diversity across genomic windows of varying size to generate a description of the single-locus coalescent process that contains far more information than the simple SFS without using a detailed model for recombination.

Results

Approach

The key fact underlying MAGIC is that the relationship between the population parameters (such as recombination rates, historical demography, and selection) and genomic data is entirely mediated by the coalescent history of the sample (Fig. 1, blue and red boxes). We therefore take it as our goal to learn the coalescence time distribution directly from the data without needing a model for the population dynamics. Once one knows the coalescent history, the genomic data contains no additional information about the population parameters, and one can fit or evaluate a wide range of models without having to re-analyze the full data set every time Gattepaille, Günther & Jakobsson, 2016.

Essentially, MAGIC uses the variability in the density of polymorphisms across a wide range of length scales to learn the genome-wide distribution of coalescent histories. This technique is inspired by Li & Durbin, 2011's method, PSMC, and its successor MSMC (Schiffels & Durbin, 2014), which use the fact that SNPs tend to be dense in regions with a long time to the most recent common ancestor (TMRCA), and sparse in regions with short TMRCAs (Fig. 1, top left and middle left). Thus, the distribution of SNPs across the genome can be used to infer the distribution of *local* coalescence times. But while PSMC and MSMC use models for coalescence and recombination to assign a coalescence time to each locus, MAGIC estimates the genomic distribution of times directly, bypassing the need for explicit modeling. To do this, MAGIC first splits the genome into windows and finds the distribution of genetic diversity across windows, i.e., the histogram of the number of polymorphic sites per window of a given length (Fig. 1, bottom left; Fig. 6). This histogram is then used to infer the statistics of *window-averaged* coalescence times (Fig. 1, bottom right; Fig. 7). For small windows, these times are essentially the true single-locus coalescence times, but the inference is noisy due to the small number of mutations in each window. For large windows, the inference is more accurate but the window-averaged distribution is far from the single-locus distribution because windows typically span multiple linkage blocks. The basic trick of MAGIC is that rather than choosing one window length, it integrates the information gathered from a wide range of different window lengths to find the small-length limit – the true single-locus distribution (Fig. 1, center right).

MAGIC's accuracy is comparable to the state-of-the-art model-driven method MSMC (Schiffels & Durbin, 2014) on small data sets that conform to MSMC's assumptions (see "Single diploid samples" in "Results"), but it also enables analyzing data from populations with features (such as gene conversion, ancestral structure, or linked selection) that violate those assumptions, as well as larger samples. MAGIC's algorithm, described in "Methods" below, is designed to be as simple and modular as possible, allowing one to analyze large samples and incorporate additional assumptions in situations where more information is available. This also enables the inference of a wide range of summary statics, including the distribution of map lengths of blocks of identity-by-descent across the genome (Ralph & Coop, 2013). Finally, MAGIC can use the dependence of the window-averaged statistics on window size to learn about the rate of recombination and the variation in recombination rate and the coalescent process across the genome.

Representing coalescence time distributions

For single diploid samples, the coalescent history is completely described by a single time at each locus. Thus the pair-wise coalescence time distribution could equivalently be described by the hazard function (the pairwise coalescence rate, as in Schiffels & Durbin, 2014) or its reciprocal (the "effective population size" $N_e(t)$, as in Li & Durbin, 2011). However, when estimating the distribution from noisy data, the procedure that minimizes the error for one description will not in general minimize the error for the others. We focus on estimating the distribution, primarily because it naturally generalizes to arbitrary sets of coalescence tree branch lengths for larger samples (see below). This also lets us emphasize that the idea that the coalescent can be described by a single $N_e(t)$ is a model that can be tested. For plotting, we show cumulative distributions rather than densities so that we can plot the actual coalescent histories of samples (black curves in Figs. 2 and 3), which consist of discrete sets of events, and also because the density estimates are very poorly constrained by the data (Ralph & Coop, 2013).

Single diploid samples

To validate our approach, we have tested MAGIC on single diploid samples generated under a range of coalescent models simulated with `ms` (Hudson, 2002). MAGIC accurately infers the distribution of coalescent times from samples with map length and polymorphism density similar to that of a human genome (Fig. 2, top two rows, solid curves; see Methods and Fig. 9 for detailed parameters). Although the simulation parameters were conducted under MSMC's assumed model, MAGIC performs nearly as well (Kolmogorov-Smirnov distance to the true distribution of $4 - 13\%$ for the simulations shown, compared to $4 - 11\%$ for MSMC). Both methods tend to smooth out sharp transitions in the coalescence distribution as a consequence of regularization. The distribution of map lengths of blocks of identity by descent (IBD) can be inferred with very high accuracy (Fig. 2, top two rows, dashed curves), improving on MSMC, which sometimes overestimates the amount of very deep coalescence, and correspondingly erroneously estimates a large number of very short blocks. The part of the block-length distribution estimated by MAGIC is complementary to the very long blocks that can be observed directly (as in, e.g., Ralph & Coop, 2013). MAGIC is also accurate on genomes simulated with `ms` under a model in which recombination is dominated by gene conversion (Fig. 2, bottom row); this can be seen as loosely corresponding to a primarily asexual population, with gene conversion representing homologous recombination. In this case, MSMC's

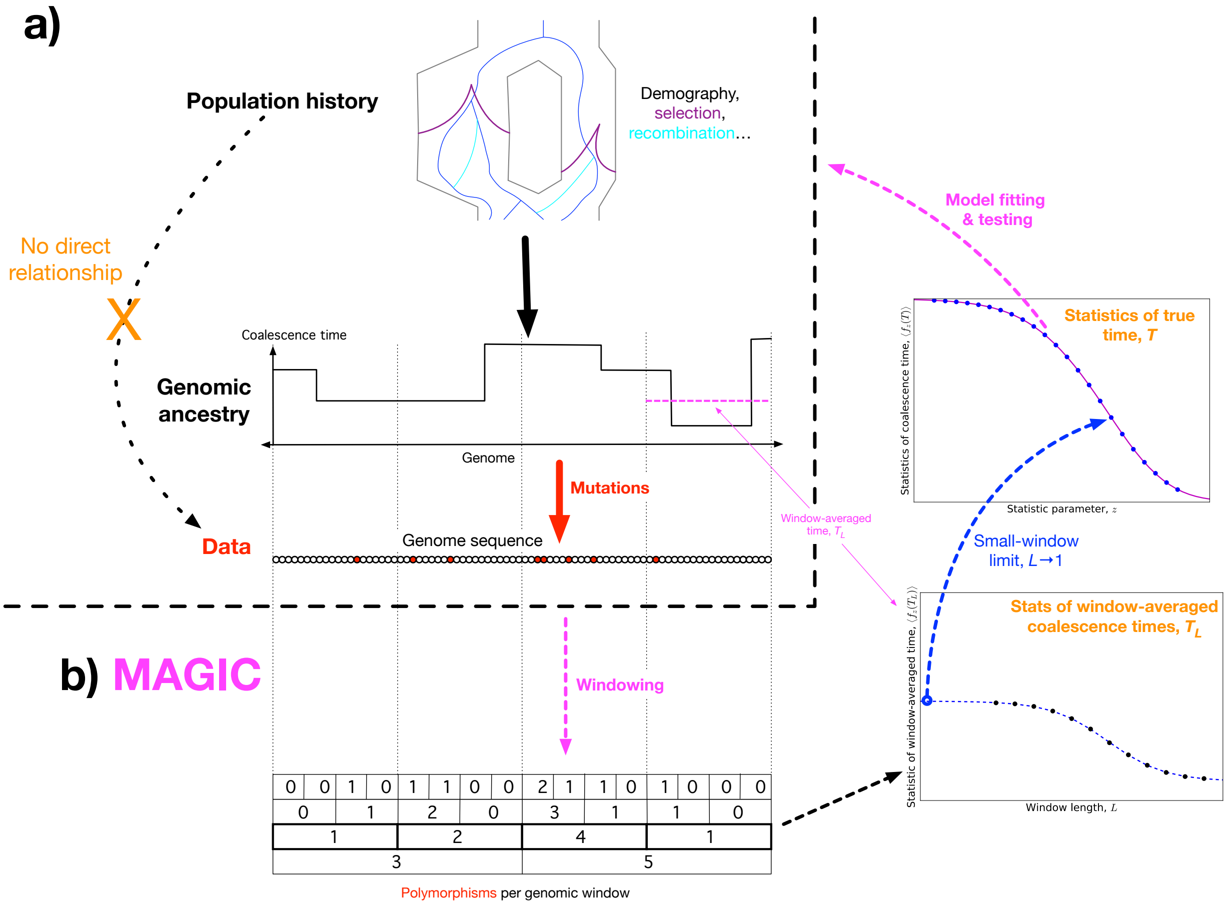

Figure 1: **a)** Concept: we would like to infer the history of the population (top) from the sequence data (bottom), but the causal connection between the two is entirely mediated by the coalescent history of the sample (middle). This suggests that it should be possible to extract much of the coalescent information from the data without making strong assumptions about the population dynamics. **b)** Schematic algorithm: MAGIC first splits the sample into small windows and counts the polymorphisms within each window, then progressively merges pairs of adjacent windows together (bottom left). For each window length L, the histogram of window diversity is used to calculate statistics of the window-averaged coalescence time T_L (bottom right). Taking the limit of these as L goes to 1 gives the statistics of the true coalescence time T (top right), which are then used to fit and test models for the underlying dynamics.

recombination model breaks down and MAGIC's inferences are more reliable.

Larger samples

For samples of more than two haplotypes, the coalescent history at each locus is described by a tree, rather than a single time. The space of possible trees grows very rapidly with the sample size, so that even with long genomes it is impossible to directly estimate the full distribution. Instead, MAGIC infers the distribution of some small set of features of the trees, such as mean pairwise distance and total branch length, chosen either because they are important in and of themselves or because they are sufficient statistics for some model of the coalescent process. For example, MAGIC can fit the basic time-dependent effective population size model by estimating the distribution of pairwise coalescence times, and then check whether the fitted model correctly predicts the distributions of other tree features.

We test this approach on a sample of six haplotypes from a recently admixed population simulated with `ms` (demographic parameters in the last column of Figs. 2 and 9). MAGIC accurately estimates the distributions of pairwise coalescence times, the total branch length of the full coalescent trees, and the total length of the tips of the coalescent trees (Fig. 3, top row). We use the first distribution to fit a time-dependent effective population size model, and then compare its predictions to MAGIC's inferences for the latter two distributions. The large differences show that the model is not a good description of the population history. MSMC's estimate of the pairwise distribution is relatively inaccurate – perhaps unsurprisingly, since it is based on an effective population size model.

MAGIC's running time on large samples is dominated by the time to read all the data through memory, so it grows only linearly with sample size, meaning that the method can be run on essentially arbitrarily large samples. MAGIC accurately estimates the distributions of pairwise times, total branch length, and total tip length in a sample of 100 haplotypes from the same admixed population (Fig. 3, bottom row). (Curves are not shown for MSMC because it cannot analyze large samples; a sample size of eight from this population caused it to crash.) The total branch length distribution remains different from that predicted by the effective population size model, but the tip distribution is close, differing mainly in the right tail, indicating that the model works better for recent times than ancient ones, as expected.

In the example above, we have only inferred the distributions of pairwise coalescence times, total branch lengths, and total tip lengths. But to be able to consider a wide range of models for the population, one must have estimates for a wide range of statistics. MAGIC can infer the distribution of the total length of any specified set of tree branches. For a given set, MAGIC first filters the polymorphisms in the original data for those that correspond to mutations on the desired branches, and then proceeds with the same analysis as in the basic case of a single diploid sample. For example, to find the distribution of total branch length, MAGIC analyzes the genomic distribution of all polymorphic sites, while to find the distribution of total tip length, it only looks at singletons. For data with no linkage information, the statistics that can be estimated are closely related to the site-frequency spectrum (SFS), but while the SFS just gives estimates of the *means* of the lengths of different sets of branches, MAGIC infers full distributions.

Human data

We use MAGIC to analyze human sequences from Complete Genomics' "69 genomes" data set (Drmanac et al., 2010). On individual diploid samples, the inferred coalescence time distributions (plotted in the upper left panel of Fig. 4 as "effective population sizes") are similar to those obtained

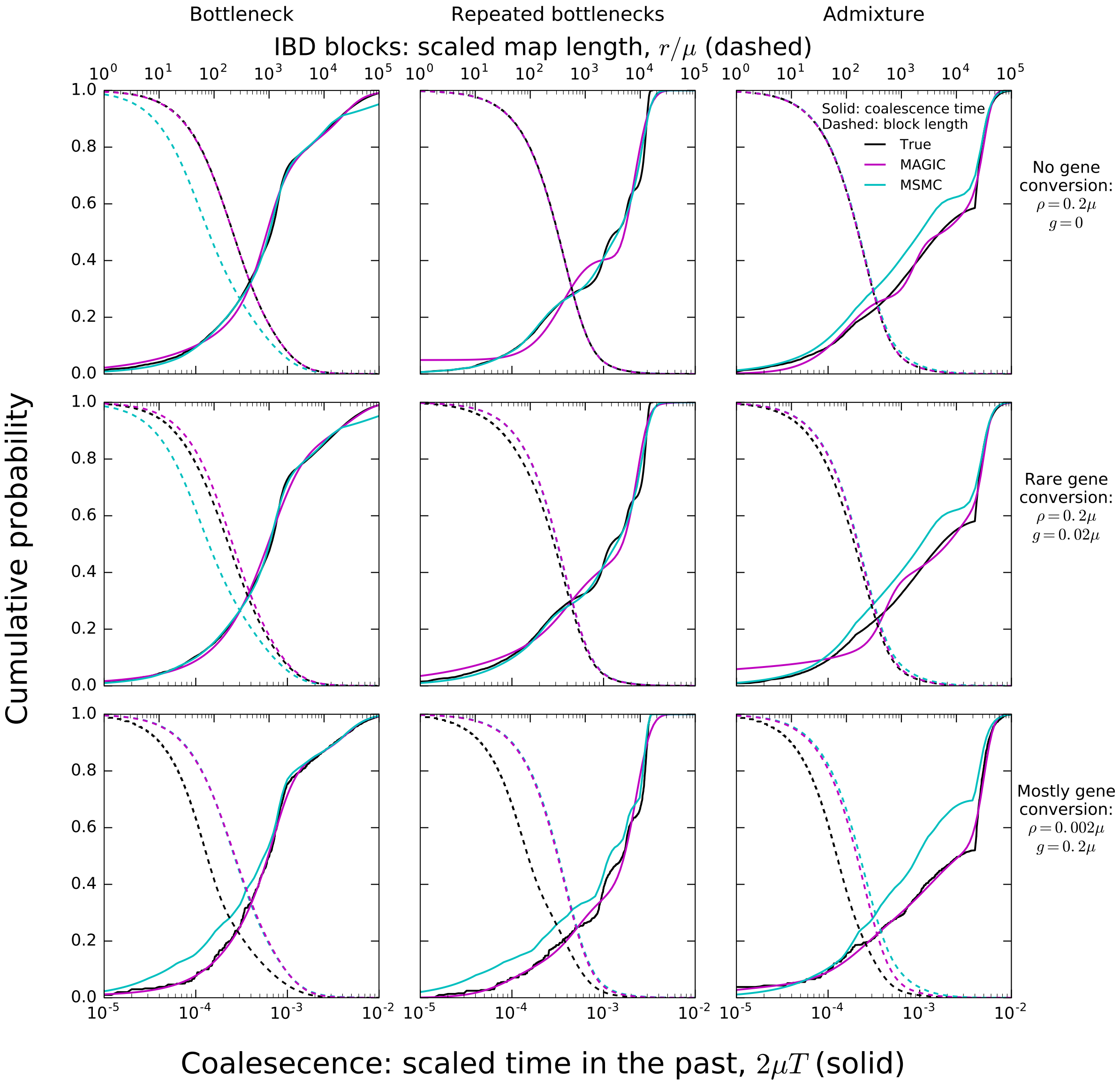

Figure 2: MAGIC accurately infers the distribution of coalescence times (solid curves) and lengths of blocks of identity-by-descent (IBD, dashed curves) for pairwise data simulated with `ms` under several demographic scenarios. When crossovers are frequent and gene conversion is rare (top two rows), MAGIC and MSMC are comparably accurate for coalescence times. MAGIC very accurately infers the IBD block length distribution, while MSMC sometimes is inaccurate (e.g., "Bottleneck" scenario). For frequent gene conversion and rare crossovers (bottom row), the details of the gene conversion process have a strong effect on the IBD block lengths, and neither method can infer their distribution, but MAGIC can still infer the coalescence times. All simulations are of a genome consisting of 100 independent chromosomes, each 10^7 base pairs long, with per-base mutation rate μ and present population size N_0 such that $N_0\mu = 10^{-3}$. Recombination is via crossovers occurring at rate ρ per base, and via gene conversion being initiated at rate g per base with mean tract length $\lambda = 200$bp. See Methods for values of other simulation parameters.

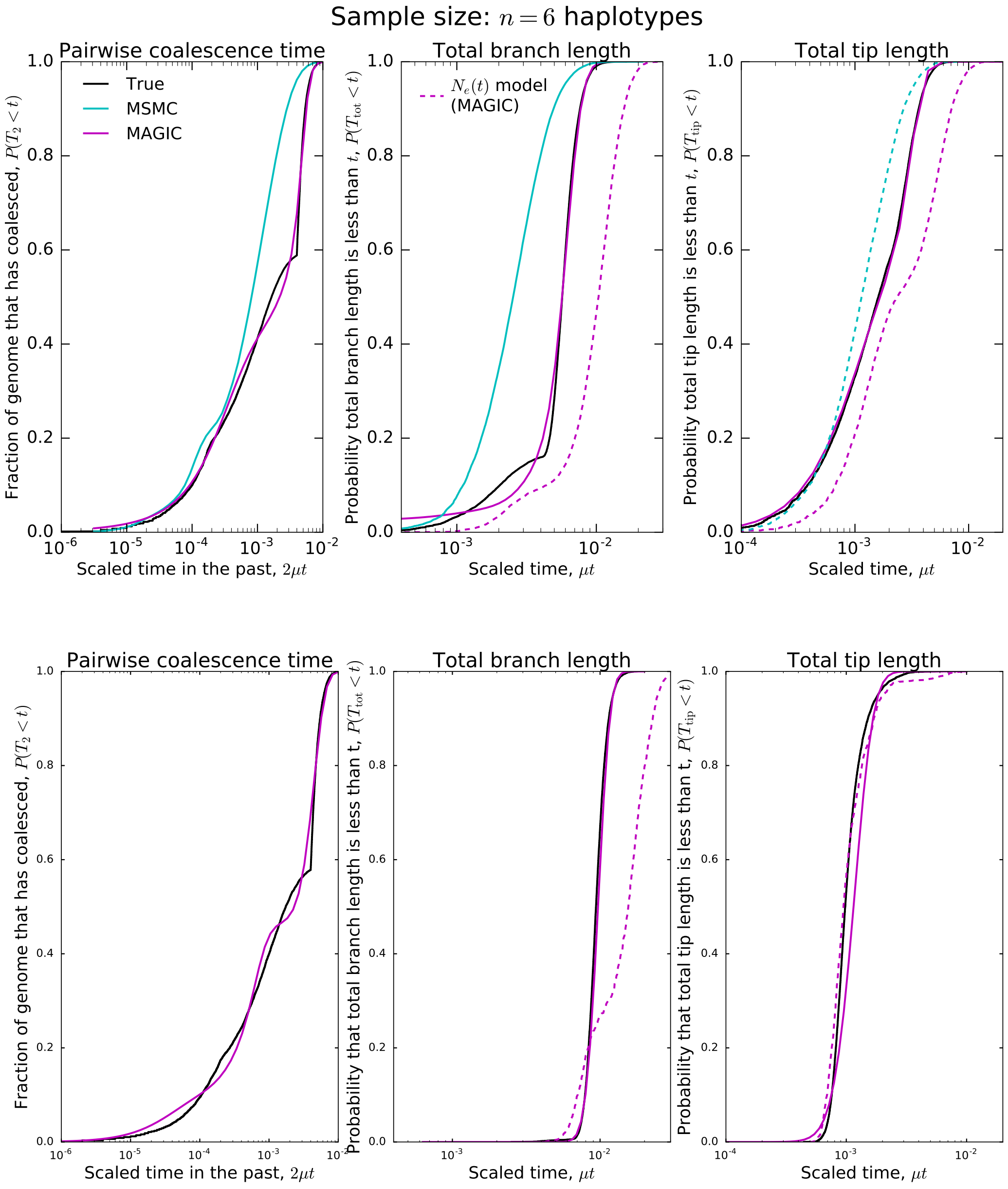

Figure 3: Coalescence time statistics for a simulated population with ancestral structure inferred from larger samples. Top row: sample size $n = 6$ haplotypes. MAGIC accurately estimates the distribution of the pairwise coalescence time, the total branch length, and the total lengths of the tips of the branches. Comparing the estimated distributions reveals that the coalescent process cannot be described by an instantaneous rate $1/N_e(t)$. Bottom row: sample size $n = 100$ haplotypes. Comparing the pairwise times and the total branch lengths still shows strong signs of population structure, but the tip lengths are close to the pairwise prediction, suggesting that the structure has recently disappeared.

with MSMC, differing mainly in the tails where the data is limited. The distribution inferred by MAGIC is closer to that inferred by MSMC using larger samples (Schiffels & Durbin, 2014, Supplementary Figure 7). We also ran MAGIC on a sample comprising all 9 unrelated Yoruban individuals in the data set (i.e., a sample of 18 haplotypes, much larger than possible with MSMC) and compared the inferred tree statistics to those predicted by the Kingman coalescent given the inferred pairwise coalescence time distribution (Fig. 4, bottom panel). The total branch length is close to that predicted by the Kingman model, but the tips of the trees are substantially longer, indicating that the inferred effective population size is too small in the recent past, i.e., it misses the recent population growth. Thus, by analyzing larger samples, MAGIC is able to probe more recent times than can be seen with pairwise comparisons. However, the observed discrepancy persists back to times at which $> 10\%$ of the genome of single individual has coalesced, by which point there should be more than enough data to accurately estimate the pairwise coalescence time distribution. This suggests that the difference is not just due to limited resolution, but also to inaccuracies of the $N_e(t)$ model, such that the pairwise coalescent does not describe the full coalescent process.

Inferring recombination rates

MAGIC's coalescence-time inference is designed to be robust to the form of recombination, but it can also be used to learn about recombination. To do this, rather than simply taking the small-window limit of the window-averaged coalescence time statistics, one can look at how they change as a function of window size. In general, besides the small-window limit in which almost all windows lie within IBD blocks, there should also be a long-window limit in which almost all windows contain many IBD blocks. In between these two there is a transitional regime where the window length lies within the bulk of the distribution of IBD block lengths; finding this transitional length gives an estimate of the recombination rate.

Because the transition from the small-window to the large-window limit is not very sharp, this estimate of the recombination rate is very rough. A more precise estimation requires a specific model of recombination and coalescence, like the one used by MSMC. But even if one does not have a good model for the dynamics of a population, one can make the assumption that all autosomes are experiencing roughly the same dynamics, whatever they are. (This assumption is implicit in all demographic inference from full genomes.) In that case, the dependence of the statistics of the window-averaged coalescence time on window length should be similar across autosomes, and differences in average recombination rates across chromosomes should be detectable as rescalings of the window lengths. MAGIC can therefore use these rescalings to precisely estimate *relative* recombination rates.

As an example of this approach, we analyze each autosome across the 9 unrelated Yoruban individuals in the data set. We find that they all show similar dependence of the statistics of the window-averaged coalescence time on window length (Fig. 5, top left). Up to a rescaling in length, the autosomes appear very similar (Fig. 5, top center), with the exception of 19. The collapse of the remaining 21 autosomes suggests that they differ primarily in the amount of very recent coalescence (rescaling the heterozygosity) and in average recombination rates (rescaling the window lengths). The scaling factors for the window lengths therefore are an estimate of relative recombination rates, and are indeed very close to values measured by Kong et al., 2002 (Fig. 5, bottom), with the exception of chromosome 19, as expected.

It is no surprise that chromosome 19 is an outlier in coalesecence: it has a much higher gene density than the other chromosomes (Grimwood et al., 2004), and is therefore likely to have a much higher fraction of loci under selection and affected by linked selection (Hernandez et al., 2011). However, the other autosomes do not have identical gene densities, and there are several

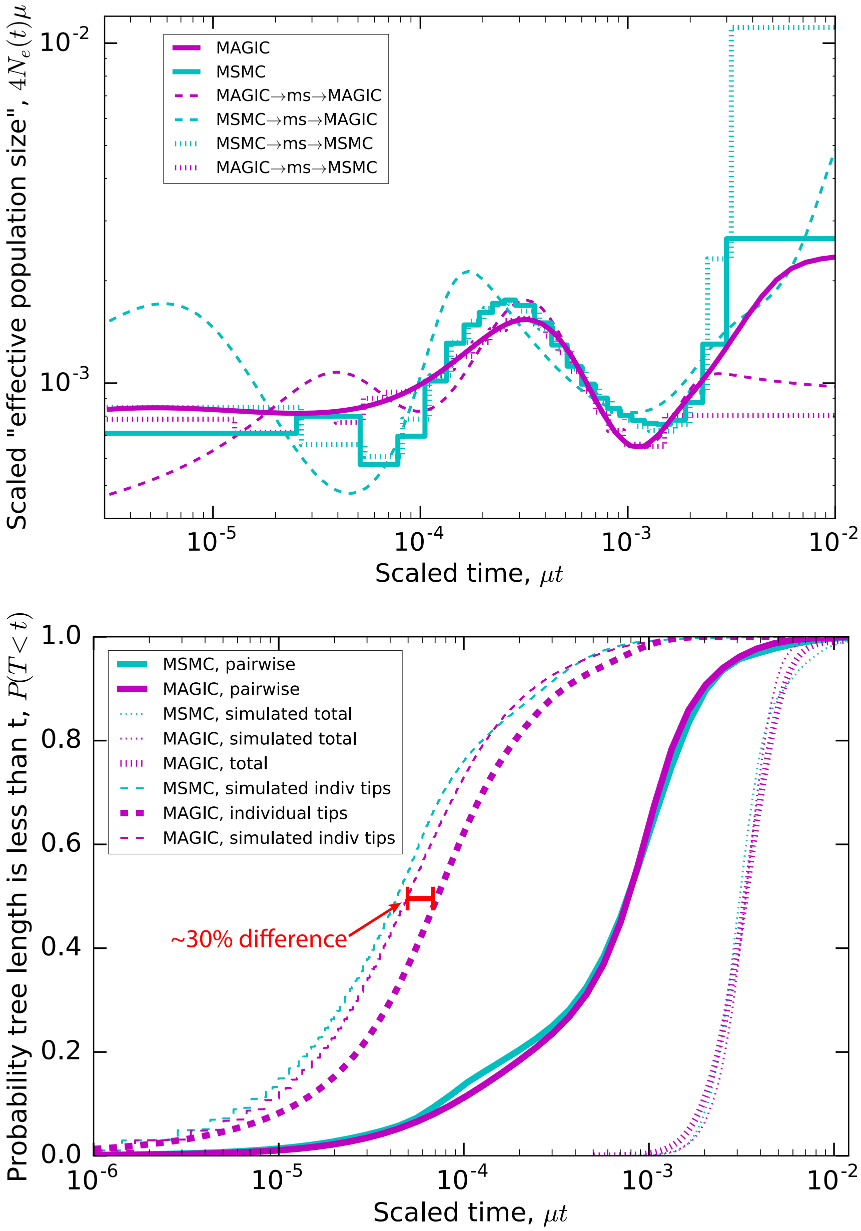

Figure 4: Inferred evolutionary history of Yoruban individuals. Top: Inferred "effective population size" $N_e(t)$ for YRI individual NA18502. MAGIC and MSMC infer similar effective population sizes (solid curves), differing mostly in the very recent and distant past where there is limited data. Running both methods on simulations (dashed and dotted curves) shows that MSMC is more accurate in the recent past, while MAGIC may be slightly more accurate in the distant past. Bottom: Distributions of coalescence times for a sample of the 9 unrelated Yoruban individuals from the 69 Genomes data set (Drmanac et al., 2010). These are compared to samples of 9 individuals simulated under the $N_e(t)$ inferred by MSMC from NA18502 and by MAGIC from the heterozygosities of all 9 Yorubans (solid curves). The pairwise $N_e(t)$ accurately describes the distribution of total branch lengths (dotted curves), but underestimates the tip lengths by $\sim 30\%$ (dashed curves).

large regions with unusual patterns of diversity, such as the MHC locus and flanking regions on chromosome 6. Indeed, even after rescaling, there is still more residual variation in coalescence across the 21 similar autosomes than would be expected by chance (Fig. 5, top right). This variation must be due to non-demographic factors driving coalescence; the fact that they are readily detectable suggests that one should be cautious in interpreting the details of inferences by MSMC and other demographic inference methods in terms of demography.

Discussion

The MAGIC algorithm bridges the gap between fast but limited SFS-based approaches to demographic inference and model-based approaches that are limited to small sample sizes, allowing far more information to be extracted from large, high-coverage samples. To see the difference between MAGIC and an SFS-based approach, consider the information that can be gained from sites with singleton polymorphisms. Under an approach that treats sites independently, these can be summarized by one number – their genomic frequency – which can only be used to estimate one number – the mean total tip length of coalescent trees. MAGIC, in contrast, also considers how clustered the sites are over a wide range of lengthscales, allowing it to estimate not just the mean, but the whole distribution of total tip lengths. MAGIC in this sense is similar in spirit to Bunnefeld, Frantz & Lohse, 2015 and Reddy et al., 2016's "blockwise SFS" approach, but differs in that it does not require prior knowledge about recombination or the range of IBD block lengths. In addition, because MAGIC converts the genomic distribution of diversity into the statistics of single-locus coalescent times, it can be checked or fit with single-locus coalescent simulations, which are much less computationally intensive than multi-locus ones.

While existing methods rely on fitting simplified demographic models while neglecting selection (Schraiber & Akey, 2015), MAGIC estimates makes no assumptions about whether coalescence is driven by demography or selection, and only minimal assumptions about mutation and recombination. We hope that this will make MAGIC useful as a first-pass analysis of genomes from species whose natural histories are not already well-known, with its results informing the choice of more detailed, model-based methods that use additional information outside of the sample sequences. Even for populations for which there are good models, the minimal-assumption approach has advantages. Because MAGIC has a modular structure and is not tailored to a specific population model, it can be used to quickly analyze many populations with very different dynamics, with each population's model incorporated in just the last step of the analysis. Similarly, for any given population, MAGIC can calculate many different statistics describing coalescence and recombination to answer multiple questions about the historical dynamics. Finally, not using any explicit model of coalescence and recombination keeps MAGIC's algorithm simple enough that it runs quickly even on very large sample sizes, and that users familiar with Python can understand and modify it.

There are a number of potential modifications to MAGIC that users could make. At a minimum, there are likely to be technical improvements to the estimation methods that would allow it to get more information out of the data More interestingly, the range of statistics estimated by MAGIC could be extended. In particular, MAGIC currently infers the distributions of features of coalescent trees that can be found from unphased, unpolarized polymorphism data, but it could be extended to take advantage of this extra information when available. It would also be possible to extend MAGIC so that it would infer *joint* distributions of different coalescence times, rather than just all the marginal distributions. This would greatly increase the amount of information that could be extracted from extremely large data sets such as are likely to be available in the near future.

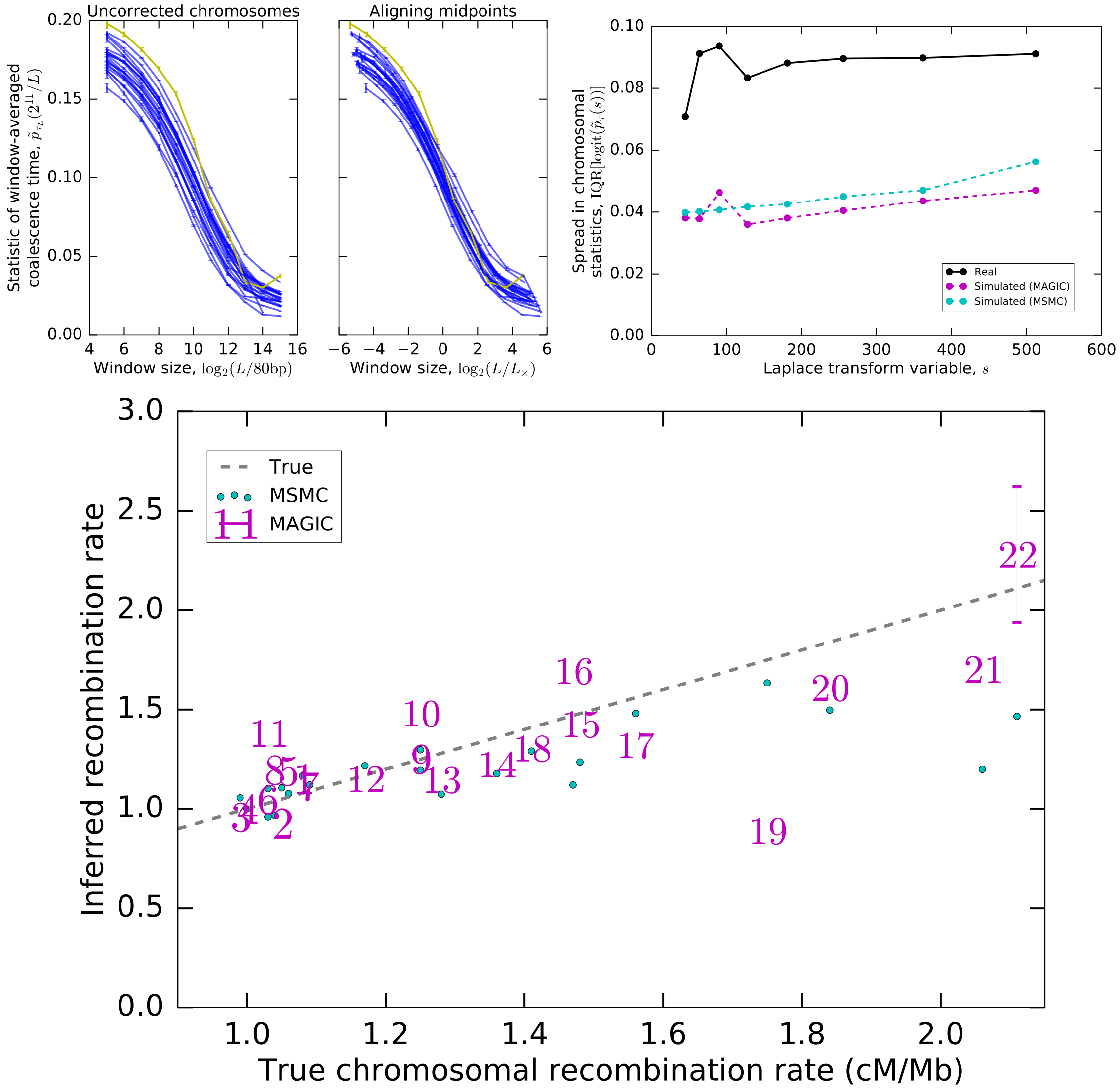

Figure 5: Estimating relative recombination rates from the genomes of Yoruban individuals. Top left: One value of the Laplace transform of the window-averaged coalescence time distribution as a function of window size for each autosome. Top center: Much inter-autosomal variation can be explained by variation in recombination rates: the curves are similar under a rescaling of window lengths, except for chromosome 19 (yellow), which appears to have a different pattern of coalescence. Top right: There is still substantial variation in the asymptotes that cannot be explained by variation in recombination rates, and is more than expected from intrinsic coalescent stochasticity. Plot shows the interquartile range of the Laplace transform of the coalescence time distribution across chromosomes for the actual data as well as simulations of the pairwise coalescent histories inferred by MAGIC and MSMC (Fig. 4). Bottom: The rescaling of window sizes needed to align the different autosomes gives an estimate of their relative recombination rates which is very close to the values obtained by Kong et al., 2002 ("True"). For chromosomes other than 22, the inferred error bars are smaller than the size of the markers.

Methods

Approach

A sample set of genomes will comprise many blocks of sequence with different coalescent histories; by looking at the distribution of genetic diversity across blocks, one can estimate the coalescence time distribution of the population the sample was drawn from. Li & Durbin, 2011 and Schiffels & Durbin, 2014 try to do this by finding the exact boundaries between the blocks using a hidden Markov model. However, this is only easy to do when mutation rates are much larger than recombination rates, which is generally not the case, and describing every block becomes impractical for larger sample sizes as the number of blocks proliferates. Instead, we simply divide the genome into windows of a fixed length L, and consider the distribution of histories of windows. MAGIC estimates the distribution of a single coalescence time (i.e., coalescence tree statistic) T across genomic positions x. For single diploid samples, $T(x)$ is the total branch length (twice the time to the most recent common ancestor at position x) and completely characterizes by the coalescent history. For larger samples, MAGIC can be used to estimate multiple statistics one at a time.

The diversity (e.g., heterozygosity in a single diploid sample) in a window of length L starting at position x_0 depends only on the *window-averaged coalescence time*, T_L defined as

$$T_L \equiv \frac{1}{L} \sum_{x=x_0}^{x_0+L-1} T(x). \tag{1}$$

If L is smaller than most block lengths, then windows will typically lie within blocks, and the cumulative distribution P_{T_L} of T_L will be close to the cumulative distribution P_T of T. For very large L, each window will average over many blocks, and P_{T_L} will have a narrow support around the mean of P_T. Usually, there will be a wide range of intermediate values of L for which windows lie inside long blocks but cover multiple short blocks.

Given that P_{T_L} approaches P_T, as L decreases, one might be tempted to take L to be as small as possible, but the problem of course is that we cannot see T_L directly; we need to infer it from the number of SNPs in the window. Given T_L, and assuming a constant mutation rate μ per base, the number of SNPs will be approximately Poisson-distributed with mean $\mu L T_L$. (Here we are assuming that μT is always small enough that each base has only a small chance of having mutated; this allows us to approximate the underlying sum of binomially-distributed random variables with a Poisson that depends only on T_L.) The smaller L is, the lower the signal-to-noise ratio will be and the less power we will have to distinguish different values of T_L. Thus, we expect that we will get the most information about P_T from an intermediate value of L, and should be able to do better still by integrating information from multiple values of L.

The total probability that there are n SNPs in a window of length L is by averaging the Poisson distribution over all possible values of T_L:

$$P_L(n) = \left\langle e^{-\tau_L} \tau_L^n \right\rangle_{\tau_L} / n!, \tag{2}$$

where τ_L is the window-averaged coalescence time scaled such that it is equal to the expected number of SNPs in the window: $\tau_L \equiv \mu L T_L$. P_L is a Poisson mixture distribution, with mixing distribution given by P_{τ_L}, the cumulative distribution function of τ_L, i.e., the fraction of the genome that is expected to have coalesced by a given scaled time. Our immediate goal is to estimate P_{T_L} from the observed SNP count distribution $\widehat{P_L}$.

If only a small, random fraction of the genome has been sequenced, this is the classic statistical question of estimating the mixing distribution of a Poisson mixture from a finite number of draws;

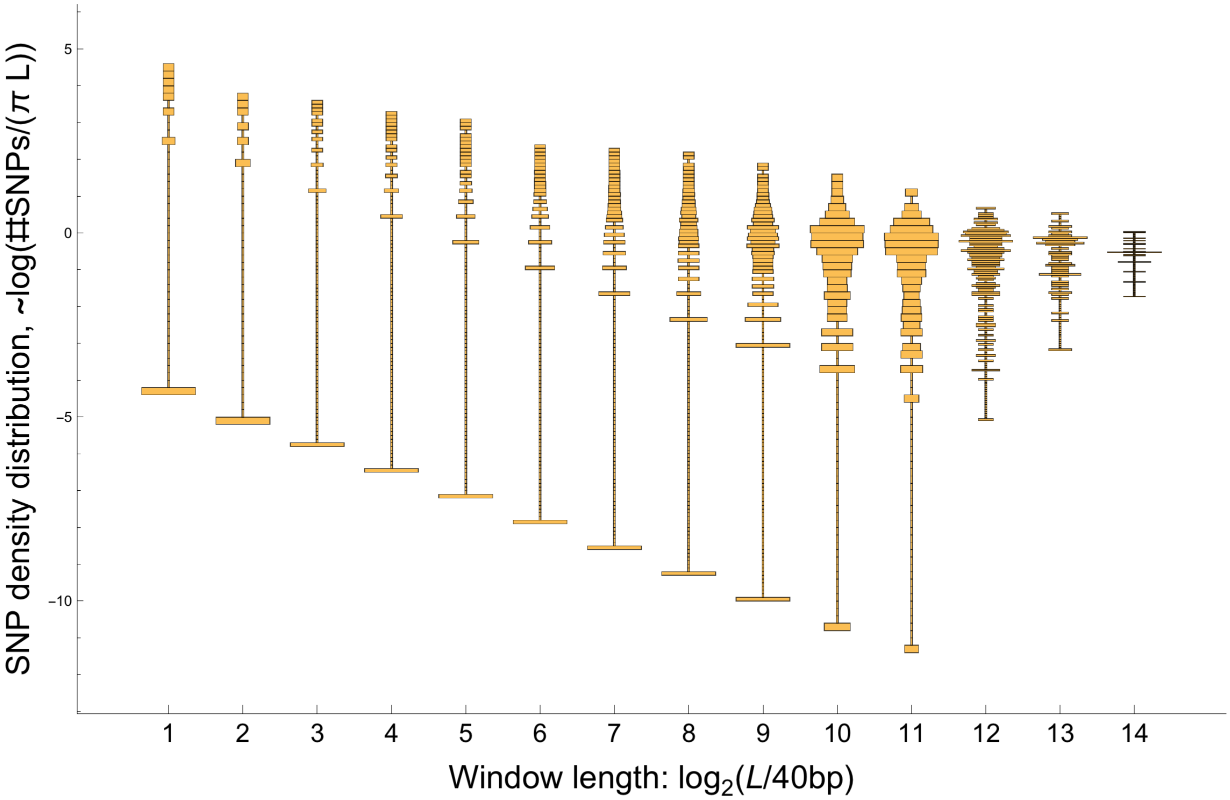

Figure 6: Distribution of SNP densities across windows at different window lengths, normalized by mean value. Data from NA18501 (YRI) chromosome 1. The bottom bars are windows with no SNPs, which are artificially put at finite values on the log scale. At short lengths, the distribution is bunched near zero, while at long lengths it is bunched near the mean. The best spread is in between, at kb scales.

in this case, the maximum likelihood estimate for P_{τ_L} is a step function with a number of steps equal to roughly half the number of distinct values of n (number of SNPs in a window) observed across the genome (Simar, 1976). This maximum likelihood function is difficult to find, however, and in any case we are typically in a slightly different situation, in which most of the genome has been sequenced and the sampling noise in the number of windows with a given T_L is small. More importantly, we are interested primarily in the true distribution P_T rather than the window-averaged distribution P_{T_L}, so we only want to estimate features of P_{T_L} that also describe P_T.

Laplace transforms

Rather than trying to estimate the full distribution P_{T_L}, we will instead estimate a set of statistics describing it. The Laplace transform $\tilde{p}_{\tau_L}(z) \equiv \langle e^{-z\tau_L} \rangle$, evaluated at a set of points $\{z_j\}$, is a natural choice, as it is closely related to the diversity distribution: (2) shows that $(-1)^n P_L(n)$ is the n^{th} Taylor coefficient of $\tilde{p}_{\tau_L}(z)$ about $z = 1$. This has two implications. First, the similarity to the proportion of homozygous windows $P_L(0) = \langle e^{-\tau_L} \rangle$ means that the Laplace transform has a natural interpretation as an estimate for the proportion of windows of length L that would be homozygous if the mutation rate were multiplied by z. (It is also closely related to the distribution of lengths of IBD blocks – see below.) Second, we can quickly and easily estimate $\tilde{p}_{\tau_L}$ using the plug-in estimator:

$$\widehat{\tilde{p}_{\tau_L}(z)} = \sum_{n=0}^{\infty} \widehat{P_L(n)}(1-z)^n. \tag{3}$$

The estimate (3) of the Laplace transform will be accurate for z close to 1, but will blow up for large z – we cannot accurately estimate the amount of very recent coalescence. To make this precise, we need to estimate the error in (3). We use Ghosh, Hwang & Tsui, 1983's formula (5) to estimate the value of T_L for every window from the observed number of SNPs, with the adjustment that for the K_0 windows with 0 SNPs, we use $\widehat{T_L} = \log(2)/(\mu K_0)$ rather than 0. We then estimate the error introduced by the stochastic mutation accumulation process under the resulting $\widehat{P_{T_L}}$. The accuracy of both (3) and the error estimate could be improved by using more sophisticated estimators, but the current ones are the easiest to compute and appear to give reasonable values for all the simulated data that we have tried.

Combining lengthscales

To combine information from different window lengths, we need to correct for the increase in window-wide mutation rate μL. We can therefore consider the quantity $\tilde{p}_{\tau_L}(z/L)$ as a function of L, holding z fixed, as shown in Fig. 7. When P_{T_L} is nearly independent of L, this quantity should be nearly constant. (To see this, note that $\tilde{p}_{\tau_L}(z/L) = \left\langle e^{-z\mu T_L} \right\rangle$, with no explicit L dependence.) This is the case for very large L, when each window averages over many coalescent blocks, and for very small L, where each window falls within a coalescent block and P_{T_L} approaches P_T, the distribution we are interested in. We therefore fit a sigmoid curve (specifically, Richards' curve) to $\widehat{\tilde{p}_{\tau_L}}(z/L)$ as a function of $\log(L)$,

$$\tilde{p}_{\tau_L}(z/L) \approx a_z + \frac{b_z - a_z}{\left[1 + (L/L_\times)^{-c_z}\right]^{1/\nu_z}}, \text{ with } a_z, b_z \in [0,1];\ c_z, d_z, \nu_z > 0;\ L_\times > 1, \tag{4}$$

and take the left asymptote a_z as an estimate of $\tilde{p}_T(\mu z)$. The right asymptote b_z is the long-window limit, and can therefore be estimated directly from the genomic density of SNPs ϖ: $b_z = e^{-\varpi z}$. To fit the remaining parameters, we us `curve_fit` function in SciPy's `optimize` package.

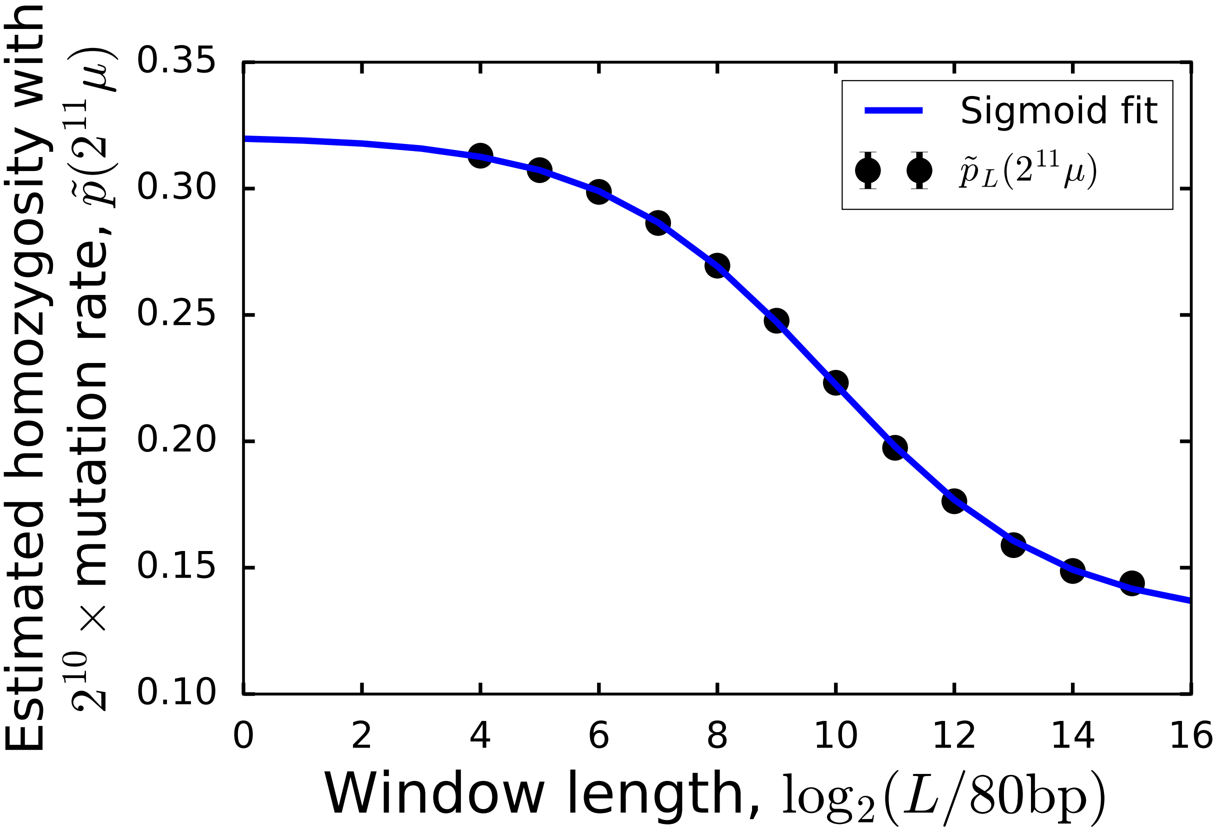

Figure 7: The window-averaged coalescence time distribution changes with window size. For short windows, the distribution approaches the true coalescence time distribution, while for long windows it approaches a distribution sharply peaked at the mean coalescence time. The length ∼ 5kb at which it begins to deviate from the small-window limit gives an estimate of the density of recombination events. Data is from YRI individual NA18502, chromosome 1.

`curve_fit` also returns a standard error $\sigma^2(\widehat{\tilde{p}_T(\mu z)})$, estimated under the assumption that the errors in $\widehat{\tilde{p}_{\tau_L}(z/L)}$ are independent for different values of L. This is obviously not true (all estimates are from the same set of mutations), but even for lengthscales that are only separated by factors of 2, the correlations in the error appear to be small in simulations, giving final error bars of roughly the right magnitude. While a_z contains the information about the single-locus coalescent, the values of other parameters, particularly $L_\times$, are informative about recombination and the relationships among loci – see "Inferring recombination rates" in main text.

The sigmoid form (4) is flexible enough to fit all of the simulated and real data that we have examined, but we only trust our estimate of $\tilde{p}_T(\mu z)$ if the data are close enough to the left asymptote so that the estimate is not very sensitive to the choice of functional form. Effectively, this means that the estimate is close to $\widehat{\tilde{p}_{\tau_L}(z/L)}$ for the smallest L for which the estimated error bars are small, with corrections based on the next few higher length scales. But this smallest L depends on z, so while for any given value of z only a few lengthscales are important, every lengthscale is important for estimating the Laplace transform for some z. Short lengthscales are useful for small z (i.e., long coalescence times), while long lengthscales are useful for large z (short coalescence times).

Coalescence-time distributions

Once we have estimates for the Laplace transform of the coalescence time distribution at a set of $\{z_j\}_{j=1,\ldots,J}$, we would like to invert the transform to obtain p_T. Unfortunately, this is a fundamentally hard problem (Epstein & Schotland, 2008), and we need to use some kind of regularization. We do this by assuming that P_T can be written as a mixture of gamma distributions:

$$p_T(t) = \sum_{i=1}^{\lfloor (J+1)/3 \rfloor} c_i \frac{t^{k_i - 1} e^{-t/\theta_i}}{\Gamma(k_i)\theta_i^{k_i}}, \tag{5}$$

where $\sum_i c_i = 1$ and all c_i, k_i, and θ_i are positive. This can also be extended to include a possible point mass at $t = 0$ when estimating the distribution of features that we expect to be exactly 0 for some coalescent trees, e.g., the length of branches that are ancestral to 5 haplotypes out of a sample of 10.

The gamma mixture distribution has the computational advantage of having a simple analytic Laplace transform:

$$\tilde{p}_T(z) = \sum_{i=1}^{\lfloor (J+1)/3 \rfloor} c_i (1 + \theta_i z)^{k_i}, \tag{6}$$

with an additional constant term if a point mass at $t = 0$ is included. This means that we can simply fit (6) to the values $\left\{\widehat{\tilde{p}_T(z_j)}\right\}$ without having to deal with inverse Laplace transforms directly. We use the L-BFGS-B algorithm implemented in SciPy to find the $\{c_i, k_i, \theta_i\}$ that minimize the scaled squared error $\sum_j \left(\tilde{p}_T(z_j) - \widehat{\tilde{p}_T(z_j)}\right)^2 / \sigma^2(\widehat{\tilde{p}_T(z_j)})$. The gamma mixture form is flexible enough to fit all the data that we have tried (see, e.g., Fig. 8).

Block-length distributions

Identical-by-descent (IBD) blocks are stretches of the genome that have not undergone recombination since the common ancestor of the block. The distribution of block lengths can be used to infer patterns of relatedness and ancestry (Li & Durbin, 2011; Ralph & Coop, 2013), but it is hard to measure except for long blocks with very recent common ancestors, because the recombination events are not directly observable. Under the standard assumptions that coalescence is mostly driven by neutral processes (rather than linked selection) and that recombination primarily occurs via crossovers, the distribution of the genetic map lengths of these blocks across the genome is closely related to the Laplace transform of the coalescence time distribution:

$$P(r_{\text{block}} > r) = -\tilde{p}'_T(r)/E[T] = \tilde{p}'_T(r)/\tilde{p}'_T(0). \tag{7}$$

We can estimate the block-length distribution in Morgans simply by approximating the derivative of $\tilde{p}_T$, a much easier problem than inverting the transform. If we want to convert map lengths to numbers of bases, we need to estimate the crossover frequency per base. The dependence of $\tilde{p}_{\tau_L}$ on L (i.e., $L_\times$ in (4) above) provides a rough estimate; MSMC gives a more precise one. (7) gives the distribution of lengths of blocks that are not interrupted by even "ghost" recombination events where the coalescent tree does not change (Marjoram & Wall, 2006's "R class" of events). In the simplest case of a single diploid sample from a well-mixed population evolving neutrally under the Kingman coalescent, these ghost events can be ignored by simply scaling all recombination rates by 2/3, but in general they must be included, even though they are not directly observable.

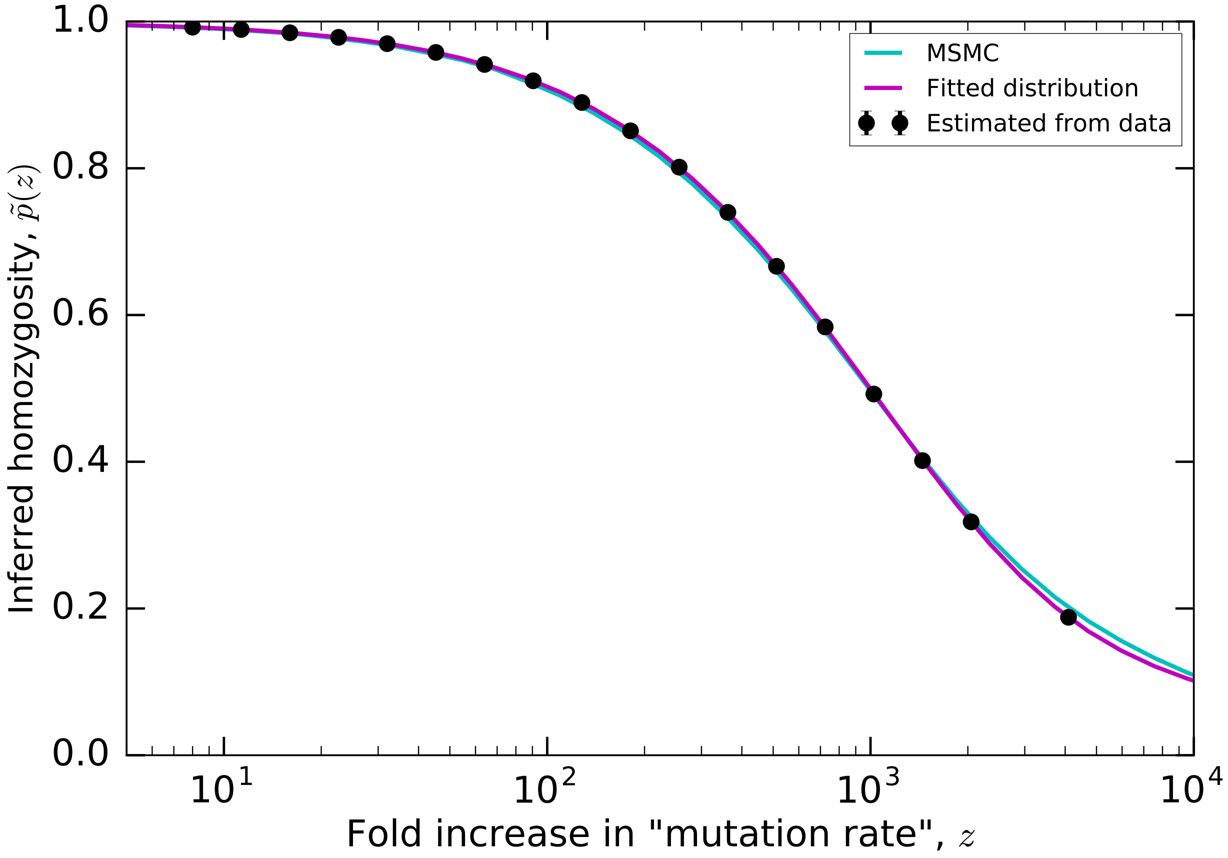

Figure 8: Laplace transform of the coalescence time distribution. Points are values estimated using the scale-dependence of the window-averaged diversity distributions, as shown in Fig. 7. Magenta curve shows MAGIC's fitted gamma mixture distribution, cyan curve shows the Laplace transform of MSMC's estimated coalescence time distribution. The two curves are close, but differ slightly for very large z, corresponding to very recent times.

Implementation

The code for MAGIC is written in Python and is available at https://github.com/weissmanlab/magic-beta. It uses the same input format as MSMC and the `msmc-tools` suite.

Data processing

We use the 69 Genomes Diversity Panel from Complete Genomics (Drmanac et al., 2010), and use msmc-tools (Schiffels & Durbin, 2014) to turn the data into a list of SNPs. We split the genome into windows of 80bp, count the number of SNPs in each window, and then repeatedly merge all windows in pairs and re-count to get the SNP count distribution at successively larger length scales (Figs. 1 and 6). (To correct for uneven sequencing coverage across windows, all windows with $<$ 80% coverage were dropped, and all with $>$ 80% coverage were down-sampled to 80%.) This gives us SNP count distributions at a range of length scales for every chromosome of every individual in the data set.

Simulated data

All coalescent simulations were done in `ms` (Hudson, 2002). To make the simulations computationally tractable, genomes were assembled from independently simulated "chromosomes" of 10^7 bases each.

For the test demographic scenarios in Figs. 2 and 3, the per-base mutation rate was $\mu = 10^{-3}/(4N_0)$. (`ms` is parametrized in terms of the present population size, $2N_0$.) For the "bottleneck" scenario, the demography was given by the command "`-eG .3 10 -eG .4 -10 -eG .6 0`"; for "repeated bottlenecks", by "`-eG .1 -10 -eG .3 10 -eG .5 -10 -eG .7 10 -eG .9 -10 -eG 1.1 10 `"; and for "admixture", by "`-es .1 1 .5 -ej 2 1 2`". See Fig. 9 for schematics. For the pairwise simulations, each sample consisted of 100 chromosomes with recombination rates as listed in Fig. 2. For the larger-sample simulations in Fig. 3, each sample consisted of 10 chromosomes with per-base rate of crossovers $\rho = \mu/5$, and per-base rate of initiation of gene conversion $g = \mu/20$ with mean tract length $\lambda = 1$kb.

Acknowledgements

We thank Stephan Schiffels for help with MSMC, Kelley Harris and Rasmus Nielsen for help with human population genetic data, Peter Ralph for discussions of the mathematical analysis, and Razib Khan for suggesting the name of the method. This work was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM115851 (O.H.) and by a Simons Investigator award from the Simons Foundation (O.H.).

References

Alfred J, Baldwin IT (2015) New opportunities at the wild frontier. eLife 4: e06956.

Bhaskar A, Song YS (2014) Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. The Annals of Statistics 42: 2469–2493.

Bunnefeld L, Frantz LAF, Lohse K (2015) Inferring bottlenecks from genome-wide samples of short sequence blocks. Genetics 201: 1157–1169.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78–81.

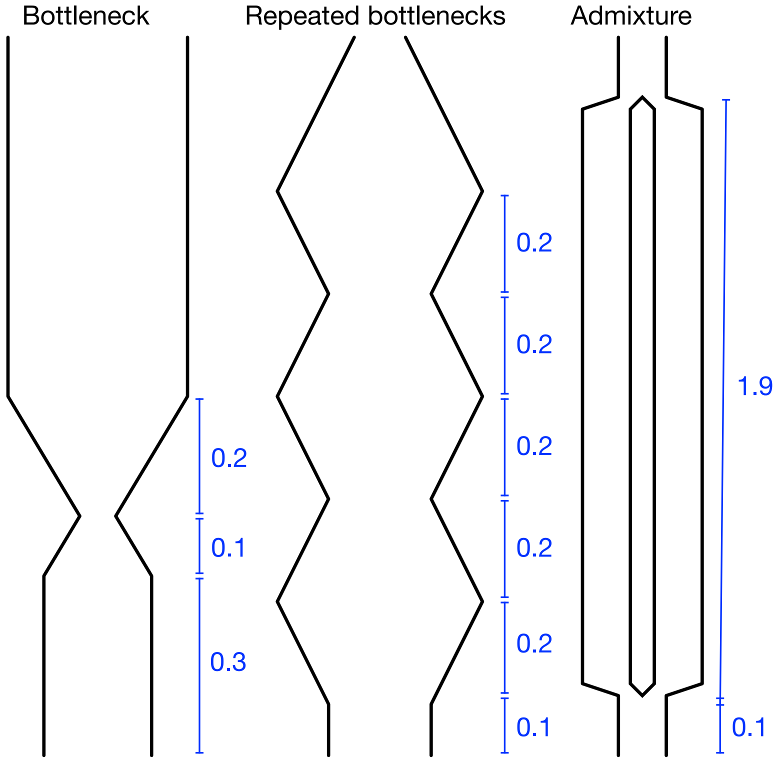

Figure 9: Demographic scenarios simulated. All time intervals are in units of $4N_0$. In the "bottleneck" and "repeated bottlenecks" scenarios, the population grows and shrinks exponentially at rate $10/(4N_0)$.

Epstein CL, Schotland J (2008) The bad truth about Laplace's transform. SIAM Review 50: 504–520.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genetics 9: e1003905.

Gattepaille L, Günther T, Jakobsson M (2016) Inferring past effective population size from distributions of coalescent times. Genetics: in print.

Ghosh M, Hwang JT, Tsui KW (1983) Construction of improved estimators in multiparameter estimation for discrete exponential families. The Annals of Statistics 11: 351–367.

Grimwood J, Gordon LA, Olsen A, Terry A, et al. (2004) The DNA sequence and biology of human chromosome 19. Nature 428: 529–535.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5: e1000695.

Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genetics 9: e1003521.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.

Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLOS Genetics 3: e7.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, et al. (2002) A high-resolution recombination map of the human genome. Nature Genetics 31: 241–247.

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

Marjoram P, Wall JD (2006) Fast "coalescent" simulation. BMC Genetics 7: 16.

McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1387–1393.

Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? Theoretical Population Biology 73: 342–348.

Paul JS, Steinrücken M, Song YS (2011) An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187: 1115–1128.

Ralph PL, Coop G (2013) The geography of recent genetic ancestry across Europe. PLoS Biology 11: e1001555.

Reddy CB, Hickerson MJ, Frantz LAF, Lohse K (2016) Approximate likelihood inference of complex population histories and recombination from multiple genomes. bioRxiv: 077958.

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

Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nature Reviews Genetics.

Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194: 647–662.

Simar L (1976) Maximum likelihood estimation of a compound Poisson process. The Annals of Statistics 4: 1200–1209.

Steinrücken M, Kamm JA, Song YS (2015) Inference of complex population histories using whole-genome sequences from multiple populations. bioRxiv: 026591.