Population Genetics Models of Local Ancestry

Simon Gravel[1]

Genetics Department, Stanford University, Stanford, California 94305-5120

DNA sequencing is an invaluable tool for understanding demographic relationships between populations. Even with a limited number of genetic markers, measured across individuals and populations, it is often possible to estimate relatedness between populations, ancestry proportions in admixed populations, or sex-biased gene flow. The availability of dense genotyping platforms and high-throughput sequencing technology has enabled refined analyses of genetic diversity.

Because of recombination, different loci along an individual genome can reveal different aspects of its ancestry. Consider a sample and its ancestral population at some time T in the past, and suppose that we give ancestral individuals subpopulation labels, defining *source* populations. These labels are typically chosen to represent subgroups that have increased genetic homogeneity due to cultural or geographic reasons. Then a simple summary of the demographic trajectory of a sampled allele is the source population from which it originated. We say that an individual is "admixed" if it draws ancestry from multiple source populations—thus admixture is not an intrinsic property of individuals, but depends on our choice of labels and time T. An example of subpopulation labels often used to study human populations in the Americas are the European, Native American,

and West African populations prior to the advent of massive intercontinental travel. Many routines have been proposed to infer the source population along the genome of admixed individuals (Ungerer *et al.* 1998; Tang *et al.* 2006; Falush *et al.* 2003; Hoggart *et al.* 2004; Patterson *et al.* 2004; Sankararaman *et al.* 2008; Bercovici and Geiger 2009; Price *et al.* 2009). These typically proceed by locally matching an admixed genome to panel populations chosen as proxies for the source populations, revealing a mosaic of tracts of continuous ancestry (Figure 1). In this work we use PCAdmix (Brisbin 2010), a heuristic approach for local ancestry inference. PCAdmix first divides the genome in windows of typical width of 10–50 kb. For each window, the probability that the sample haplotype originates from any of the panel populations is estimated on the basis of the position in PCA space. Finally, PCAdmix uses these probabilities as emission probabilities of a hidden Markov model and ancestry is inferred via Viterbi decoding.

Local ancestry patterns have been used to identify disease loci (see Seldin *et al.* 2011 and references therein) and to search for regions experiencing selection (Tang *et al.* 2007; Bhatia *et al.* 2011). They also provide hints about the history of migration (Pool and Nielsen 2009). The purpose of this article is to understand and model the observed ancestry patterns on the basis of detailed demographic models, to learn about human demography, and to empower selection and association scans. In particular, we are interested in the length distribution of the continuous ancestry tracts, and the variation in ancestry proportions across chromosomes and individuals.
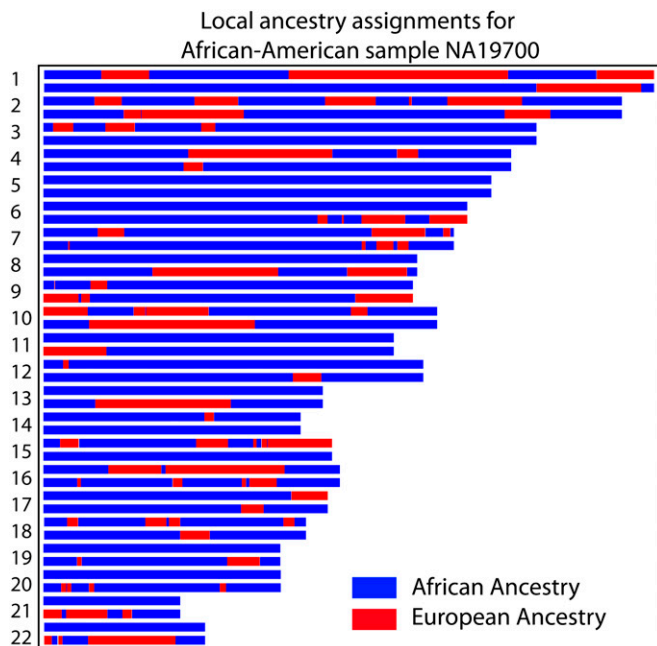
Figure 1 Local ancestry across 22 autosomes for an African-American individual inferred by PCAdmix, a local ancestry inference software (Brisbin 2010) using HapMap European (CEU) and Yoruba (YRI) as source populations. The majority of the genome is inferred to be of African origin (blue), but a significant fraction of the genome is inferred to be of European origin (red). The purpose of this article is to model the distribution of ancestry assignments in such admixed individuals.

A dominant stochastic process leading to these patterns is recombination, which tends to break down segments of continuous ancestry in admixed individuals. As a result, the length of continuous ancestry tracts tends to be shorter for more ancient admixture. The tract-length distribution is sensitive to details of recent migration (*i.e.*, tens of generations) and is thus complementary to analysis based on the joint site-frequency spectrum (Gutenkunst *et al.* 2009; Gravel *et al.* 2011), which is more sensitive at time scales of hundreds to thousands of generations.

Recently, Pool and Nielsen (2009) proposed a model in which a target population receives migrants from a source population, initially at a constant rate m_2. Starting at a time T in the past, the rate changes to m_1. In this model, back migrations are not allowed, recombinations within migrant chromosomes are neglected, and tracts shorter than a cutoff value are forgotten (since migration occurs over an infinite period, this is necessary to avoid having a genome completely replaced by migrants). Assuming that recombinations occur according to a Poisson process, these approximations allow for an analytical solution for the distribution of tract lengths, which was used to infer demographic events in mice (Pool and Nielsen 2009). This model is limited to admixture proportions weak enough that recombinations between migrant chromosomes are unlikely. A second limitation is that the model assumes two epochs of constant migration rate, which might or might not be the most appropriate for a given population. The special case $m_2 = 0$ has been used to infer de-

mographic histories in humans for North African individuals (Henn *et al.* 2012).

Here we propose a more general approach to predicting the distribution of tract lengths that can accommodate both time-dependent and strong migration. This approach builds on that of Pool and Nielsen (2009) but introduces multiple improvements. First, general time-dependent migrations can be considered. Second, recombinations between tracts of the same ancestry are not neglected, allowing for the modeling of strong migration and the simultaneous study of tracts of multiple ancestries. Third, chromosomal end effects are modeled explicitly. Fourth, our model can be modified to incorporate errors in tract assignments. As in the Pool and Nielsen approach, we model recombination as a Poisson process with a unit rate per morgan, and the recombination map is taken to be identical across populations [a reasonable approximation at the centimorgan scale (Wegmann *et al.* 2011)]. To perform demographic inference, we further require that local ancestry inference can be performed to high accuracy using one of the methods mentioned above. Whether this can be done depends on the degree of divergence of the ancestral populations (or sources), the availability of data for panel populations that are good proxies for the sources, and the possibility of accurately phasing diploid genomes.

Admixture history also leaves a trace in the variance in admixture proportions across individuals, as stochastic mating and recombination tend to uniformize ancestry proportions with time (Verdu and Rosenberg 2011). Generalizing the models of (Ewens and Spielman 1995; Verdu and Rosenberg 2011) to include the effects of recombination in a finite genome and drift, we show that after a discrete admixture event, the variance decays in time in three consecutive regimes, first exponentially as differences in individual genealogies average out, then linearly as recombination creates shorter tracts, and finally exponentially again as drift fixes local ancestry along a chromosome. A simple approximate equation captures all three regimes accurately. By contrast, variance in continuous migration models is dominated by the first regime, and the expressions from the model of Verdu and Rosenberg (2011) are reasonably accurate (see Appendix 3).

In general, distinguishing the effects of population structure and time-dependent patterns of gene flow is not straightforward, and the inference problem is prone to overfitting, as is the case, *e.g.*, for inference based on the site-frequency spectrum (Myers *et al.* 2008). However, our analysis shows that tract lengths, and more generally ancestry correlation patterns, can help resolve subtle differences in patterns of historical gene flow. An implementation of the proposed methods for tract length modeling, called *tracts*, is available at http://tracts.googlecode.com.

Theory

Admixture models: definitions and global properties

We construct a model for the admixture of diploid individuals that takes into account recombination, drift, migration,
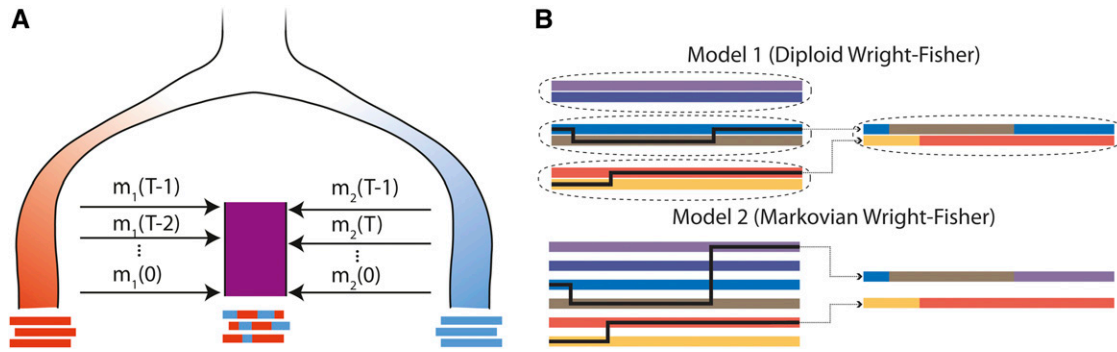
Figure 2 (A) Illustration of an admixture model starting at generation $T - 1$, where the admixed population (purple) receiving $m_i(t)$ migrants from diverged red ($i = 1$) and blue ($i = 2$) source populations at generation t. If these are statistically distinct enough, it is possible to infer the ancestry along the admixed chromosomes. Independent of our statistical power to infer this detailed local ancestry, the mosaic pattern may leave distinct traces in genome-wide statistics, such as global ancestry or linkage patterns. (B) Gamete formation in two versions of the Wright–Fisher model with recombination. In model 1, diploid individuals are generated by randomly selecting two parents and generating gametes by following a Markov paths along the parental chromosomes. In model 2, gametes are generated by following a Markovian path across the parental allele pool. Both models have the same distribution of crossover numbers and are equivalent for genomic regions small enough that multiple crossovers are unlikely. Model 1 is more biologically realistic and is used in the simulations, whereas model 2 is more tractable and is used for inference and analytic derivations.

and finite chromosome length. Since a full coalescent treatment of these effects is computationally prohibitive (Griffiths and Marjoram 1996), we simplify the model to consider only the demography of our samples up to the first migration event, T generations ago. We label generations $s \in \{0, 1, 2, \ldots, T - 1\}$, and the total fraction of the population $m(s)$ that is replaced by migrants in a generation s can be subdivided in contributions $m_p(s)$ from M migrant populations: $p \in \{1, \ldots, M\}$. We treat the replacement fraction $m_p(t)$ as deterministic, while the replaced individuals are selected at random (see Figure 2). Generations follow a Wright–Fisher model with random mating in a population with $2N$ genome copies, each with K finite chromosomes of morgan length $\{L_i\}_{i=1\ldots, K}$. We consider two different variations of the Wright–Fisher model with recombination.

The first variation (model 1) is meant to be the most biologically motivated and will be used for all simulations. Starting from a finite parental diploid population of size N, we first replace $m(s)N$ randomly selected individuals with diploid migrants. Diploid offspring are generated by drawing one gamete from each of two randomly selected diploid parents. Gamete formation is a Markov path with transition rate of one transition per morgan across the two parental chromosomes (see Figure 2B).

Model 1 results in long-range, non-Markovian correlations along the genome. This complicates the modeling without necessarily having a large effect on most global statistics. We therefore also consider a more tractable model (model 2) in which gametes are drawn from the migrant populations with probability $m(s)$ and are otherwise generated by following a Markov path along all nonmigrant parental gametes (see Figure 2). The reason for singling out new migrants is that it is possible to generate their gametes as in the more realistic model 1, without sacrificing tractability. Model 2 may not capture all long-range correlations in ancestry but it has the correct distribution of crossovers and

for small portions of the chromosomes is very similar to model 1: the only difference is that each draw from the parental gamete pool is independent in model 2, whereas the fact that a diploid individual can have multiple offspring induces a small degree of correlation between draws in model 1. Unless otherwise stated, we calculate all population-wide statistics after the migration step, but before gamete generation.

Model 2 is reminiscent of the Li and Stephens (2003) copying model used in HAPMIX (Price *et al.* 2009), as it also neglects back-and-forth recombinations due to multiple crossovers during a single meiosis. The purpose of the models are different, in that the current Markov models attempt to simulate gamete formation from parental chromosomes and represent evolution in time, whereas the Li and Stephens model attempts to simulate an unobserved haplotype on the basis of haplotypes from the same generation. The Markov ancestry transition model used in HAPMIX (and many other local ancestry inference software) corresponds to a special case of model 2 when each population contributes migrants at a single generation.

Local ancestry patterns are sensitive to the three stochastic processes of migration, recombination, and random genetic drift. Where possible, we take all three effects into account. By contrast, we do not model the effects of population structure, of selection, and of population size fluctuations. We derive our results under the assumption that local ancestries can be determined exactly; the effects of misidentification are discussed throughout, together with possible correction strategies.

Given a history of migrations, it is relatively straightforward to calculate the expected population averages for ancestry proportions and tract lengths. If $m(s)$ is the total fraction of the population that is replaced by migrants, s generations ago, with $m_i(s)$ from population i, the expected ancestry from population i at a time t in the past is the sum

over generations s of migrant contributions $m_i(s)$ weighted by the survival probability $\prod_{s'=t}^{s-1}(1-m(s'))$ to time t. After the migration step, the ancestry proportions are

$$\alpha_i(t) = \sum_{s=t}^{\infty} m_i(s) \prod_{s'=t}^{s-1}(1-m(s')).$$

We can follow a similar procedure to obtain the expected density w_{ij} of ancestry switch points from population i to population j per morgan, replacing the amount of new migrants $m_i(s)$ by the density of new switch points, which are proportional to the recombination rate (assumed constant with unit rate in genetic units) and the expected fraction of the genome $h_{ij}(s)$ that is heterozygous with respect to ancestries i and j after generation s. In the gamete pool, we find

$$\mathbb{E}\big[w_{ij}(t)\big] = \sum_{s=t}^{\infty} h_{ij}(s) \prod_{s'=t}^{s-1}(1-m(s')).$$

The ancestry heterozygocity h_{ij} can be evaluated using a recursive equation (such as Equation A1), as in the case of allelic heterozygosity. In the absence of drift, $h_{ij}(s) = (1-m(s))\alpha_i(s+1)\alpha_j(s+1)$. In the population (before gamete generation), the sum over s starts at $t+1$ rather than t. The expected number of switch points per morgan at time 0 is therefore

$$w_{ij} \equiv \mathbb{E}\big[w_{ij}(0)\big] = \sum_{s=1}^{\infty} h_{ij}(s) \prod_{s'=0}^{s-1}(1-m(s')).$$

To estimate the expected tract length $\mathbb{E}[x_i(t=0)]$ for ancestry i on a chromosome of length L, we divide the expected length covered by this ancestry, $\alpha_i(0)*L$, by the expected number of tracts of this ancestry, which is $\frac{L}{2}\sum_j w_{ij} + \alpha_i(0)$ since each tract must begin and end by an ancestry switch or by the end of the chromosome. We thus find

$$\mathbb{E}[x_i(t=0)] = \frac{2\alpha_i(0)L}{L\sum_j w_{ij} + 2\alpha_i(0)}.$$

If the demographic model under consideration has a single parameter, such as the timing of a single pulse of migration, demographic inference can proceed from this single estimate. However, the mean tract length may be largely dependent on the number of very short tracts that are difficult to detect; this statistic is therefore sensitive to false-positive and false-negative ancestry switches. Here we are interested in studying more detailed models of migration and their impact on tract-length distribution.

Tract-length distribution

For illustration, we first consider a source population (Blue), and a target population (Red), with a single, infinitely long diploid chromosome. At generation $t = T - 1$, a fraction m
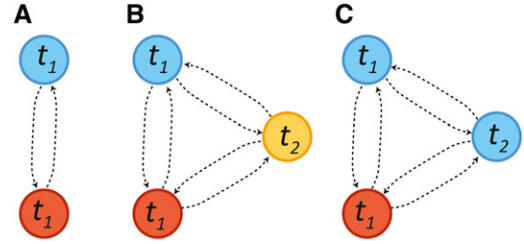

Figure 3 (A) A two-state Markov model for ancestry along a chromosome for a single pulse of migration at time t_1. Tract-length distributions are exponential. (B) A three-population Markov model with a pulse of blue and red ancestry at time t_1 followed by a pulse of migration from the yellow population at time t_2. All tract-length distributions are exponential. (C) A two-population model in which the blue population contributes migrants at generation t_1 and t_2. The distribution of blue ancestry tracts is no longer exponential, as we cannot detect transitions between blue states.

of population Red is replaced by individuals from population Blue. Consider the Markovian Wright–Fisher model discussed above (model 2). In this model, the position of the closest recombination to either side of a point along an infinite chromosome is exponentially distributed and there is no memory of previously visited states along a chromosome. The chromosomes resulting from this admixture process can therefore be modeled as a continuous-time Markov model with a Red and a Blue state (Figure 3A), where each recombination event corresponds to a Markov transition and the continuous Markov "time" corresponds to the position along the chromosome. The transition rate out of a state in this model is proportional to the number of recombinations, namely $t - 1$ per morgan: since recombinations within first-generation migrants do not induce ancestry changes, and we suppose that we sequence somatic cells at generation 0, recombination can occur only during gamete formation at generations $1, ..., t - 1$. If a recombination occurs, the probability of transitioning is m to the Red state and is $(1 - m)$ to the Blue state.

We are interested in the length distribution of continuous segments in the Blue or Red ancestry, independent of the number of within-ancestry transitions, which are difficult to detect. We avoid these complications by setting the self-transition rates to zero: this does not affect the trajectories, but now all transitions change the ancestry. We therefore have the model shown in Figure 3A, and the distribution of tract lengths $\phi_i(x)$ is equal to the exponential distribution of distance between Markov transitions:

$$\begin{aligned} \phi_R(x) &= m(t-1)e^{-m(t-1)x} \\ \phi_B(x) &= (1-m)(t-1)e^{-(1-m)(t-1)x}. \end{aligned} \tag{1}$$

Note that the distribution is ill-defined for $t = 1$, since this situation produces tracts that are infinite in the infinite-chromosome limit.

Multiple populations, discrete migration: As long as the migration from each population is limited to a single generation and the target population is infinitely large, model 2

produces Markovian trajectories along ancestry states. To see this, consider a point x along the genome in a segment from ancestry p that arrived t generations ago. As before, the distance to the first recombination event downstream from x is exponentially distributed (with rate $t-1$), and the timing τ of the recombination is uniform on $(1, t-1)$. Moreover, since gametes in model 2 are formed by following a Markov path in the parental gamete pool, the probability of observing ancestry p' downstream from the recombination is proportional to the ancestry proportions in the parental pool at the time τ of the recombination. Thus we have the discrete transition rate

$$M(p \to p') = \sum_{\tau=1}^{t-1} P(p'|\tau)P(\tau|p) = \sum_{\tau=1}^{t-1} \frac{\alpha_{p'}(\tau+1)}{t-1},$$

which depends only on the time of arrival t of ancestry p. We note that the Markov property over ancestry states would be lost in model 1, because the state downstream of the recombination is correlated with upstream states. Drift reduces the transition rates and also breaks the Markov property: mitigation strategies are discussed in Appendix 1. The Markov property over ancestry states is also lost if a population contributes migrants over many generations, and our next step is to restore the Markov property in this situation by extending the state space.

General incoming migration in the absence of drift: We now allow for general incoming migration histories that start at a time $T-1$ in the past. For each generation $t \in \{0, \ldots, T-1\}$, a fraction $m_p(t)$ of the individuals from the target populations are replaced by individuals from the source population p, with $m(t) = \sum_p m_p(t) \le 1$. We further impose that the first generation is composed of nonadmixed individuals: $m(T-1) = 1$. Since the ancestry switches are no longer Markovian in the general migration case, it is convenient to consider states defined by both ancestry p and time of arrival t. Intuitively, we may imagine that we have a large number of migrant populations (p, t), each contributing migrants over a single generation (see Figure 3, B and C). Here the Markov property is maintained, but ancestry states can now correspond to multiple Markov states.

We first calculate the transition rates between states (p, t) as we did for the discrete migration case. First, the probability of encountering state (p, t) downstream from a recombination that occurred at time τ is

$$P(p, t|\tau) = \Theta(t - (\tau+1))m_p(t) \prod_{t'=\tau+1}^{t-1} (1 - m(t')),$$

where

$$\Theta(s) = \begin{cases} 1 & s \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

is the Heaviside function.

As before, given a point x in state (p, t), the position of the next downstream recombination is exponentially distributed with rate $t-1$, and the time of this recombination is uniformly distributed on $(1, t-1)$. In the two Wright–Fisher models considered here, states on either side of the recombination are uncorrelated, and we can write the discrete transition probabilities

$$R(p, t \to p', t') = \sum_{\tau=1}^{\min(t,t')-1} \frac{P(p', t'|\tau)}{(t-1)},$$

which is independent of p. The continuous transition rate is obtained by multiplying the discrete transition rate by the continuous overall transition rate $t-1$:

$$Q(p, t \to p', t') = m_{p'}(t') \sum_{\tau=1}^{\min(t,t')-1} \prod_{s=\tau+1}^{t'-1} (1 - m(s)). \quad (2)$$

These transition probabilities are valid for both Wright–Fisher models in the infinite-population size limit. Since model 2 is Markovian, these transition rates are sufficient to fully specify the ancestry state model.

Given the transition matrix Q, we can use standard tools for the study of Markov chains to efficiently estimate the length distribution of excursions on Markov states corresponding to a single ancestry. In Appendix 2, we first derive results under the approximation that chromosomes are infinitely long. We account for finite chromosomes by studying the distribution of tract lengths in finite windows, randomly chosen along the infinite chromosomes. We thus obtain a distribution of tracts $\phi_p(x)$ for each population p. To compare these predictions to observed data, a computationally efficient strategy is to bin data by tract length and treat the observed counts in each bin as an independent Poisson variable with mean obtained by integrating $\phi_p(x)$ over the bin range.

Short ancestry tracts are likely to have both elevated false-positive and false-negative rates, and inference based on such tracts is likely to be biased, whereas longer tracts can be detected with increased confidence. Following Pool and Nielsen (2009), we therefore perform inferences using only tracts longer than a cutoff value C. We should emphasize that a large number of uniformly distributed spurious short tracts may still affect the distribution of longer tracts, making non-exponential distributions look more exponential. Therefore, significant assignment error may cause an underestimation of the amount of continuous migration. By contrast, drift would tend to reduce the transition rates and cause underestimates of the time since admixture (see Appendix 1).

Variance among individuals

We now consider the variance across individuals in total migrant ancestry X^p from population p, measured as a proportion of the morgan length of the genome whose origin is from p. The variance in ancestry can be separated in two components, which we label the genealogy variance and

assortment variance. The genealogy variance is due to a different number of migrant ancestors; if a randomly chosen fraction m of the population is replaced by migrants at each generation, a fraction m^2 of individuals will have two migrant parents, $2m(1-m)$ will have one migrant parent, and $(1-m)^2$ will have none. The assortment variance accounts for the fact that two individuals with the same genealogy can vary in their genetic ancestry proportions, since not all ancestors contribute the same amount of genetic material to an individual. Recombination and the independent assortment of chromosomes tend to reduce such variance.

We can use the law of total variance, conditioning over the genealogies g, to isolate these two contributions to the variance $\mathrm{Var}(X^p)$:

$$\mathrm{Var}(X^p) = \mathrm{Var}_g[\mathbb{E}(X^p|g)] + \mathbb{E}_g[\mathrm{Var}(X^p|g)].$$

Here $\mathbb{E}[X^p|g]$ is the fraction of migrant ancestry from population p, based on the genealogy g. Alternatively, this can be thought of as the infinite-sites expectation for the ancestry proportions. The first term therefore represents the genealogy variance in ancestry, whereas the second term represents the assortment variance. Because of random chromosome assortment, the variance in ancestry among chromosomes is informative of the assortment variance. We discuss in Appendix 3 how, in the absence of drift, the variance can thus be broken down in these two components without requiring a demographic model. We discuss below how to obtain expectations for each components given a specific demographic model.

Genealogy variance: To ease calculations of the genealogy variance, we neglect correlations due to overlap between individual genealogies and describe each individual as being sampled from an independent genealogy (in a randomly mating population, this amounts to neglecting drift). In this model, the genealogy variance $\mathrm{Var}_g(\mathbb{E}[X^p|g])$ is easily calculated. Considering the genealogy g of a nonmigrant sample up to T generations ago (we label the current generation 0, and the generation with the first migrants $T-1$), we first note that

$$\mathbb{E}[X^p|g] = \frac{1}{2^{T-1}}\sum_{i=1}^{2^{T-1}} z_i^p,$$

where z_i^p is 1 if there has been a migrant on the lineage leading from the root to leaf i and 0 otherwise. Results with continuous admixture since time immemorial can be obtained by taking a limit $T \to \infty$. In such cases, the approximation of independent pedigrees eventually breaks down, but the resulting expression might remain approximately correct if the majority of present-day genomes originate from recent migrants.

The expectation over genealogies g and assortments $\mathbb{E}_g[\mathbb{E}[X|g]]$ is then $\alpha_p(0)$. The calculation of $\mathbb{E}_g[\mathbb{E}[X|g]^2]$ is also straightforward if we can calculate the expectation $\mathbb{E}_g[z_i^p z_j^p]$. For $z_i^p z_j^p$ to be nonzero, we must have had a migrant

either on the common branch leading to the two leafs i and j, or one migrant on each of the separate branches,

$$\mathbb{E}_g\left[z_i^p z_j^p\right] = \sum_{s=0}^{T-1-d_{ij}} m_p(s) \prod_{s'=0}^{s-1}(1-m(s')) + \alpha_p^2(T-d_{ij}) \prod_{s=0}^{T-1-d_{ij}}(1-m(s))$$

$$\equiv e(d_{ij}),$$

$$(3)$$

with d_{ij} half the tree distance between leafs i and j. Then we can write the sum over half distances, weighted by the number of leaf pairs at each distance:

$$\mathbb{E}_g\left[\mathbb{E}[X|g]^2\right] = \sum_{d=1}^{T-1} 2^{d-T}e(d) + \alpha(0)/2^{T-1}. \quad (4)$$

Since $\mathbb{E}_g[\mathbb{E}[X|g]] = \alpha(0)$, we have

$$\mathrm{Var}_g(\mathbb{E}[X|g]) = \sum_{d=1}^{T-1} 2^{d-T}e(d) + \alpha(0)\left(\frac{1}{2^{T-1}} - \alpha(0)\right).$$

In the two-population pulse model, with $m_{p=1}(t) = m\delta_{t,T-1}$, we have the expected $\mathrm{Var}_g(\mathbb{E}[X|g]) = m(1-m)/2^{T-1}$, with a rapid exponential decay of the variance as a function of T. By contrast, if we have continuous migration of population p in a target population, with, $m_i^p = m\Theta(T-i-1)$, the variance reads

$$\mathrm{Var}_g(\mathbb{E}[X|g]) = \frac{2^{-(T-1)}m(1-m)^T\left([2(1-m)]^T - 1\right)}{1-2m}, \quad (5)$$

with a more complex dependence of the variance on T. Finally, in the case in which two populations provide respectively pm and $(1-p)m$ migrants to a target population at each generation since the beginning of time, we have the simple expression

$$\mathrm{Var}_g(\mathbb{E}[X|g]) = \frac{2p(1-p)m}{1+m}. \quad (6)$$

This expression supposes that the variance is calculated after migration occurs. If variance is calculated before replacement by migrants, the factor of 2 disappears, and we recover Equation 47 in Verdu and Rosenberg (2011).

Assortment variance: To study the global ancestry variance due to assortment, a natural starting point is to consider the ancestry variance at a particular point in the genome. In a randomly mating population with two ancestries, the variance in ancestry at a site is $h/2$, where h is the ancestry heterozygocity at that site. The ancestry heterozygocity can be calculated using the same recursive strategy commonly used for allelic heterozygocity (Equation A1). The case of three or more ancestries can be reduced to two ancestries by singling out one ancestry and pooling the others. As a specific example, in the case of a pulse migration with migration rates m and $1-m$ at generation $T-1$, the heterozygocity at generation 0 is

$$h_0 = \left(1 - \frac{1}{2N}\right)^{T-1} 2m(1-m). \qquad (7)$$

We wish to combine these local variances into an expression for the genome-wide variance. In Appendix 3 we provide a derivation of the expected ancestry variance using Markov models. Here, to obtain a simple approximation for the migration pulse model, we imagine that the length of the genome is divided in n tracts by uniformly drawing $n-1$ separators. We suppose that the ancestry is chosen independently on each segment, with variance $h_0/2$. Then the variance in ancestry in the large-n limit is

$$\mathbb{E}_g[\mathrm{Var}(X^p|g)] \simeq \frac{h_0}{n}.$$

The effect of drift is therefore captured by the decay of ancestry heterozygosity with time, whereas the effect of recombination is captured by the number of independent tracts n, which is proportional to the number of recombinations. In the case of a pulse of migration T generations ago without drift, we write $n = 1 + (T-2)L_i$ for a single haploid chromosome (the 1 accounts for the chromosome edge and can be neglected for large TL_i), and $2K + 2(T-2)L$ for a diploid genome with K chromosome pairs of total length $L = \sum_i L_i$. Thus the total variance reads

$$\mathrm{Var}(X^p) = \frac{m(1-m)}{2^{T-1}} + \frac{2m(1-m)(1-1/2N)^{T-1}}{2K + 2(T-2)L}. \qquad (8)$$

Even though it neglects the effect of drift on the number of independent tracts n, this expression provides excellent quantitative agreement with simulations over multiple regimes (Figure 6). If we model the variation over time of the population ancestry proportion as a random walk with decreasing step size $\mathrm{Var}(X^p)/N$, the variation will be dominated by the genealogy variance, which after an infinite time contributes a finite variance of $\sigma^2 = (m(1-m))/N$. Thus for an initial population of 100 individuals divided equally between two ancestries, we can expect the final ancestry proportions to be $0.5 \pm 2\sigma = 0.5 \pm 0.1$, a relatively modest uncertainty given the small population size. Assortment variance for continuous migration models is discussed in Appendix 3.

Comparison with Simulation and Experimental Data

In this section we first present results of Wright–Fisher simulations, comparing our model predictions to the simulation results. We then consider the HapMap African-American panel, for which we performed local ancestry inference and analyzed the tract length distribution.

Tract lengths

We performed a 30-generation diploid Wright–Fisher simulation (using model 1; see Figure 2) of 10000 chromosomes of length 1 M with continuous gene flow from population 1 into a population initially composed of individuals from pop-
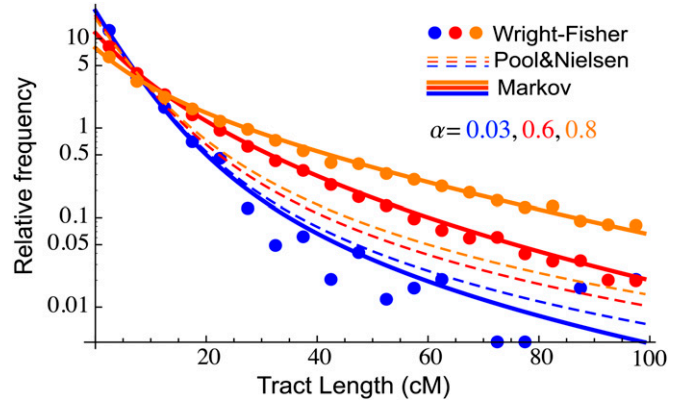

Figure 4 Comparison of the Markov model, the Pool and Nielsen (2009) prediction, and Wright–Fisher simulation for migrant tract length distributions. Each dot represents the normalized number of ancestry tracts whose length is contained in one of 20 bins. The simulation followed 10000 chromosomes over 30 generations, with constant migration rates $m = 0.001, 0.03, 0.05$ giving rise to final ancestry fractions of $\alpha = 0.03, 0.6, 0.8$. Since recombination between migrant tracts were neglected in (Pool and Nielsen 2009), the results depart significantly from simulation at high migration, whereas the Markov model is accurate in the three regimes.

ulation 2. We considered three different migration intensities, namely $m_1 = 0.001, 0.03$, and 0.05 per generation. We kept track of the ancestry of each segment during the simulation, so that the continuous ancestry tracts could easily be tabulated. In Figure 4, we compare the observed histograms of tract lengths for population 1 (dots) to predictions from Equation 10 in Pool and Nielsen (2009) (dashed lines) and to predictions from the Markovian Wright–Fisher model (model 2 on Figure 2), using rates from Equation 2 and implemented as described in Appendix 2 to account for finite chromosome length (solid lines). As expected, the predictions of the two models are similar when migration rates are low and differ substantially when we depart from the weak migration assumptions of the Pool and Nielsen model (see Figure 4). The Markov model predictions are in good agreement with the simulations over the range of models considered, including when the migrant population becomes the majority population.

We now consider the HapMap African-American panel (ASW) (International HapMap 3 Consortium 2010) and focus on 20 unrelated samples that were trio phased, to reduce biases due to phasing errors. We obtained local ancestry inferences using PCAdmix (Brisbin 2010), using 132 unrelated HapMap samples from Europe (CEU) and 204 from West Africa (YRI) as reference panels. We used windows of size 0.3 cM for the HMM and based our inferences on the number of tracts longer than 10 cM. We pooled tracts in 50 bins according to tract length (chromosomes with no ancestry switches were in a separate bin independent of the chromosome length) and calculated model likelihood assuming that counts in each bin are Poisson distributed with mean given by the model predictions for this bin.
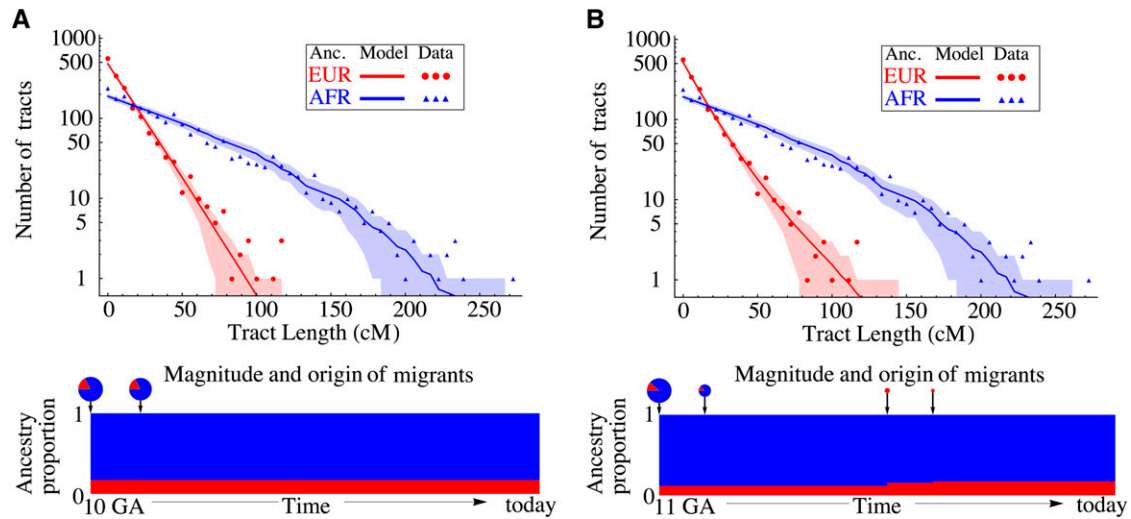
Figure 5 Distribution of continuous ancestry tract lengths in 20 HapMap African-American (ASW) trio individuals [as inferred by PCAdmix (Brisbin 2010), a local ancestry inference software], compared with predictions from a single-pulse migration model (top) and a model with subsequent European migration (bottom). Each dot represents the number of continuous ancestry tracts whose length is contained in one of 50 bins. The shaded area marks the 68.3% confidence interval based on the model. The second model, in which over 30% of European origin in the ASW samples is quite recent, provides a sufficiently better fit to justify the extra parameters (likelihood-ratio test, $P = 0.002$).

We compared inferences on the basis of two different models; (a) a "pulse" model, with a single migration event, and (b) a two-pulse model, with a subsequent migration of Europeans (Figure 5). Model b has two additional parameters, corresponding to time and proportion of the subsequent European migration. A likelihood-ratio test shows that $\ln(\mathcal{L}_b/\mathcal{L}_a) \simeq 7$. To establish the significance of the extra two parameters, we simulated 1000 random tract-length distributions from the maximum-likelihood model a, and obtained maximum-likelihood estimates for both models. The probability of obtaining such a likelihood ratio under model a is $P = 0.002$.

Ancestry proportions and variance

Simulations of 80 individuals, each with 22 autosomal chromosomes of realistic lengths (namely 2.78, 2.63, 2.24, 2.13, 2.04, 1.93, 1.87, 1.70, 1.68, 1.79, 1.59, 1.73, 1.27, 1.16, 1.26, 1.35, 1.30, 1.19, 1.08, 1.08, 0.62, 0.73 M, for chromosomes 1–22, respectively) and 30% of initial admixture proportion, illustrate many of the effects predicted in variance models. The global ancestry proportions and fraction of sites heterozygous for ancestry fluctuate considerably over the first few generations, but the fluctuations decrease in time as ancestry proportions approach a fixed value and ancestry heterozygocity decays following Equation 7.

Figure 6 shows that the variance in ancestry across individuals follows three different regimes; first, the variance is dominated by the genealogy variance, with a rapid exponential decay. After about 10 generations, the assortment variance starts to dominate, and decays polynomially due to recombination until drift becomes important, where an exponential decay is resumed, although at a much reduced rate.

Equation 8 captures these three regimes in quantitative detail—the average variance over 50 independent simula-

tions follows the model prediction closely. The continuous migration case, in which genealogy variance tends to dominate, is discussed in Appendix 3.

Comparing the ancestry variance from the African-American data to those predicted by the demographic models, we find that the pulse model predicts a genealogy variance of 0.0005, whereas the variance in the model with two distinct pulses is 0.002. The total variance in the African-American sample is 0.0047, of which we infer that 0.0041 is due to genealogy variance (using the method described in Appendix 3). Thus the model with two pulses of migration is again more realistic than the single pulse model; the fact that it still underestimates the variance can be due to a combination of factors that have not been modeled: our demographic model may be underestimating low level, very recent migration because of the parameterization as two discrete pulses of migration, and both population structure and errors in ancestry assignment may be adding to the observed variance.

Discussion

Limitations and possible improvements

A limitation of all demographic inference methods is that the model space is very large, and the information available to learn about the models is limited. Thus we need to coarsely parameterize model space at the risk of introducing biases. This is similar to the modeling of allele frequency distributions: even though the vast majority of scenarios are inconsistent with the data, the number of models that are consistent with the data remains large, and model fitting often requires simplifying assumptions. When applied to HapMap trio-phased African-American data (ASW), inferred parameters were reasonable and we found evidence for
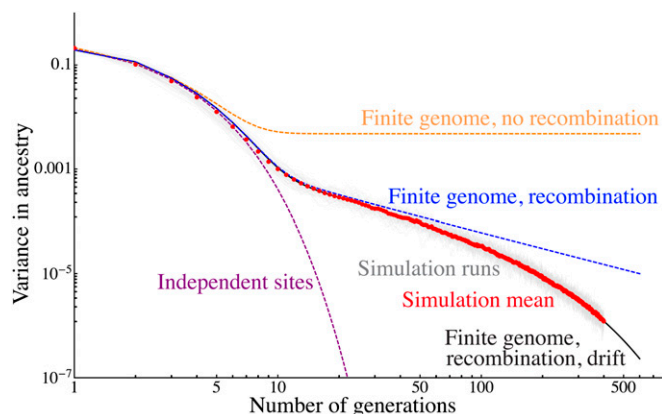
Figure 6 Comparison of 50 independent Wright–Fisher simulations of a population of 80 samples and 30% admixture proportion to predictions from increasingly detailed models. We show the variance in ancestry across individuals for each simulation in pale gray, and the average over the simulations is shown as red dots. These are compared to predictions for an independent sites model (purple) for a finite genome with 22 nonrecombining chromosomes (orange), for a model with recombination (blue), and finally for a model with recombination and drift given by Equation 8 (black). The latter model captures the variance in quantitative detail over three qualitative regimes.

migration patterns that depart from the migration pulse with subsequent random mating, which is at the heart of many approaches. However, distinguishing between continuous migration and nonrandom mating remains challenging.

The demographic inference strategy we presented requires accurate local ancestry assignments. Since longer tracts contain more ancestry information, we expect the most significant types of misassignment to be short, spurious ancestry tracts and the failure to identify real, short ancestry tracts. In the HapMap ASW data used here, the source populations are diverged enough that assignments are relatively reliable down to relatively short tracts, and indeed we find that the number of tracts predicted by the model is in good agreement with the data for the shortest tracts, even though these were not used in the fitting procedure. If the expected number of misidentified short tracts is large enough that it will strongly affect the distribution of longer tracts (by introducing spurious breaks in longer tracts), the Markov models should be modified to include misidentification states, and transition rates could be estimated via simulated admixed individuals.

Alternatively, it is possible to circumvent the local ancestry inference step altogether by focusing on a derived statistic, such as the decay of correlation in ancestry informative markers with genetic distance. Such a method was proposed in Reich *et al.* (2009), for the case of pairwise ancestry correlations in a pulse migration model. Even though such an approach avoids possible biases due to local ancestry assignment, pairwise ancestry correlations become noisy as distance is increased and are thus less sensitive to continuous gene flow patterns. The Markov models presented here provide a natural framework to generalize linkage-based models for more general admixture scenarios, as

arbitrary order linkage statistics can be derived in the Markov framework. Furthermore, HMM approaches could be developed to model the complete mosaic pattern without the need to focus on summary statistics such as the tract-length distributions and ancestry variances. Even though such approaches would be more computationally intensive, they may increase the accuracy of the inference, especially when assignment errors are important.

Conclusion

Overall, we found that the proposed models accurately describe the distribution of ancestry tract lengths and variances when compared to Wright–Fisher simulations. The models we used allow for general migration histories, yet are tractable and can be used for inferring demographic parameters in real data. They are therefore useful to improve our understanding of the consequences of gene flow and our ability to infer demography in populations with complex histories. Such populations have often been underrepresented in medical genetic studies, in part because of complications in the modeling of genetic heterogeneity. As medical genetics sampling efforts strive to reduce this disparity, detailed models for genetic diversity will be increasingly important to make the most out of the resulting data.

Literature Cited

Bercovici, S., and D. Geiger, 2009 Inferring ancestries efficiently in admixed populations with linkage disequilibrium. J. Comput. Biol. 16(8): 1141–1150.

Bhatia, G., N. Patterson, B. Pasaniuc, N. Zaitlen, G. Genovese *et al.*, 2011 Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am. J. Hum. Genet. 89(3): 368–381.

Brisbin, A., 2010 Linkage analysis for categorical traits and ancestry assignment in admixed individuals. Ph.D. Thesis, Cornell University, Ithaca, NY.

Ewens, W. J., and R. S. Spielman, 1995 The transmission/disequilibrium test: history, subdivision, and admixture. Am. J. Hum. Genet. 57(2): 455–464.

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.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108 (29): 11983–11988.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3(4): 479–502.

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.

Hoggart, C. J., M. D. Shriver, R. A. Kittles, D. G. Clayton, and P. M. McKeigue, 2004 Design and analysis of admixture mapping studies. Am. J. Hum. Genet. 74(5): 965–978.

International HapMap 3 Consortium, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.

Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165: 2213–2233.

Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73(3): 342–348.

Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler *et al.*, 2004 Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. 74(5): 979–1000.

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., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature 461 (7263): 489–494.

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin, 2008 Estimating local ancestry in admixed populations. Am. J. Hum. Genet. 82(2): 290–303.

Seldin, M. F., B. Pasaniuc, and A. L. Price, 2011 New approaches to disease mapping in admixed populations. Nat. Rev. Genet. 12 (8): 523–528.

Stewart, W., 1994 *Introduction to the Numerical Solution of Markov Chains*, Vol. 41. Princeton University Press, Princeton, NJ.

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.

Ungerer, M. C., S. J. Baird, J. Pan, and L. H. Rieseberg, 1998 Rapid hybrid speciation in wild sunflowers. Proc. Natl. Acad. Sci. USA 95(20): 11757–11762.

Verdu, P., and N. A. Rosenberg, 2011 A general mechanistic model for admixture histories of hybrid populations. Genetics 189: 1413–1426.

Wegmann, D., D. E. Kessner, K. R. Veeramah, R. A. Mathias, D. L. Nicolae *et al.*, 2011 Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. 43(9): 847–853.

Appendix 1: The Effect of Drift on Ancestry Transitions

Drift increases the probability that recombinations occur between segments of the same ancestry. In the infinite-time limit, ancestry will be fixed at every site, no more ancestry switches are created, and the tract-length distribution is constant in time. In the presence of drift, the ancestry switches are no longer Markovian; if a recombination occurs between two IBD segments, it increases the posterior probability that the next recombination will also be between IBD segments. However, it is likely that a Markovian approximation will remain accurate for moderate drift if we take into account the reduced probability of ancestry-switching recombinations.

We first wish to obtain the fraction of recombinations that occur within segments (p, t), of ancestry p having migrated at generation t, as these recombinations do not induce ancestry switches and will be most affected by drift. In other words, we want to find the fraction of sites that are homozygous for the ancestry (p, t), and contrast this to the case with no drift. For this purpose, we consider all other ancestries as a single allele, and in the first step we compute the total homozygocity of nonmigrants in this system s generations ago: $f_{p,t}^s$. We write the usual recursive relation over generations, noting that a homozygous state in a Wright–Fisher model can be obtained in one of four parental situations: drawing the same nonmigrant parent twice, drawing two nonmigrant parents with the same ancestry, drawing one last-generation migrant and a nonmigrant with the same ancestry, and finally drawing two last-generation migrants:

$$f_{p,t}^s = \left(\frac{1}{2N(1-m(s+1))} + \left(1 - \frac{1}{2N(1-m(s+1))} \right) f_{p,t}^{s+1} \right)(1-m(s+1))^2 + m(s+1)(1-m(s+1))(1-\alpha_{p,t}(s+2)) + m^2(s+1).$$

(A1)

This recursion can be initiated with the homozygocity one generation after t, namely $f_{p,t}^{t-1} = m_p(t)^2 + (1-m_p(t)^2)$. Finally, to get the fraction $c_{p,t}$ of nonmigrant sites that are homozygous for the p, t ancestry at generation s, we write

$$2\alpha_{p,t}(s+1) = 2c_{p,t}(s) + 1 - f_{p,t}^s \qquad (A2)$$

and solve for $c_{p, t}$

$$c_{p,t}(s) = \frac{f_{p,t}^s - 1}{2} + \alpha_{p,t}(s+1), \qquad (A3)$$

which reduces to $\alpha_{p,t}^2(s+1)$ in the driftless limit.

In the driftless case, the probability of the state to the right of a recombination depended only on the time of the recombination. Due to the possibility of recombining within

segments identical-by-descent, this is no longer the case when drift is present. However, consider a given point x in state (p, t) along the genome. The distribution of the distance to the first recombination encountered upstream (or downstream) from x is unaffected by drift. Thus the relationship between transition rates Q and discrete transition probabilities R is maintained: $Q(t, p \rightarrow t', p') = (t - 1)R(t, p \rightarrow t', p')$ for $(t, p) \neq (t', p')$. If we indicate the state to the left or right of a recombination by a left- and right-pointing arrow, respectively, we write

$$
\begin{aligned}
R(t, p \rightarrow t', p') &\equiv P((t',p')_{\rightarrow} | (t,p)_{\leftarrow}) \\
&= \sum_{\tau=1}^{\min(t,t')-1} \frac{P((t',p')_{\rightarrow} | \tau, (t,p)_{\leftarrow})}{t-1} \\
&= \sum_{\tau=1}^{\min(t,t')-1} \frac{P((t,p)_{\leftarrow}, (t'p')_{\rightarrow} | \tau)}{(t-1)P((t,p)_{\leftarrow})}.
\end{aligned} \quad (A4)
$$

We can then write the rate matrix as

$$
Q(t, p \rightarrow t', p') = \sum_{\tau=1}^{\min(t,t')-1} \frac{c_{p,t,p',t'}(\tau)}{2\alpha_{p,t}(\tau+1)}, \quad (A5)
$$

where $c_{p,t,p't'}$ is the proportion of nonmigrant (diploid) sites with joint ancestry (p, t) and (p', t'), which can be obtained using a recursive equation, as in Equation A3. In the driftless case, this reduces to

$$
Q(t, p \rightarrow t', p') = \sum_{\tau=1}^{\min(t,t')-1} \alpha_{p',t'}(\tau+1),
$$

as obtained in Equation 2.

A case of particular interest is the pulse migration, with proportions m and $1 - m$ for populations 1 and 2, respectively. We then get $\alpha_1(\tau) = m$, and

$$
a_{p,t,p't'}(\tau) = 2m(1-m)\left(1 - \frac{1}{2N}\right)^{T-1-\tau}.
$$

We can therefore calculate the transition probabilities, which are still proportional to the migration rates, but now exhibit a more complex time dependence:

$$
Q(i \rightarrow j \neq i) = m_j(2N-1)\left(1 - \left(1 - \frac{1}{2N}\right)^{T-2}\right).
$$

The limit $N \rightarrow \infty$ yields the driftless case

$$
Q(i \rightarrow j \neq i) = m_j(T-2),
$$

and the limit $T \rightarrow \infty$ reveals a linear dependence of the transition rate on the population size:

$$
Q(i \rightarrow j \neq i) = m_j(2N-1).
$$

The infinite-time tract lengths are thus inversely proportional to the effective population size.

Appendix 2: Numerical Estimation of Tract Length Distribution

In this section we describe how to obtain the expected distribution of tract lengths, given a set of Markov transition rates. A straightforward numerical solution strategy is to uniformize the transition matrix (Stewart 1994). Uniformization uses the fact that self-transition probabilities can be adjusted without affecting the trajectory statistics and in such a way that the total transition rate from each state is equal to the rate of the state with the highest transition rate, Q_0. Once all states have the same outgoing rate Q_0, the problem can be decomposed in two steps: a discrete calculation of the number of transitions in a given excursion and a calculation of the trajectory lengths given the number of transitions.

In the first step, we establish the distribution $\{b_n\}_{n=1, \ldots, \infty}$ of the number of steps spent in tracts of a given ancestry p, which is a standard discrete Markov excursion problem. In principle, the number of steps can be arbitrarily large, but the probability of very long tracts decays rapidly, and after a certain number of steps the expected length of the excursion is more than the chromosome length. We therefore calculate $\{b_n\}_{n=1,\ldots,\Lambda}$ up to a cutoff Λ, such that $\sum_{i=1}^{\Lambda} b_i \simeq 1$ (we usually also choose Λ such that $\Lambda Q_0 > L$, the length of a chromosome). To ensure a proper probability distribution, we then set $b_{\Lambda+1} = 1 - \sum_{i=1}^{\Lambda} b_i$. There are many ways to obtain the b_n. For our purposes, we have found it convenient to evolve the state vector by repeated multiplication with a transition matrix modified to have a single, absorbing state corresponding to the non-p ancestries and recording the amount of absorbed probability per multiplication.

The second step is straightforward since the length of the trajectories with k steps follows the Erlang distribution

$$
\mathbb{E}_{k,Q_0}(x) = \frac{Q_0^k x^{k-1} e^{-Q_0 x}}{(k-1)!},
$$

leading to the following expression for the tract-length distribution:

$$
\phi(x) \simeq \sum_{k=1}^{\Lambda+1} b_k E_{k,Q_0}(x). \quad (A6)
$$

Ancestry tract-length distributions obtained in the infinite-chromosome limit may not be appropriate for finite genomes, particularly if many tracts have a length comparable to the chromosome length. For example, predicted tracts may be longer than the full chromosome length L, and these will not be observed. To model the tract-length distribution on a finite chromosome, we consider a general tract length distribution $\phi(x)$ on an infinite chromosome and ask for the distribution of tract lengths observed in a given window of length L. To this end, we first calculate the probability that the intersection of a tract of length x_0 and a window of length L has length x. The probability $P(I)$ that

a tract of length x_0 intersects the window of length L is proportional to $x_0 + L$. Given I, and assuming that $x_0 < L$, the probability that the intersection is of length x is

$$P(x|x_0 \leq L, I) = \frac{2}{x_0 + L}\Theta(x_0 - x) + \left(1 - \frac{2x_0}{x_0 + L}\right)\delta(x - x_0),$$

(A7)

with Θ the Heaviside function and δ Dirac's delta function.

The result for $x_0 > L$ can be obtained by the permutation $x_0 \leftrightarrow L$, so that

$$P(x|x_0 \geq L, I) = \frac{2}{x_0 + L}\Theta(\eta - x) + \left(1 - \frac{2\eta}{x_0 + L}\right)\delta(x - \eta)$$

(A8)

with $\eta = \min(x_0, L)$. This yields

$$P(x|x_0) \propto P(x|x_0, I) * (L + x_0).$$ (A9)

As a result, we can write the expected new tract distribution, ranging from 0 to L, as

$$\phi'(x) \propto 2 \int_x^\infty dx_0 \phi(x_0) + (L - x)\phi(x) \\ + \delta(L - x) \int_L^\infty (x_0 - L)\phi(x_0)dx_0.$$

(A10)

The first term corresponds to the tracts that contact the edges of the window, the second term describes tracts that are strictly included in the window, whereas the third term describes all tracts that span the full window. Note that the "edge" tracts therefore generally have a different length distribution compared to the "inner" tracts and that as L goes to infinity, the second term dominates and is proportional to $\phi(x)$.

The normalizing factor is

$$Z = L + \int_0^\infty x\phi(x)dx,$$ (A11)

if all tracts are taken into account, and

$$Z^{<L} = L\left(2 - \int_0^L \phi(x)dx\right) + \int_0^L x\phi(x)dx,$$ (A12)

if only tracts with length $x < L$ are considered. Finally, if only tracts of length $> C$ are considered, we have

$$Z_C = (L - 2C)\int_C^\infty \phi(x) + \int_C^\infty x\phi(x)$$

and

$$Z_C^{<L} = (L - 2C)\int_C^\infty \phi(x) + L\int_L^\infty \phi(x)dx + \int_C^L x\phi(x).$$

We now apply these results to the tract-length distributions from Equation A6. First, we note the Erlang distribution is related to the generalized incomplete gamma function by

$$\int_{x_1}^{x_2} dx' E_{k,T}(x') = \frac{\Gamma(Tx_1, Tx_2, k)}{(k-1)!},$$

$$\int_{x_1}^{x_2} dx' x' E_{k,T}(x') = \frac{\Gamma(Tx_1, Tx_2, 1 + k)}{T(k-1)!}.$$

(A13)

This way, using our series expansion (A6), everything can be calculated in terms of gamma functions. For example,

$$Z = L + \sum_{k=1}^{\Lambda+1} b_k k / T$$

and

$$Z_C = Z - 2C + (2C - L)\sum_{k=1}^{\Lambda+1} b_k \frac{\Gamma(0, TC, k)}{(k-1)!} - \sum_{k=1}^{\Lambda+1} b_k \frac{\Gamma(0, TC, k+1)}{T(k-1)!}.$$

We can thus write separately the probabilities of having inside, edge, or full tracts of various lengths:

$$\phi_i(x) = \frac{(L-x)}{Z}\sum_{k=1}^{\Lambda+1} b_k E_{k,T}(x)$$

$$\phi_e(x) = \frac{2}{Z}\sum_{k=1}^{\Lambda+1} b_k \frac{\Gamma(Tx, \infty, k)}{(k-1)!}$$

(A14)

$$\phi_f(x) = \frac{\delta(L-x)}{Z}$$

$$\times \sum_{k=1}^{\Lambda+1} b_k \frac{L\Gamma(TL, \infty, k) + \Gamma(TL, \infty, k+1)}{(k-1)!}.$$

Appendix 3: Ancestry Variance in the Absence of Drift

Ancestry variance under a Markov model of ancestry

We consider in this section the assortment variance, in the absence of drift, where ancestry in two individuals is modeled as independent realizations of a two-state Markov process. Let the Markov states representing ancestry be labeled by $k = 1, 2$ with rates q_1 and q_2 out of states 1 and 2, respectively. The generalization to multistate Markov processes is discussed below. We first consider a single chromosome of length L and are interested in the variance in X, the length of this chromosome covered in state $k = 1$. We have $X = \int_0^L dx \psi_k(x)$, with $\psi_k(x)$ the indicator function of state k at position x along the genome. Changing the order of the expectation and the integrals, we have

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_0^L \int_0^L dxdy \mathbb{E}[\psi_k(x)\psi_k(y)] \\ - \mathbb{E}[\psi_k(x)]\mathbb{E}[\psi_k(y)].$$

(A15)

All these expectations are independent of the position along the chromosome. We therefore have

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_0^L \int_0^L dxdy\, \alpha_k (P(y|x) - \alpha_k), \qquad \text{(A16)}$$

with $\alpha_k = q_{1-k}/(q_1 + q_2)$ and $P_k(y|x)$ is the probability that y is in ancestry k given that x is in ancestry k. In a Markov process, $P(y|x) = (1-\alpha_k)e^{-(q_2+q_1)|x-y|} + \alpha_k$. The integral yields

$$\frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{L^2} = \frac{2\alpha_k(1-\alpha_k)}{(q_1+q_2)L}\left(1 - \frac{1-e^{-(q_1+q_2)L}}{(q_1+q_2)L}\right).$$

In the absence of drift, $q_1 + q_2 = T - 2$, and in the limit $(q_1 + q_2)L \gg 0$, we recover our estimate

$$\frac{\mathbb{E}[X^2] - \mathbb{E}[X]^2}{L^2} \simeq \frac{2\alpha_p(1-\alpha_p)}{(T-2)L}.$$

Assortment variance for nonconstant migration

The generalization to arbitrary one-way migrations is straightforward, in the absence of drift. We evaluate Equation A16 by expanding on arrival times s for ancestry p:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_0^L \int_0^L dxdy \left(\sum_s P(y_p|x_{(p,s)})\alpha_{p,s} - \alpha_p\alpha_p\right). \qquad \text{(A17)}$$

The probability that y is in the ancestry p, given that x is in state (p, s), can be written as $P(y_p|x_{(p,s)}) = \sum_{i,\nu} a_i^s e^{\kappa_i r} v_{i\nu}\rho_{\nu p}$, where r is the distance between x and y in morgans, ν represents a Markov state (p', s'), $\rho_{\nu p}$ is the indicator that $p' = p$, and $(\kappa_i, v_{i\nu})$ are the eigenvalues and eigenvectors of the transition matrix $Q_{\nu\nu'}$. To obtain the a_i^s, we set $\sum_i a_i^s v_{i\nu} = \delta_{\nu,(p,s)}$. For computational efficiency, we can first perform the sum over s in Equation A17 and solve only once for $a_i = \sum_s a_i^s(\alpha_{p,s}/\alpha_p)$. Assuming that the Markov chain has a unique stationary distribution (which is the case if the number of generations is finite and last-generation migrants are not allowed), there is a unique i_0 with $\kappa_{i_0} = 0$. The corresponding term cancels out in (A17), so that we can finally write

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \alpha_p \sum_{i \neq i_0, s} a_i v_{is}\rho_{sp} \int_0^L \int_0^L dxdy\, e^{\kappa_i r},$$

$$= \alpha_p \sum_{i \neq i_0, s} a_i v_{is}\rho_{sp}\frac{\left(-1 + e^{L\kappa_i} - L\kappa_i\right)}{\kappa_i^2}. \qquad \text{(A18)}$$

Distinguishing the two components of the ancestry variance from interchromosomal variance

We argued that, due to random chromosome assortment, the variance in ancestry between chromosomes is informative of the assortment variance. If all chromosomes had the same length, we could expect that the assortment variance in ancestry proportion across individuals would be pro-
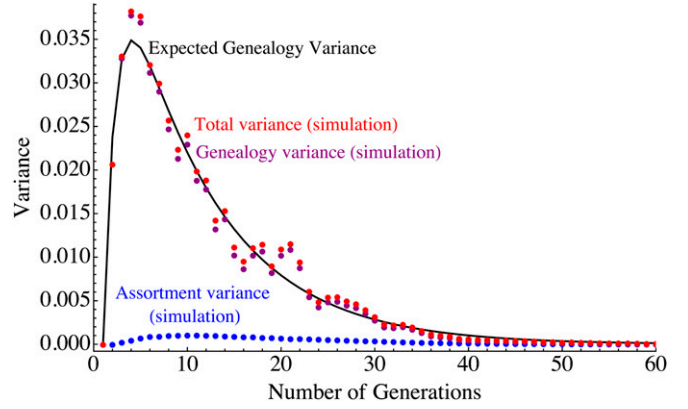

Figure A1 Time evolution of the variance for a population of 200 diploid individuals for a constant migration rate of 5% starting at generation 1. As the fraction of genetic ancestry originating from the migrant populations grows from 0 to 1, the variance reaches a maximum before the migration frequency reaches 0.5. Using the assumptions of Equation A19, we decompose the observed variance (red dots) in a genealogy (purple) and an assortment (blue) contribution. As expected, the genealogy contribution dominates.

portional to the variance across chromosomes and inversely proportional to the number of chromosomes per individual. However, the different chromosomes have different lengths, and to combine the information we need an idea of how ancestry variance depends on chromosome length. We assume that the assortment variance on ancestry is inversely proportional to the chromosome length in morgans; in effect, we suppose that the number of independent ancestry observations is proportional to the chromosome length. The proportionality factor σ_g depends on the pedigree, so that

$$\text{Var}(X^p|g) = \text{Var}\left(\sum_i L_i X_i^p|g\right)/L^2$$

$$= \sum_i L_i^2 \text{Var}(X_i^p|g)/L^2$$

$$\simeq \frac{\sigma_g^p}{L^2} \sum_i L_i \qquad \text{(A19)}$$

$$= \frac{\sigma_g^p}{L}.$$

Furthermore, since we are interested in the average variance over all pedigrees, we get

$$\mathbb{E}_g[\text{Var}(X^p|g)] = \frac{\mathbb{E}_g[\sigma_g^p]}{L}.$$

We therefore obtain an expression for $\mathbb{E}_g[\sigma_g^p]$ derived from the data. For each individual and each chromosome, we can obtain an estimate for this variance by comparing the ancestry proportion in that chromosome to the individual mean. We can then obtain the best-fitting σ_g^p. An average over all sequenced individuals provides us with an estimate for $E_g[\sigma_g^p]$. This procedure is used to decompose the simulated variances in Figure A1.