The Lengths of Admixture Tracts

Mason Liang[1] and Rasmus Nielsen
University of California, Berkeley, California 94108

ABSTRACT The distribution of admixture tract lengths has received considerable attention, in part because it can be used to infer the timing of past gene flow events between populations. It is commonly assumed that these lengths can be modeled as independently and identically distributed (iid) exponential random variables. This assumption is fundamental for many popular methods that analyze admixture using hidden Markov models. We compare the expected distribution of admixture tract lengths under a number of population-genetic models to the distribution predicted by the Wright–Fisher model with recombination. We show that under the latter model, the assumption of iid exponential tract lengths does not hold for recent or for ancient admixture events and that relying on this assumption can lead to false positives when inferring the number of admixture events. To further investigate the tract-length distribution, we develop a dyadic interval-based stochastic process for generating admixture tracts. This representation is useful for analyzing admixture tract-length distributions for populations with recent admixture, a scenario in which existing models perform poorly.

THERE has been interest in analyzing population genomic data by using methods that partition an admixed individual's genome into blocks originating from different ancestral populations. An early version of the popular program Structure (Falush *et al.* 2003) accomplished this with a hidden Markov model (HMM), indexed along the genome, with hidden states corresponding to the ancestral population each position was inherited from. The contiguous blocks of the genome inherited from a population are called "admixture/migrant tracts/segments," depending on the context. For consistency, we use the term "admixture tract" in this article. Admixture tracts are unobservable, and their existence can be inferred only from genomic data. The process of doing so is called "admixture deconvolution" or "ancestry painting" and has been used in a number of different contexts, such as in admixture mapping for identifying human disease-associated genes (Hoggart *et al.* 2003; Reich *et al.* 2005), population-genetic inferences aimed at understanding human ancestry (Bryc *et al.* 2010; Henn *et al.* 2012), or identifying regions affected by natural selection (Tang *et al.* 2007).

The technique of using HMMs to partition an individual's genome into admixture tracts has been used in subsequent methods. Hoggart *et al.* (2003) and Smith *et al.* (2004) used HMMs for inferring admixture tracts with the purpose of admixture mapping and controlling for population stratification, similar to the method of Falush *et al.* (2003). More recent publications have focused on admixture deconvolution for more general population-genetic purposes, such as Tang *et al.* (2006) and Sundquist *et al.* (2008).

In HapMix (Price *et al.* 2009), the HMM model of Li and Stephens (2003) for modeling linkage disequilibrium is extended to include admixture between two populations. HapMix uses a genotype-based state space and so does not require phased data.

LAMP (Sankararaman *et al.* 2008; Paşaniuc *et al.* 2009; Baran *et al.* 2012) is similar to HapMix, in that it also can be considered an extension of the Li and Stephens model. However, the size of its state space does not depend on the number of reference haplotypes, which allows it to run faster than HapMix.

PCAdmix (Bryc *et al.* 2010; Brisbin *et al.* 2012; Henn *et al.* 2012) also uses an HMM to identify admixture tracts, but replaces observed data with admixture scores inferred from principle component analyses (PCA). As in the case of LAMP, it is applicable to multiple populations. Brisbin *et al.* (2012) argue that the method performs better than LAMP in simulations and has performance comparable to that of HapMix, which is limited to two populations.

There are also methods for estimating population-genetic parameters of admixture events from genomic data without

doi: 10.1534/genetics.114.162362
Manuscript received January 30, 2014; accepted for publication April 21, 2014; published Early Online April 26, 2014.
[1]Corresponding author: University of California—Berkeley, 736 Commercial St., #1, San Francisco, California 94108. E-mail: wmliang@berkeley.edu

first inferring admixture tracts, such as ROLLOFF (Moorjani *et al.* 2011). Other more general methods for estimating population-genetic parameters, such as $\partial a \partial i$ (Gutenkunst *et al.* 2009), can also be used to estimate time and the strength of admixture events. Finally, there are a many pregenomic methods for analyzing divergence and gene flow exemplified by the IM methods developed in Hey and Nielsen (2004) and Hey (2010). However, these methods do not directly use the information contained in the distribution of admixture tract lengths.

As a result of these efforts, there has been considerable interest in the relationship between admixture tract lengths and the time of admixture (T) and admixture fraction (m), to be defined mathematically later. Pool and Nielsen (2009) derived the admixture tract-length distribution under the assumptions that inbreeding is not significant and that tracts are so rare that they are unlikely to recombine with each other. Gravel (2012) relaxed this second assumption to model tracts descended from multiple migrant ancestors, but under simplified model of reproduction called the Markovian Wright–Fisher (MWF).

The methods for ancestry deconvolution discussed above use an HMM, assuming that the spacing between recombination events is independent and exponentially distributed and that ancestries of these recombination segments are independent. This is equivalent to assuming that admixture tracts have lengths that are independent and exponentially distributed. Population-genetic models that are designed to be Markov along the genome, such as the MWF, sequentially Markov coalescent (SMC) (McVean and Cardin 2005), or SMC′ (Marjoram and Wall 2006) models generate admixture tracts with these properties. Under the Wright–Fisher (WF) model with recombination, which is not Markov along the genome, we show that admixture tract lengths do not have an exponential distributionand, furthermore, that these lengths can be highly correlated. When T is small, these properties are a result of inheritance from a small, fixed sample pedigree, and when T is large, they are a result of inbreeding (in the sense of identity-by-descent due to genetic drift, as opposed to nonrandom mating). This former cause was first discussed by Wakeley *et al.* (2012) in examining the convergence of the ancestral recombination graph (Hudson 1983; Griffiths and Marjoram 1996) to the WF genealogical process. Because of this integration over pedigrees, the ancestral recombination graph diverges from the WF model when T is small and, like the Markov population-genetic models, generates independent, exponential tract lengths.

Parallel to the literature on inference methods for admixture deconvolution is a well-developed literature on the segregation of tracts in pedigrees. This starts with Fisher's theory of junctions (Fisher 1949). A junction is defined with respect to an ancestral population and is a point in the chromosome where, due to a crossover, the segments to the left and right trace their descent back to different members of the ancestral population. The distribution of the distances between junctions is of prime interest in this body of theory and is closely related to the distribution of admixture tract lengths. Fisher (1949) was interested in determining the expected number of junctions under different models of inbreeding. Stam (1980) extended Fisher's original results by considering a randomly breeding population of constant size and derived a number of different results under the assumption of independent and exponentially distributed tract lengths. Many studies have subsequently focused on the amount of genetic material passed from an individual to its descendants, given a known pedigree. Donnelly (1983) showed that the probability that an individual contributes no genes to a descendant T generations in the future is $\sim\exp(-TR/2^T)$, where R is the recombination map length. Barton and Bengtsson (1986) looked at the inheritance of blocks of loci under selection in hybridizing populations. Other studies have subsequently studied properties of the distribution of junctions and the distances between junctions, for fixed pedigrees including Guo (1994); Bickeböller and Thompson (1996a,b); Stefanov (2000); Cannings (2003); Dimitropoulou and Cannings (2003); Ball and Stefanov (2005); Walters and Cannings (2005); Rodolphe *et al.* (2008).

Baird *et al.* (2003) also consider the distribution of surviving tracts among the descendants of an individual. They model the number of descendants as a branching process and the lengths of inherited material carried by all descendants as a branching random walk. Assuming complete crossover interference (*i.e.*, at most one recombination event per chromosome), they derive the generating function for these lengths as a function of T and the map length. They also derive expressions for the mean number of tracts of a certain length under both the complete crossover interference model and a Poisson process of recombination. Baird *et al.* (2003) note that their results can be used to understand the process of genetic fragments between introgressed species, similar to the admixture problem considered here. In particular, they note that the standard deviations of both tract lengths and number of tracts are comparable to their means, indicating a high degree of variability. These results have been extended in other applications, for example, to derive the distribution of reproductive values (Barton and Etheridge 2011).

Chapman and Thompson (2002) derive general expressions for the mean and variance of the number of junctions. Their results can be applied under different demographic models because they show that these two moments depend only on the recombination map length and the one- and two-locus probabilities of identity-by-descent (IBD).

Beyond the fact that we focus on the effect on an admixed population, these approaches differ from our work in two ways. First, we consider the backward-in-time process of the ancestry of a sample, instead of considering the forward-time process describing the descendants of an individual. We also consider the merger of multiple fragments inherited from a group of individuals (migrants), instead of the contributions from just one. The effect of such mergers is particularly important when the number of migrants is large.

As no models other than the full WF model are available for accurate analyses of tract lengths for recent admixture

times, we present a new model of genealogical structure that can be used to analyze and approximate tract-lengths distributions and short-term pedigree-based processes more generally. This model assumes that the sample has a full pedigree and represents the genealogical history of a sample in terms of dyadic intervals. It is accurate for time scales and population sizes in which pedigree structure is important but inbreeding is not.

Methods

For simplicity, we consider a simple admixture scenario in which, T generations ago, two source populations contributed to form a third, admixed, population. Founders of this admixed population come from the "migrant" population with probability m and from the "nonmigrant" population with probability $1 - m$. Note that the labels on the two source populations are arbitrary.

Each of the population-genetic models analyzed in this article model the reproduction and recombination in this monecious population of $2N$ chromosomes subsequent to the admixture event. We assume that recombination events follow a Poisson process with rate 1 crossover/Morgan. This assumption of no crossover interference is not biologically accurate, but it is mathematically tractable. We later argue that this assumption is conservative with respect to the major conclusions of this article and show how our results can be extended to incorporate some models of interference.

Haploid Wright–Fisher with recombination

This is the standard haploid version of the WF model with recombination considered by Gravel (2012), Wakeley *et al.* (2012), and others. Each chromosome is produced by recombining two parents from the previous generation, chosen independently and uniformly at random. We consider this to be the more appropriate model for understanding tract-lengths distributions and compare the following models to it.

Markovian Wright–Fisher

Gravel (2012) introduced this mathematically tractable approximation of the diploid WF model. It assumes that chromosomes are formed from the recombination of *all* $2N$ chromosomes from the previous generation, instead of just two. At each recombination point, the offspring copies from one of the $2N$ chromosomes from a previous generation, uniformly at random. Additionally, it assumes that $2N$ is large, so that each crossing-over results in a new parent contributing genetic material. As its name implies, the MWF model is a Markov process along the genome.

Coalescent with recombination

In the coalescent limit ($2N \to \infty$ with time measured in units $2N$ generations and recombination distance in units of crossovers/$4N$), Griffiths and Marjoram (1996) showed that the genealogical process of a sample from the haploid WF model converges in distribution to the ancestral recombination graph (ARG), which can be constructed as a Markov process going backward in time. Wiuf and Hein (1999) presented a sequential construction of the ARG along the genome. This sequential process is not Markov. Instead, the conditional distribution of a marginal trees depends on all the trees that have appeared to the left of it. The case of admixture tracts is slightly different than other uses of the coalescent, because here we start with one lineage and stop the process at the fixed time, $T/2N$, instead of the more common case, where we start with more than one lineage and stop the process when only one lineage is left.

Sequentially Markov coalescent

McVean and Cardin (2005) developed an approximation of the coalescent in which the sequence of marginal trees form a Markov process along the sequence. In the SMC, the only allowed coalescence events are for lineages with overlapping ancestral material. The model is otherwise identical to the coalescent.

Majoram and Wall's SMC′

Marjoram and Wall (2006) presented a related model (SMC′) that loosens the restrictions of the SMC while retaining its Markov property. In addition to the coalescence events allowed in the SMC, the SMC′ further allows coalescence events for lineages with abutting ancestral material. This extra possibility allows for back coalescences in the ancestral recombination graph, which produces a significant improvement for this model's predictive powers when these events are likely.

Perfect binary tree model

As we argue in *Results*, none of the four previous models approximate the tract-length distribution well when T is small relative to $2N$. We therefore introduce the perfect binary tree model (PBT), so named because it assumes that the sample has 2^T distinct great^{T-2} grandparents, *i.e.*, that the pedigree of the sample, up to generation T, is a perfect binary tree with depth T. From simulations, we found that this approximation produces accurate results when $2^T < N$, which is the parameter space for which the coalescent approximation does not. For most biological populations, this restricts T to a rather limited set of parameter values, but often, this is a region of great interest. Some definitions and properties of this process are discussed in the following section, which can be skipped by the less mathematically interested reader.

Our goal is to characterize the stochastic process by which segments of ancestral genetic material are recombined to form the genome of a particular person of interest (the proband). We call this the ancestor-copying process, which represents the line of descent of the proband's genome as a function of the genomic position. Label the parents of an individual as the left and right parent, respectively. The ancestry of an individual in a particular position in the genome is then determined by the choices of left and right parents back in time on the pedigree.

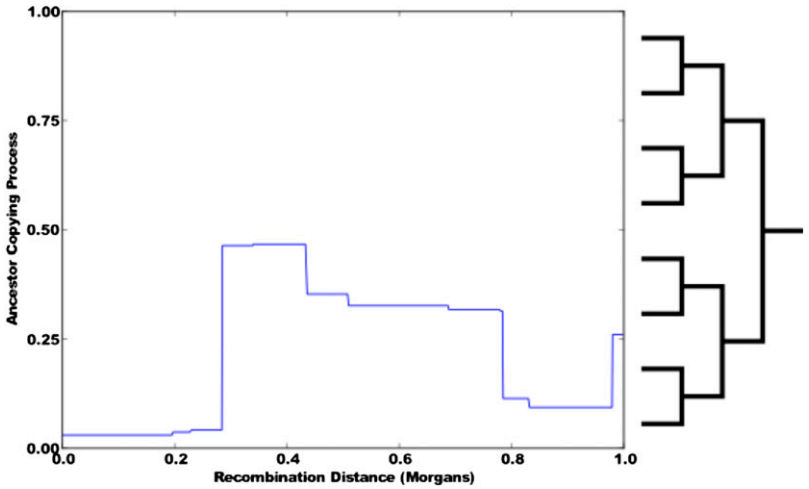

Figure 1 A realization of the ancestor-copying process. In this case, the process stays in the interval $\left[0,\frac{1}{2}\right)$, indicating that this length of chromosome was inherited entirely from the proband's left parent. The process jumps between $\left[0,\frac{1}{4}\right)$ and $\left[\frac{1}{4},\frac{1}{2}\right)$ three times, indicating that each left grandparent contributed two blocks to the proband. The pedigree, up to the proband's eight great-grandparents, is shown on the right. Each ancestor has been placed in its corresponding dyadic interval.

In investigating IBD probabilities, Donnelly (1983) considered this ancestry as a random walk on a hypercube, with each vertex corresponding to the set choices of left or right parents for *every* individual in the pedigree. For a perfect binary tree, the size of this state space is superexponential in T, which Donnelly (1983) was able to considerably reduce by using symmetries in the transition matrix. For the ancestry-copying process, we cannot use these symmetries in the same way and instead directly integrate over hidden recombination events.

We instead represent this ancestry using dyadic intervals. At a position in the genome, x, the ancestor-copying process N_x takes a value from the half-open interval [0, 1). The dyadic intervals in which N_x is contained correspond to the ancestors this position was inherited from. We define dyadic intervals to be half-open intervals of the real line of the form $I_{j,k} = [k2^{-j}, (k+1)2^{-j})$ for $j, k \in \prime$, $k < 2^j$. Dyadic intervals are isomorphic to the nodes of binary trees in that every dyadic interval is the union of two unique disjoint dyadic intervals. We use the following notation to denote the left and right halves of a dyadic interval $I_{j,k}$:

$$I^{\ell}_{j,k} = \left[k2^{-j}, (2k+1)2^{-j-1}\right)$$
$$I^{r}_{j,k} = \left[(2k+1)2^{-j-1}, (k+1)2^{-j}\right).$$

We denote the length of a dyadic interval by $|I_{j,k}| = 2^{-j}$ and define the distance between two dyadic intervals, $d(I, J)$, to be the length of the shortest dyadic interval containing both. For a dyadic interval I, we define I' to be the dyadic interval with $2|I| = |I'|$ such that $I \subset I'$ and I^* to be the set difference of I' and I.

We associate an ancestor to each dyadic interval in [0, 1): the proband to $I_{0,0}$, the left parent to $I_{1,0}$, the right parent to $I_{1,1}$, the left parent's left parent to $I_{2,0}$, etc. The value of the ancestor-copying process at a particular position represents the ancestors the proband inherited that position from; *e.g.*, if the ancestor copying process is $<\frac{1}{2}$, then the proband inherited that position from the left parent, or if is $\geq\frac{3}{4}$, then the proband inherited that position from the rightmost grandparent (and consequently the right parent). A realization of the ancestor-copying process is given in Figure 1.

The defining property of the ancestor-copying process is that its distribution does not change after a generation of recombination. The process of recombination between two parental genomes can be described by a two-state Markov process, R_x, which switches between 0 and 1 at rate 1. If N_x and $N_x^{'}$ are the independent ancestor-copying processes of the two parents, which are jointly independent of R_x, then

$$N_x \overset{d}{=} \tfrac{1}{2} R_x N_x + \tfrac{1}{2}(1 - R_x)\left(1 + N_x^{'}\right). \tag{1}$$

This property makes it clear that conditional on R_x, the behavior of N_x in the range $\left[0,\frac{1}{2}\right)$ is independent of its behavior in $\left[\frac{1}{2}, 1\right)$. In fact, this property can be extended to any mutually disjoint collection of dyadic intervals:

Theorem 1. *For a dyadic interval A, the processes* $N_x\mathbf{1}\{N_x \in A\}$ *and* $N_x\mathbf{1}\{N_x \notin A\}$ *are conditionally independent given* $\mathbf{1}\{N_x \in A\}$.

An intuitive explanation for this theorem is that because there is no inbreeding, ancestors that are not lineal descendants will be unrelated, and hence independent. The mathematical proof, as with all others in the article, is presented in the *Appendix*.

To characterize the ancestor-copying process, we want to find the rate at which N_x leaves a dyadic interval I,

$$n_I = \lim_{x \downarrow 0} \frac{1 - \mathbb{P}_I(N_x \in I | \mathcal{N}_0)}{x},$$

and the transition rates between disjoint dyadic intervals I and J,

$$n_{I,J} = \lim_{x \downarrow 0} \frac{\mathbb{P}_I(N_x \in J | \mathcal{N}_0)}{x},$$

where $\mathbb{P}_I$ is the measure induced by conditioning on $N_0 \in I$ and $\mathcal{N}_0 = (\{N_x : x \leq 0\}$.

Theorem 2. *The length over which* N_x *remains in a dyadic interval is exponentially distributed, with rate given by*

$$n_{I_{j,k}} = j.$$

Theorem 3. *The transition rates between disjoint dyadic intervals is given by*

$$n_{I,J} = \prod_{i \in P(I,J)} \frac{1}{2} + \left(\mathbf{1}\{T_i > T_{i^*}\} - \frac{1}{2} \right) \exp(-2T_i)$$

with

$$T_I = \sup\{x < 0 : N_x \in I\}$$

and

$$P(I, J) = \{i \in \mathcal{I} : |i| < d(I, J), J \subset i\}.$$

The rate at which N_x leaves dyadic intervals depends only on the length of the dyadic interval, which is in accord with the results of Baird *et al.* (2003), Pool and Nielsen (2009), and Gravel (2012) regarding the exponential distribution of genetic distance between recombination events. However, the process is not Markov, because the transition rates depend on the values of N_x for $x \leq 0$ and not just N_0.

The MWF and SMC models assume that segments are inherited from distinct ancestors, but for the PBT model, multiple segments can be inherited from the same ancestor. The probability of this event decreases as T increases, confirming the prediction given in Baird *et al.* (2003).

Simulations

As we explain in the results, when there is a single pulse of admixture, the Markov models (MWF, SMC, and SMC′) produce admixture tracts whose lengths are independent and exponentially distributed. For the other models, we first wrote Monte Carlo simulations that assigned an ancestor to each recombination segment. For the coalescent model, we used code that was essentially identical to the program ms (Hudson 2002), with two modifications: the backward process stops at the time of admixture, instead of when only one lineage remains, and the simulation starts with just one lineage. The extant lineages at the time of admixture are then traced forward in time to find which recombination segments they contribute.

For the PBT model, we used the transition rates from Theorem 3 to efficiently simulate N_x on the dyadic intervals with size at least 2^{-T} in the following manner: The stationary distribution of N_x is uniform on [0, 1), so we put N_0 in a dyadic interval, I, with length 2^T, chosen uniformly at random. The length for which N_x remains in this interval has an Exp(T) distribution. Note that $n_{I,I^*} = n_{I,(I')^*} = n_{I,(I'')^*} = \cdots = 1$, and that $I, I^*, (I')^*, \ldots$ form a partition of $I_{0,0}$ so we first determine which of these dyadic intervals N_x jumps to. Conditional on this, we then recursively determine which of the left and right dyadic intervals contain N_x, until we have narrowed N_x down to a dyadic interval of length 2^{-T}. As we do this, we also update the values of the T_I's. One of the advantages of the dyadic interval representation is that it allows efficient simulations of pedigree structure by simulating a stochastic process on [0, 1) instead of representing full pedigrees for each segment of the genome as a linked list in the computer memory.

The WF model is the same as the PBT model, with the exception that inbreeding is allowed. We still represent the pedigree as a perfect binary tree, with the caveat that some of the nodes are taken to represent the same ancestor. For the simulation, this means that some of the T_i's for different dyadic intervals that represent the same ancestor will in fact be equal. Generating the entire pedigree is computationally expensive for large T, so we extend the pedigree only as is needed, *i.e.*, as N_x jumps to previously unvisited dyadic intervals.

After assigning an ancestor to each recombination segment, we then independently label each ancestor as migrant or nonmigrant, with probabilities m and $1 - m$, respectively, allowing us to demarcate admixture tracts. For each set of admixture parameters, we used a simulated a segment of genome 30 times longer than the average tract length. To minimize edge effects, we examine only the tracts from the middle third of this segment.

Models of multiple admixture pulses

The Markov models (MWF, SMC, and SMC′) predict that admixture tracts resulting from one pulse of admixture will have exponentially distributed lengths, while those resulting from two (or more) pulses of admixture will have length distributions that are the mixture of two (or more) exponentials. On the other hand, the Wright–Fisher model produces admixture tracts that are nonexponential, even in the one-pulse scenario. As a result, when analyzing the data using a Markov model, it is possible to mistakenly conclude that the observed tract-length distribution cannot be explained by just one pulse of admixture, when in fact it can be, but only by using the more complex Wright–Fisher model.

We investigated the probability of this happening when using a likelihood ratio test to distinguish between an exponential distribution *vs.* a mixture of two exponentials. To draw from the null distribution, we simulated 10^4 admixture tracts with exponentially distributed lengths and found the maximum log-likelihood of these under a mixture model, with two exponentials, *i.e.*,

$$\mathcal{L}(p, a, b|x) = \prod_{i=1}^{10^4} [pae^{-ax_i} + (1-p)be^{-bx_i}],$$

where each x_i is the length of a admixture tract. This maximization was done by a standard expectation maximization (EM) algorithm. The 100 initial random values p_0, a_0, and b_0 were repeatedly updated by first computing the posterior probabilities,

$$r_{i,t} = \frac{p_t a_t e^{-a_t x_i}}{p_t a_t e^{-a_t x_i} + (1-p_t) b_t e^{-b_t x_i}},$$

and then the likelihood-maximizing posterior means

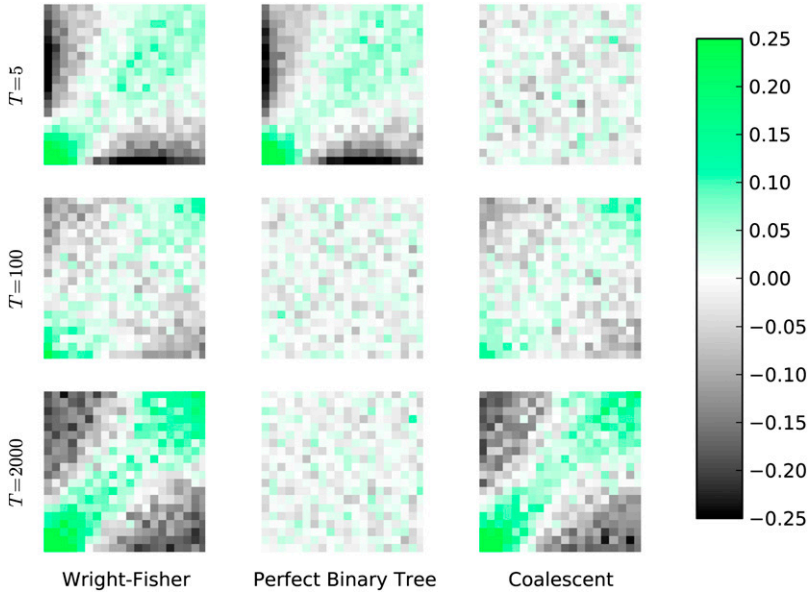

Figure 2 The correlation of the lengths of consecutive admixture tracts for the WF with $2N = 1000$ (red), PBT (green), and coalescent (blue) models. In all cases the admixture fraction is $m = 0.95$. Admixture tract lengths were transformed into the unit interval by their empirical quantiles, so uncorrelated lengths would produce an entirely white square. The simulations were run with a population size of $2N = 2000$.

$$\hat{p}_{t+1} = \frac{\sum_{i=1}^{10^4} r_{i,t}}{10^4}$$

$$\hat{a}_{t+1} = \frac{\sum_{i=1}^{10^4} r_{i,t}}{\sum_{i=1}^{10^4} r_{i,t} x_i}$$

$$\hat{b}_{t+1} = \frac{\sum_{i=1}^{10^4} (1 - r_{i,t})}{\sum_{i=1}^{10^4} (1 - r_{i,t}) x_i}.$$

The values were updated until the log-likelihood improvement was $<10^{-3}$. We took the highest log-likelihood value resulting from these 100 optimizations to be the maximum log-likelihood under the mixture model for this sample.

Tests of a single admixture pulse

To test the null hypothesis of a single admixture event, we define a likelihood ratio test statistic, S, by subtracting the maximum log-likelihood value under the full model with two admixture events from that obtained for a model allowing only a single admixture event. The asymptotic distribution for this test statistic is not known, because some parameters of the alternative hypothesis are not estimable under the null hypothesis. This implies that the general asymptotic likelihood theory is not applicable. To obtain critical values for this test statistic we instead used parametric simulations under the null hypothesis and assuming independent exponentially distributed tract lengths. We simulated 10^5 samples to approximate the critical values corresponding to significance levels of $P = 0.05$ and $P = 0.02$ a range of values for T and for $m = 0.1$, 0.3, and 0.5. We then compared this distribution of log-likelihood ratios to log-likelihood ratios obtained in the same way for simulated datasets of 10^4 tracts generated under the Wright–Fisher model with a single admixture event.

Results

The models described in the *Methods* predict that the sampled chromosome can be viewed as a mosaic of recombination segments from chromosomes in generation T. The models agree in predicting that the distance between recombination events, and hence the length of a recombination segment, is exponentially distributed, with scale T^{-1}, but differ in their predictions regarding how recombination segments are inherited from ancestors from the admixing generation. In the following, we use simulations to illuminate these differences.

Admixture tracts lengths are neither independent and identically distributed nor exponentially distributed

Recombination fragments are exponentially distributed in the WF model. Under the assumption that all ancestors are distinct, Theorem 2 shows that the distribution of the length of fragments in which an individual has any particular ancestor T generations ago is also exponentially distributed, with scale T^{-1}. If admixture tracts are assumed to be so rare that they are unlikely to recombine with each other, then admixture tract lengths will also be exponentially distributed, and the process will be well modeled using the independence assumption of Pool and Nielsen (2009). However, admixture tracts are different from recombination segments, as multiple recombination segments can recombine to form a single admixture tract. This was the situation considered by Gravel (2012). In general, if the lengths of recombination tracts are independent and identically distributed (iid) exponential random variables, and each segment is migrant independently and with probability m, then the length distribution of admixture tracts would be found as a geometric mixture of exponential random variables and consequently exponentially distributed with scale $[T(1 - m)]^{-1}$. However, the second condition is not true. There are two reasons for this. First, as shown by Theorem 3 the ancestry-copying process is not

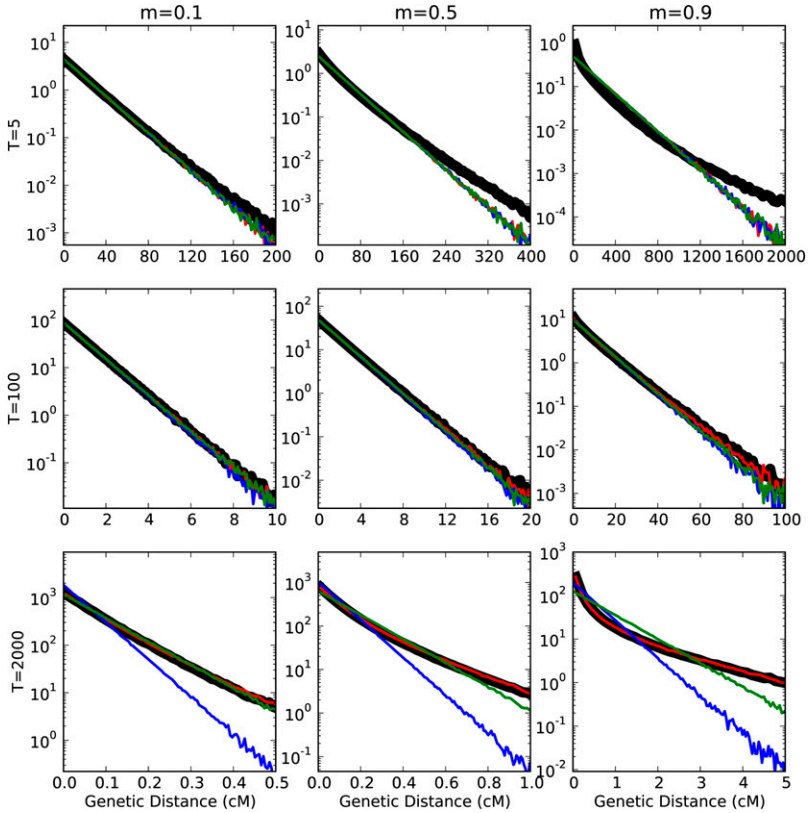

Figure 3 Admixture tract-length distributions for the MWF and SMC (both blue), SMC′ (green), coalescent (red) models compared to the distribution under the WF model (thick black). Note that the *y*-axis is shown on a logarithmic scale. The simulations were run with a population size of $2N = 2 \times 10^3$. For $T = 5$, the former three models give exponential distributions and do not match the WF distribution. For $T = 2000$ the coalescent and WF distributions are the same.

Markov. An individual has a finite number of ancestors and recombination can bring together recombination fragments inherited from the same ancestor. As a result, the lengths of migrant tracts are correlated when T is small. Another factor that contributes to this correlation is the variance in the number of migrant ancestors an individual has. For instance, an individual with one migrant grandparent will have admixture tracts that tend to be shorter than those for an individual with three migrant grandparents. The effect of this is illustrated in Figure 1 for $T = 5$. In addition, when T is large, the number of genetic ancestors will be significantly smaller than 2^T. It might be useful to think of this effect forward in time as an effect of inbreeding, in which admixture tracts introduced into the population are broken up by recombination but also joined again by inbreeding. As a result, many fragments in the population segregating after time T will likely be descendants of a relatively few number of larger fragments. The location of smaller fragments will therefore be correlated in the genome, corresponding to the location of the initial admixture fragments, and back recombination has a higher probability than under the iid assumption. This effect is illustrated in Figure 1 for $T = 2000$.

Baird *et al.* (2003) also simulated and commented on the clustering of tracts in the genome. A single tract spanning a larger region may survive the first generations and then be broken up into smaller fragments in different individuals in the same region of the genome. Martin and Hospital (2011) also examined the problem of correlated tract lengths, but in the context of recombinant inbred lines, and similarly concluded that tract lengths are not independent.

As a consequence of the correlation in tract lengths along the chromosome, admixture tracts are not accurately modeled as a geometric mixture of iid recombination fragments. This effect is illustrated in Figure 2. The strongest deviations occur when T is large or when the admixture proportion is large. The length distribution of admixture fragments when the admixture proportion is m corresponds to the distribution of distances between fragments when the admixture proportion equals $1 - m$. In terms of HMM modeling, deviations from exponential distribution of either admixture fragments, or distances between admixture fragments, will violate the model assumptions.

Related results have previously been obtained relating to the theory of junctions. Chapman and Thompson (2002) examined an assumption of independent Poisson-distributed junctions among individuals and independence of junctions within individuals. They noted that this assumption tends to underestimate the true variance when $T/N > 1$. Although the assumptions in their study are different from ours—in particular we consider descent from multiple migrant individuals and the possibility of recombination between tracts from these individuals—the conclusion reached by Chapman and Thompson (2002) is essentially similar to the one reached here: tracts are not exponentially distributed when T is large relative to N. Martin and Hospital (2011) examined this problem further in the context of recombinant inbred lines and similarly concluded that tract lengths are not exponential.

The interplay of the nonindependence and nonexponentiality of the admixture tract distribution can be illustrated

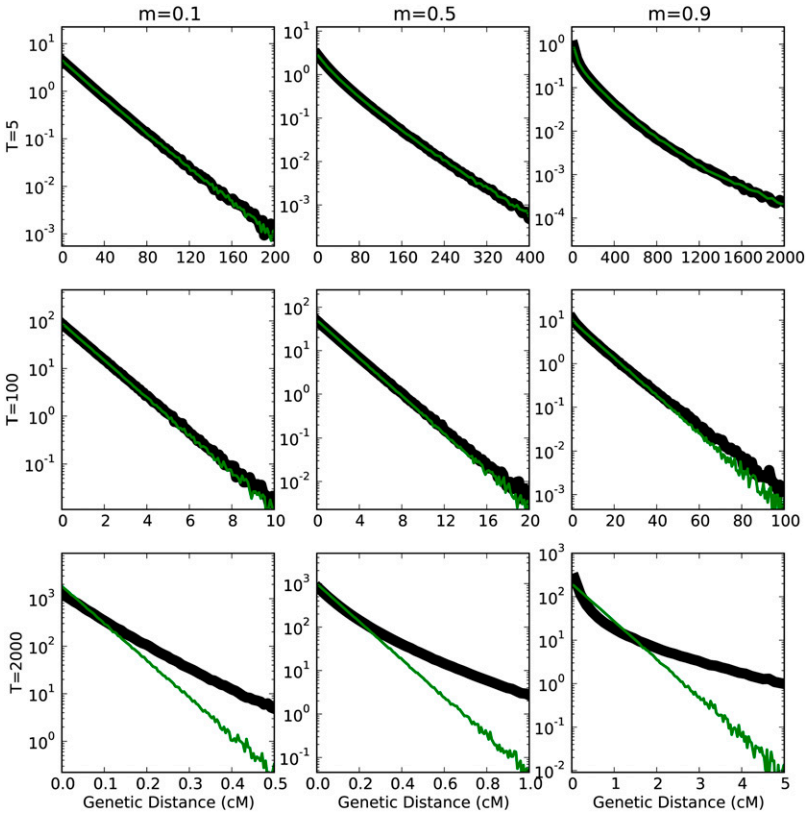

Figure 4 Admixture tract-length distributions for the PBT model (green) and the WF model (thick black). The simulations were run with a population size of $2N = 2 \times 10^3$. Note that the *y*-axis is shown on a logarithmic scale. For $T = 5$, the PBT model matches the WF model closely, while for $T = 2000$, it does not and has an exponential distribution instead.

by looking at the distribution of admixture proportions, the proportion of a window that is inherited from migrant ancestors. This is presented in Figure 4, using a window size of 1 cM, in an admixture scenario in which the pattern of admixture tracts is expected to have fixed in the population. The PBT, MWF, and SMC models do not account for the effect of inbreeding, so they predict that admixture tracts will become ever smaller as T becomes larger. As a result, they predict degenerate admixture proportions, *i.e.*, an atom on m. Consequently, these models were not included in Figure 4. The coalescent, SMC′, and WF models do take inbreeding into account, and consequently predict nondegenerate limiting distributions for the admixture proportion.

For both values of m, the distribution predicted by the WF and coalescent models has a larger variance than that predicted by SMC′, while having the same mean. For small values of m, this is because admixture tracts are likely to be clustered and have either zero or a larger number of tracts than predicted by SMC′. For large values of m, this higher variance is better explained by the fat tails of the admixture tract-length distribution.

Coalescent with recombination

The coalescent provides an approximation to the WF model that is in general excellent, but may be less so when considering the dynamics shortly before the time of sampling (Wakeley *et al.* 2012). In the present context this means that the coalescent approximates the WF model well when T is large, but not necessarily so for small values of T. The correlation that arises due to inbreeding is well modeled by the ARG, but the correlation due to a small number of ancestors in the pedigree in the very recent ancestry is not. This is shown in Figure 1. For small values of T, the coalescent does not accurately capture the correlation structure. As a consequence, the distribution of admixture tract lengths is not well modeled when T is small (Figure 2), particularly for large migration fractions ($m = 0.9$). In an admixed population, the distribution of tracts originating from the population contributing most of the genetic material are far from exponentially distributed. However, the effect rapidly diminishes as T increases.

Markovian models

The MWF, SMC, and SMC′ models all generate admixture tracts with exponentially distributed lengths. In these models, admixture tracts follow a geometric mixture of iid exponential random variables. In each of these Markovian models, the ancestry of a recombination segment depends only on the ancestry of the recombination segment to its left. As a result, the number of recombination segments that make up an admixture tract will be a geometric random variable. The geometric mixture of iid exponential random variables results in another exponential. Under the MWF model, each recombination segment is inherited from a distinct ancestor in generation T. Each of these ancestors is from the admixing population with probability m, so admixture tract lengths are exponentially distributed with scale $[T(1 - m)]^{-1}$, as previously discussed. In the SMC, the recombined lineage cannot coalesce back to the

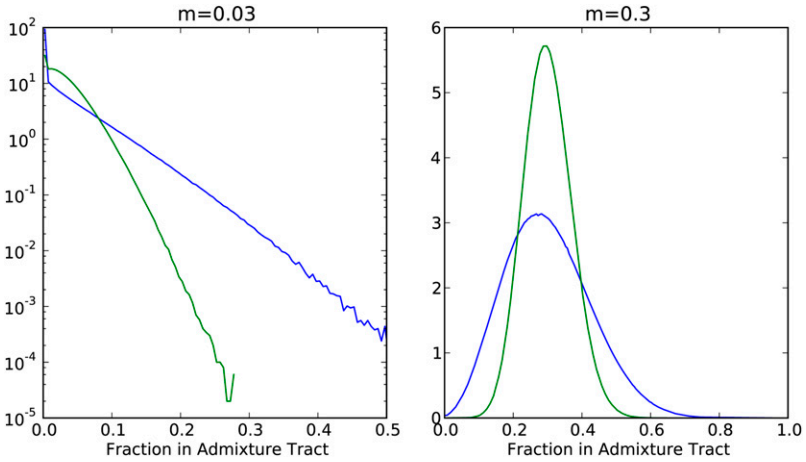

Figure 5 Distributions of the fraction of 1-cM windows that are parts of admixture tracts, for two values of m. Parameters for the two simulations were otherwise the same, with $N = 5 \times 10^3$ and $T = 2 \times 10^4$. The distribution under the SMC′ model is in green and the distribution under the coalescent and Wright–Fisher models is in blue. Note that the left graph is plotted on a log scale.

current marginal tree, so as in the Markovian WF model, each recombination segment will be descended from a distinct ancestor and admixture tracts lengths will again be exponentially distributed with scale $[T(1-m)]^{-1}$. In SMC′, back coalescences to the current marginal tree are possible and occur with probability $1-2N\left(1-e^{-T/2N}\right)/T$. In this event, the recombination segment is migrant if and only if the previous segment was. Therefore, the probability that the segment on the right of a recombination point is migrant, given that the segment on the left was, is

$$\left[1-\frac{2N}{T}\left(1-e^{-T/2N}\right)\right] + \left[\frac{2N}{T}\left(1-e^{-T/2N}\right)\right]m = 1-\frac{2N}{T}(1-m)\left(1-e^{-T/2N}\right),$$

so admixture tract lengths will have an $\mathrm{Exp}\left[2N(1-m)\left(1-e^{-T/2N}\right)\right]$ distribution. When $2N \gg T$, this is approximately the same distribution given by the other two models, but for fixed $2N$ and as $T \to \infty$, SMC′ makes the more accurate prediction that the average tract length goes to the nonzero value of $[2N(1-m)]^{-1}$.

These models may fail to give accurate predictions for both small and large values of T. These are two separate effects. When T is small they give inaccurate predictions for the same reasons as the coalescent. In particular, they do not accurately model the correlation due to a fixed number of ancestors in the pedigree and the possibility of back recombination. For this reason, tract-length distributions do not fit well, especially for large values of m.

For large values of T they fail because they do not accurately model the effect of inbreeding. The MWF model and the SMC give identical predictions (Figure 3). When T is large, they underestimate the length of admixture tracts for small values of m. For large values of m they underestimate the variance in tract length. In either case, the fit of tract-length distribution to that expected under the WF model, or the coalescent, is poor. In the coalescent and WF models, nonadjacent segments may be descendants of the same ancestor, an event that occurs with higher probability as T increases. The overall effect of this is that the Markovian models are too likely to assign more distinct ancestors to a given length of chromosome, which increases the probability that some section was inherited from a nonmigrant ancestor. The error for the SMC′ is less than that of the SMC and Markovian Wright–Fisher model (Figure 3).

Perfect binary tree

In the *Methods*, we derived a genealogical model that can be used to study tract-length distributions when T is small. This process captures the correlation structure and admixture tract-length distribution of the full WF model for small T (Figure 2 and Figure 4), something that the other approximative models explored here fail to do. However, the model does not accurately describe the dynamics when T is large, as it assumes that all ancestors from generation T are distinct. For $T > \log_2 N$, this is not possible, and some ancestors must necessarily be the same.

This is consistent with the result of Baird *et al.* (2003), which found that, asymptotically for large T, the probability that an individual inherits multiple blocks from one ancestor goes to zero. In this limit, where every recombination segment is inherited from a distinct ancestor, admixture tracts lengths will be idd exponential, as in the case of the Markov models.

Admixture tracts as distances between junctions

We further compare our results with the results of Baird *et al.* (2003) to illustrate the effect of considering multiple ancestors of an individual and the effect of assumptions regarding crossover interference. Baird *et al.* (2003) consider the distribution of the lengths of genetic material inherited from one individual, in a branching-process model with complete interference, *i.e.*, assuming at most one recombination event on a chromosome each generation. They found that the density, in z, for this distribution is given by

$$\frac{(1-z)^{T-1}\left(2T+T(T-1)\left[y-z/1-z\right]\right)}{1+yT},$$

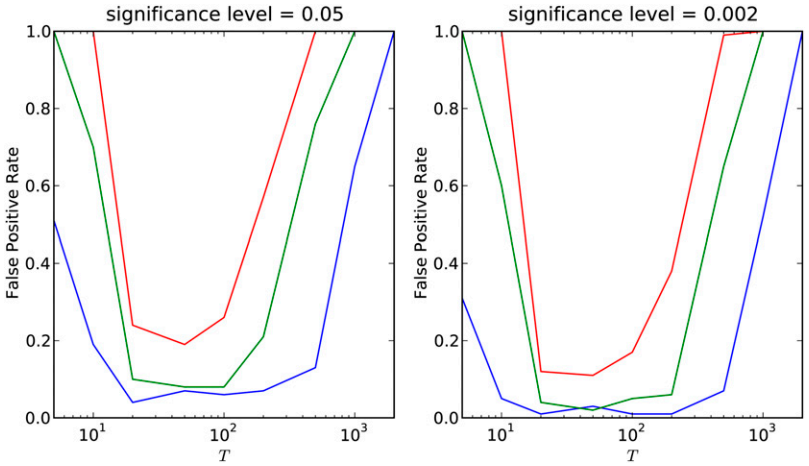

Figure 6 Probability of erroneously inferring two pulses of admixture as a function of T, when using a MWF or SMC′ null model. The red, green, and blue lines correspond to $m = 0.5$, 0.3, and 0.1, respectively. The left plot is for a likelihood-ratio test with $\alpha = 0.05$ and the right plot is with $\alpha = 0.002$.

where y is the recombination probability and T is the number of generations. When m is small, *e.g.*, 0.01, most admixture tracts will be inherited from just one migrant ancestor. In this scenario, the Baird distribution is comparable to the admixture tract-length distribution (Figure 7). When $T = 5$, the Baird distribution differs from the WF and PBT models because it uses a different model of interference. Under its assumption of complete interference, no tract can span more than a map distance of y, whereas the other two models have no such maximum. In the bottom row, where $T = 2000$, both the Baird distribution and the PBT model fail to account for the back coalescence of different fragments and consequently predict tracts that are shorter than under the WF model. However, there are no effects with regard to their different assumptions about recombination interference. For $T = 100$, when the effects of back coalescence are negligible, all three models predict the same distribution, despite their different assumptions.

When m is not small, the Baird distribution fits less well (shown in the right column of Figure 7). This is mainly because each admixture tract is now more likely to be composed of genetic material inherited from multiple migrant ancestors.

Likelihood ratio test of the number of admixture pulses

To determine the effect of wrongly assuming iid exponential tract lengths for inferences for real data, we implemented a likelihood ratio test and tested the null hypothesis of one admixture pulse, against the alternative of two admixture pulses, on data simulated under the null hypothesis. The false-positive rate, defined as a fraction of these log-likelihood ratios that exceeded the critical value (obtained using simulations), was plotted as a function of T and is shown in Figure 6. Note that there is a strong excess of false positives, particularly when T is large or small. The false-positive rate is less for intermediate values. This is explained by the observations from the previous sections, showing that the assumption of iid exponential tract lengths is particularly poor when T is very small (due to finite number of ancestors in the pedigree) or larger than N (due to inbreeding).

Discussion

We have found that under many scenarios, the Wright–Fisher model produces admixture tracts whose lengths are not well approximated as independent, exponential random variables. There are two major effects that are important to distinguish: the effect of a finite number of ancestors in the pedigree for small values of T and the effect of inbreeding for large values of T. Both of these effects cause deviations from the idd exponential assumption.

When using an HMM for ancestry deconvolution, the Markov model provides a prior on tract lengths. If there if signal regarding local ancestry in the data, then misspecification of this prior may not matter a great deal. However, for parametric population-genetic data analysis, *i.e.*, estimating the number of timing of admixture events, it may be desirable to consider possible biases incurred due to assumptions regarding exponential tract lengths. One way to verify inferences of multiple admixture pulses would be to compare the simulated tract-length distribution under the WF model to the data.

The magnitude and direction of the estimation bias will depend on the model and the values of m and T. For small values of T, Figure 3 shows that the Markov models underestimate the number of long tracts. Consequently, estimates of T based on the number these longer tracts will be downwardly biased.

The biases can be avoided by using the Wright–Fisher, instead of a Markov, model to construct a prior for the local ancestry distribution. However, there are no known computationally efficient algorithms for integrating over this prior. However, efficient inference under the perfect binary tree model may be possible, because of the conditional independence given by Equation 1. When T is small, this would be a good approximation to inference under the Wright–Fisher model. As the simulations show, when $20 < T \ll 2N$, all of the models produce approximately the same tract-length distributions, so in this region of the parameter space, there is minimal bias from using a Markov model.

The deviations from a Markov model explored here may also affect methods that do not directly attempt to estimate

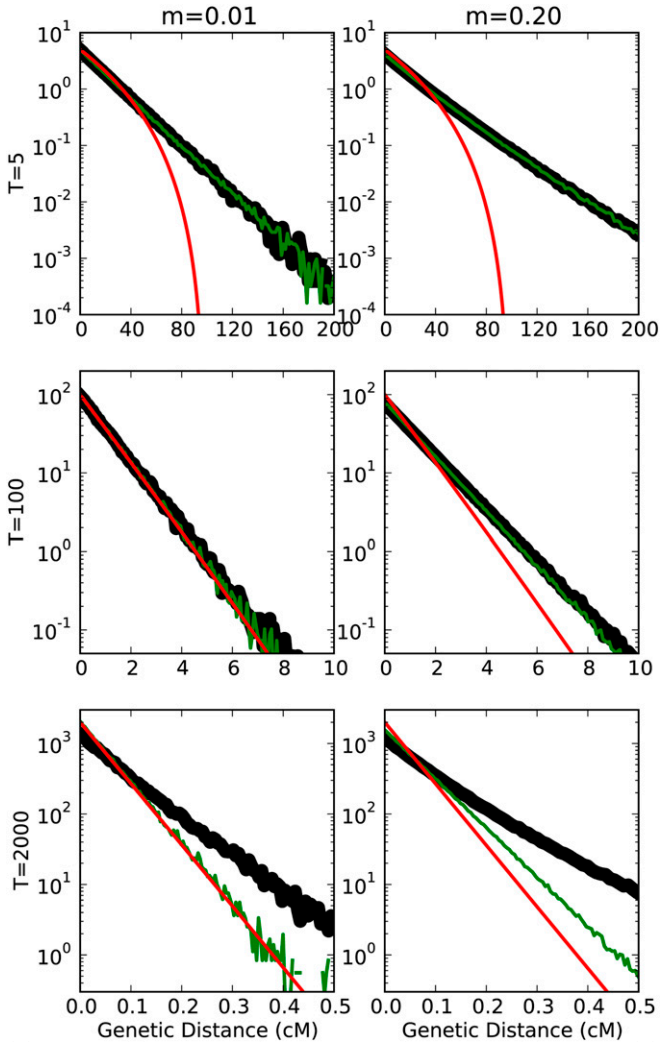

Figure 7 Tract-length distributions for the Baird distribution (red), PBT model (green), and the WF model (thick black). The WF simulations were run with a population size of $2N = 2 \times 10^3$. Note that the *y*-axis is shown on a logarithmic scale. When m is small and at intermediate time scales, all three models agree.

admixture tract distributions. For example, ROLLOFF (Moorjani *et al.* 2011) assumes that the probability that two sites a distance r apart are linked after T generations is given by $\exp(-rT)$ and uses this to make a prediction about the value of a correlation coefficient. Under the PBT model, this probability is $((1 + \exp(-2r))/2)^T$, and under the WF model, this probablity is $(1 - 1/N)^T((1 + \exp(-2r))/2)^T$. For some values of N, r, and T, these probabilities are approximately equal, but for others they are not. This suggests that further analyses might be warranted on the statistical properties of methods such as ROLLOFF (Moorjani *et al.* 2011).

Throughout this article, we have assumed that admixture occurred in a single generation. This is a highly restrictive and, in most cases, unrealistic assumption. In real data analysis, the effects of such assumptions should be carefully considered. However, the basic conclusions regarding distributions of tract length as functions of T are still valid. Our results can be extended to more complicated scenarios of multiple admixture events, or continuous gene flow, by integrating over admixture times as in Pool and Nielsen (2009). For the PBT model, continuous gene flow, as well as overlapping generations, results in pedigrees that are still binary trees, but of uneven depth. Consequently, this same technique also allows us to relax the assumption of nonoverlapping generations.

In our mathematical analysis and simulations, we have assumed that recombination events occur according to a Poisson process and have ignored the possibility of crossover interference. For large values of T this approximation may be quite accurate, but for small values of T, crossover interference could potentially have a strong effect on the results, as illustrated in Figure 7. However, the transition rates of the ancestor-copying process are simple functions of the mapping function induced by the model of crossover interference. The binary tree process under other models of crossover interference with known mapping functions would typically still be mathematically tractable. Future methods for ancestry deconvolution and parametric admixture inference should seek to incorporate such mapping functions in addition to the non-Markovian properties of the ancestry process, which has been the main focus of topic of this article.

The Python programs used for the simulations can be obtained by contacting the author.

Acknowledgments

We thank members of the Nielsen and Slatkin labs for their help and encouragement and Nick Barton and two anonymous reviewers for their helpful comments. Funding was provided by National Institutes of Health grant 2R01HG003229-09.

Literature Cited

Baird, S., N. H. Barton, and A. M. Etheridge, 2003 The distribution of surviving blocks of an ancestral genome. Theor. Popul. Biol. 64(4): 451–471.

Ball, F., and V. T. Stefanov, 2005 Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. Math. Biosci. 196(2): 215–225.

Baran, Y., B. Pasaniuc, S. Sankararaman, D. G. Torgerson, C. Gignoux *et al.*, 2012 Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28(10): 1359–1367.

Barton, N. H., and B. O. Bengtsson, 1986 The barrier to genetic exchange between hybridising populations. Heredity 57: 357.

Barton, N. H., and A. M. Etheridge, 2011 The relation between reproductive value and genetic contribution. Genetics 188: 953–973.

Bickeböller, H., and E. A. Thompson, 1996a Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theor. Popul. Biol. 50(1): 66–90.

Bickeböller, H., and E. A. Thompson, 1996b The probability distribution of the amount of an individual's genome surviving to the following generation. Genetics 143: 1043–1049.

Brisbin, A., K. Bryc, J. Byrnes, F. Zakharia, L. Omberg *et al.*, 2012 PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed

ancestry from two or more populations. Hum. Biol. 84(4): 343–364.
Bryc, K., A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser *et al.*, 2010 Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA 107(2): 786–791.
Cannings, C., 2003 The identity by descent process along the chromosome. Hum. Hered. 56(1–3): 126–130.
Chapman, N. H., and E. A. Thompson, 2002 The effect of population history on the lengths of ancestral chromosome segments. Genetics 162: 449–458.
Dimitropoulou, P., and C. Cannings, 2003 RECSIM and INDSTATS: probabilities of identity in general genealogies. Bioinformatics 19(6): 790–791.
Donnelly, K., 1983 P. The probability that related individuals share some section of genome identical by descent. Theor. Popul. Biol. 23(1): 34–63.
Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
Fisher, R. A., 1949 *The Theory of Inbreeding*. Oliver & Boyd, Edinburgh, Scotland.
Gravel, S., 2012 Population genetics models of local ancestry. Genetics 191: 607–619.
Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3(4): 479–502.
Guo, S.-W., 1994 Computation of identity-by-descent proportions shared by two siblings. Am. J. Hum. Genet. 54(6): 1104.
Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5(10): e1000695.
Henn, B. M., L. R. Botigué, S. Gravel, W. Wang, A. Brisbin *et al.*, 2012 Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 8(1): e1002397.
Hey, J., 2010 Isolation with migration models for more than two populations. Mol. Biol. Evol. 27(4): 905–920.
Hey, J., and R. Nielsen, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.
Hoggart, C. J., E. J. Parra, M. D. Shriver, C. Bonilla, R. A. Kittles *et al.*, 2003 Control of confounding of genetic associations in stratified populations. Am. J. Hum. Genet. 72(6): 1492.
Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23(2): 183–201.
Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18(2): 337–338.
Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165: 2213–2233.
Marjoram, P., and J. Wall, 2006 Fast "coalescent" simulation. BMC Genet. 7(1): 16.
Martin, O. C., and F. Hospital, 2011 Distribution of parental genome blocks in recombinant inbred lines. Genetics 189: 645–654.
McVean, G. A., and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360(1459): 1387–1393.
Moorjani, P., N. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao *et al.*, 2011 The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. 7(4): e1001373.
Paşaniuc, B., S. Sankararaman, G. Kimmel, and E. Halperin, 2009 Inference of locus-specific ancestry in closely related populations. Bioinformatics 25(12): i213–i221.
Pool, J. E., and R. Nielsen, 2009 Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719.
Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5(6): e1000519.
Reich, D., N. Patterson, P. L. De Jager, G. J. McDonald, A. Waliszewska *et al.*, 2005 A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. Nat. Genet. 37(10): 1113–1118.
Rodolphe, F., J. Martin, and E. Della-Chiesa, 2008 Theoretical description of chromosome architecture after multiple backcrossing. Theor. Popul. Biol. 73(2): 289–299.
Sankararaman, S., G. Kimmel, E. Halperin, and M. I. Jordan, 2008 On the inference of ancestries in admixed populations. Genome Res. 18(4): 668–675.
Smith, M. W., N. Patterson, J. A. Lautenberger, A. L. Truelove, G. J. McDonald *et al.*, 2004 A high-density admixture map for disease gene discovery in African Americans. Am. J. Hum. Genet. 74(5): 1001–1013.
Stam, P., 1980 The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35: 131–155.
Stefanov, V. T., 2000 Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. Genetics 156: 1403–1410.
Sundquist, A., E. Fratkin, C. B. Do, and S. Batzoglou, 2008 Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res. 18(4): 676–682.
Tang, H., M. Coram, P. Wang, X. Zhu, and N. Risch, 2006 Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. 79(1): 1–12.
Tang, H., S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron *et al.*, 2007 Recent genetic selection in the ancestral admixture of Puerto Ricans. Am. J. Hum. Genet. 81(3): 626–633.
Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190: 1433–1445.
Walters, K., and C. Cannings, 2005 The probability density of the total IBD length over a single autosome in unilineal relationships. Theor. Popul. Biol. 68(1): 55–63.
Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. 55(3): 248–259.

Communicating editor: N. H. Barton

Appendix

Most of these proofs are by induction on the length of the dyadic interval(s) in question. Toward this end, we couple the two sides of Equation 1 by introducing independent ancestry-copying processes S_x and D_x and letting

$$N_x \equiv \frac{1}{2}R_x S_x + \frac{1}{2}(1 - R_x)(1 + D_x). \tag{A1}$$

By Equation 1, N_x is also an ancestry-copying process.

Proof of Theorem 1.

The theorem is trivially true in the case when this length is 1, *i.e.*, $A = I_{0,0}$.

Suppose that the theorem holds for dyadic intervals with length greater than or equal to 2^{-j} and let A be a dyadic interval with size 2^{-j-1}. Without loss of generality, assume that $A \subseteq [0, \frac{1}{2})$. Note that $|2A| = 2^{-j}$, so by the inductive hypothesis, $S_x\mathbf{1}\{S_x \in 2A\}$ is conditionally independent of $S_x\mathbf{1}\{S_x \notin 2A\}$ given $\mathbf{1}\{S_x \in 2A\}$. We use notation

$$S_x\mathbf{1}\{S_x \in 2A\} \perp S_x\mathbf{1}\{S_x \notin 2A\} \,|\, \mathbf{1}\{S_x \in 2A\}$$

to denote this. Since R_x is independent of S_x, it follows that

$$S_x\mathbf{1}\{R_x = 1, S_x \in 2A\} \perp S_x\mathbf{1}\{R_x = 1, S_x \notin 2A\} \,|\, \mathbf{1}\{R_x = 1, S_x \in 2A\}.$$

Finally, since $\mathbf{1}\{R_x = 0\} = \mathbf{1}\{R_x = 1, S_x \in 2A\} + \mathbf{1}\{R_x = 1, S_x \notin 2A\}$ and D_x is independent of everything in the above expression,

$$S_x\mathbf{1}\{R_x = 1, S_x \in 2A\} \perp S_x\mathbf{1}\{R_x = 1, S_x \notin 2A\} + \mathbf{1}\{R_x = 0\}(1 + D_x) \,|\, \mathbf{1}\{R_x = 1, S_x \in 2A\}.$$

By the definition of N_x, $N_x \in A \Leftrightarrow R_x = 1$, $S_x \in 2A$, so the theorem holds for dyadic intervals of length 2^{-j-1}, and consequently all dyadic intervals.

Proof of Theorem 2.

By Equation 1, the rate at which N_x leaves $I_{1,0}$ or $I_{1,1}$ is this same as the rate at which R_x switches from 1 to 0 or 0 to 1, respectively. This latter rate is equal to one, so the theorem holds for $j = 1$.

Assume that the theorem holds for all dyadic intervals with length 2^{-j}. Let I be a dyadic interval with length 2^{-j-1}. Note that $\mathcal{N}_0 \subset \sigma(\mathcal{R}_0, \mathcal{S}_0, \mathcal{D}_0)$ and without loss of generality, assume that $I \subset [0, 1/2)$, so that

$$\frac{1}{2}R_x S_x + \frac{1}{2}(1 - R_x)(1 + D_x) \in I \Leftrightarrow R_x = 1, S_x \in 2I.$$

We can use the law of total probability to find that

$$n_I = \lim_{x \downarrow 0} \frac{1 - \mathbb{P}_I(N_x \in I | \mathcal{N}_0)}{x} \tag{A2}$$

$$= \lim_{x \downarrow 0} \frac{1 - \mathbb{E}\left(\mathbb{P}(R_x = 1, S_x \in 2I | R_0 = 1, S_0 \in 2I, \mathcal{R}_0, \mathcal{S}_0, \mathcal{D}_0) \,|\, \mathcal{N}_0\right)}{x} \tag{A3}$$

$$= \lim_{x \downarrow 0} \frac{1 - \mathbb{E}\left(\mathbb{P}(R_x = 1 | R_0 = 1)\mathbb{P}(S_x \in 2I \,|\, S_0 \in 2I, \mathcal{S}_0) \,|\, \mathcal{N}_0\right)}{x} \tag{A4}$$

$$= \lim_{x \downarrow 0} \frac{1 - \left(\frac{1}{2} + \frac{1}{2}e^{-2x}\right)\mathbb{E}\left(\mathbb{P}(S_x \in 2I \,|\, S_0 \in 2I, \mathcal{S}_0) \,|\, \mathcal{N}_0\right)}{x} \tag{A5}$$

$$= \lim_{x \downarrow 0} \frac{\frac{1}{2} - \frac{1}{2}e^{-2x}}{x} + \lim_{x \downarrow 0} \left(\frac{1}{2} + \frac{1}{2} e^{-2x} \right) \frac{1 - \mathbb{E}\left(\mathbb{P}(S_x \in 2I \,|\, S_0 \in 2I, \mathcal{S}_0) \middle| \mathcal{N}_0 \right)}{x} \tag{A6}$$

$$= 1 + \mathbb{E}\left(\lim_{x \downarrow 0} \left(\frac{1}{2} + \frac{1}{2} e^{-2x} \right) \frac{1 - \mathbb{P}(S_x \in 2I |\, S_0 \in 2I, \mathcal{S}_0)}{x} \middle| \mathcal{N}_0 \right) \tag{A7}$$

$$= 1 + j, \tag{A8}$$

where the interchange of limits follows from the dominated convergence theorem and the inductive hypothesis that the limit n_{2I} is equal to j.

Proof of Theorem 3.

We show this by induction on the length of J. By Equation 1, the rate at which N_x enters J is the rate at which R_x switches from 1 to 0 or 0 to 1, which is 1. For $|J| = \frac{1}{2}$, $P(I, J) = \emptyset$, so $n_{I,J} = 1$ and the theorem holds.

To complete the proof by induction, we need a lemma:

Lemma 1. *For a dyadic interval I,*

$$\mathbb{P}(N_x \in I | \mathcal{N}_0, N_0 \in I', N_x \in I') = \frac{1}{2} + \left(\mathbf{1}\{N_0 \in I\} - \frac{1}{2} \right) \exp(-2x).$$

Proof of Lemma 1. We prove both claims by induction on the length of the dyadic interval I. For $I = \left[0, \frac{1}{2}\right)$, by Equation 1, the left-hand side reduces to $\mathbb{P}(R_x = 1 | R_0)$, which is equal to the right-hand side. The case of $I = \left[\frac{1}{2}, 1\right)$ is analogous, so the lemma is true for dyadic intervals of length $\frac{1}{2}$.

Assume that the lemma holds for dyadic intervals of length 2^{-j} and let I be a dyadic interval with length 2^{-j-1}. Without loss of generality, assume that $I \subset \left[0, \frac{1}{2}\right)$, so that by Equation 1,

$$N_x \in I \Leftrightarrow R_x = 1, S_x \in 2I.$$

Additionally, since $I' \subseteq \left[0, \frac{1}{2}\right)$, we also have that

$$N_x \in I' \Leftrightarrow R_x = 1, S_x \in 2I'.$$

Therefore,

$$\mathbb{P}(N_x \in I | \mathcal{N}_0, N_0 \in I', N_x \in I') = \mathbb{P}\left(R_x = 1, S_x \in 2I \middle| \mathcal{N}_0, S_0 \in 2I, R_0 = 1, S_x \in 2I', R_x = 1\right) \tag{A9}$$

$$= \mathbb{P}\left(S_x \in 2I \middle| \mathcal{N}_0, S_0 \in 2I, S_x \in 2I', R_0 = 1\right) \tag{A10}$$

$$= \mathbb{E}\left(\mathbb{P}\left(S_x \in 2I \middle| \mathcal{S}_0, S_0 \in 2I, S_x \in 2I'\right) \middle| \mathcal{N}_0, R_0 = 1\right). \tag{A11}$$

Since $2I$ has length 2^{-j} and S_x has the same distribution as N_x, the inductive hypothesis implies that

$$\mathbb{P}\left(S_x \in 2I \middle| \mathcal{S}_0, S_0 \in 2I, S_x \in 2I'\right) = \frac{1}{2} + \left(\mathbf{1}\{S_0 \in 2I\} - \frac{1}{2} \right) \exp(-2x).$$

Furthermore, since we are conditioning on $R_0 = 1$, $\{S_0 \in 2I\} = \{N_0 \in I\} \in \mathcal{N}_0$. As a result, the conditional expectation evaluates to

$$\mathbb{P}\left(N_x \in I \middle| \mathcal{N}_0, N_0 \in I', N_x \in I'\right) = \frac{1}{2} + \left(\mathbf{1}\{N_0 \in I\} - \frac{1}{2} \right) \exp(-2x),$$

so the lemma holds for dyadic intervals of length 2^{-j-1} and, consequently, all dyadic intervals with length less than 1. □

Assume that the rate at which N_x transitions from any dyadic interval to a disjoint dyadic intervals of length 2^{-j} is as the theorem states and let J be a dyadic interval with length 2^{-j-1}. To each dyadic interval I, we associate the random variable

$$T_I = \sup\{x < 0 : N_x \in I\}.$$

Note that $\max(T_I, T_{I^*}) = T_{I'}$ and $\{N_{T_{I'}} \in I \Leftrightarrow T_J > T_{J^*}\}$, so by the lemma,

$$\begin{aligned}\mathbb{P}(N_x \in I \,|\, \mathcal{N}_{T_{I'}}, N_x \in I') &= \frac{1}{2} + \left(\mathbf{1}\{N_{T_{I'}} \in I\} - \frac{1}{2}\right)\exp(2(T_{I'} - x)) \\ &= \frac{1}{2} + \left(\mathbf{1}\{T_J > T_{J^*}\} - \frac{1}{2}\right)\exp(2(T_{I'} - x)).\end{aligned}$$

Additionally, for $T_I < x < 0$, $N_x \notin I$, so by Theorem 1, the left-hand side also equals $\mathbb{P}(N_x \in I | \mathcal{N}_0, N_x \in I')$. So for J, a dyadic interval of size 2^{-j-1},

$$\begin{aligned}n_{I,J} &= \lim_{x \downarrow 0} \frac{\mathbb{P}_I(N_x \in J | \mathcal{N}_0)}{x} \\ &= \lim_{x \downarrow 0} \frac{\mathbb{P}_I(N_x \in J \,|\, \mathcal{N}_0, N_x \in J')\mathbb{P}_I(N_x \in J' \,|\, \mathcal{N}_0)}{x} \\ &= \lim_{x \downarrow 0} \mathbb{P}(N_x \in J | \mathcal{N}_0, N_x \in J') \lim_{x \downarrow 0} \frac{\mathbb{P}_I(N_x \in J' \,|\, \mathcal{N}_0)}{x} \\ &= \left(\frac{1}{2} + \left(\frac{1}{2} - \mathbf{1}\{T_J > T_{J^*}\}\right)\exp(-2T_{J'})\right) \prod_{i \in P(I, J')} \frac{1}{2} + \left(\mathbf{1}\{T_i > T_{i^*}\} - \frac{1}{2}\right)\exp(-2T_{i'}) \\ &= \prod_{i \in P(I,J)} \frac{1}{2} + \left(\mathbf{1}\{T_i > T_{i^*}\} - \frac{1}{2}\right)\exp(-2T_{i'}).\end{aligned}$$