INFERRING ADMIXTURE PARAMETERS FROM ADMIXTURE FRACTIONS

MASON LIANG, MIKHAIL SHISHKIN, VLADIMIR SHCHUR AND RASMUS NIELSEN

Abstract. Estimation of admixture fractions has become one of the most commonly used computational tools in population genomics. However, there is remarkably little population genetic theory on their statistical properties. We develop theoretical results that can accurately predict means and variances of admixture proportions within a population using models with recombination and genetic drift. Based on established theory on measures of multilocus disequilibrium, we show that there is a set of recurrence relations that can be used to derive expectations for higher moments of the admixture fraction distribution. We obtain closed form solutions for some special cases. Using these results, we develop a method for estimating admixture parameters from estimated admixture proportion obtained from programs such as Structure or Admixture. We apply this method to HapMap data and find that the population history of African Americans, as expected, is not best explained by a single admixture event between people of European and African ancestry. A model of constant gene flow for the past 11 (3) generations until 2 (3) generations ago gives a better fit.

Introduction

It is common in population genetic analyses to consider individuals as belonging fractionally to two or more discrete source populations. The proportion of an individual's genome that belongs to a population is called that individual's 'admixture fraction' or 'admixture proportion'. Programs such as `Structure` (Pritchard et al., 2000), `Eigenstrat` (PCA?) (Price et al., 2006), `Frappe` (Tang et al., 2005), or `Admixture` (Alexander et al., 2009) can jointly estimate these admixture fractions for multiple individuals in a sample, along with the corresponding allele frequencies in each of the source populations. These admixture fractions are often presented in a 'structure plot,' an example of which is shown in Figure 5. We will henceforth refer

Date: May 19, 2023.

to these methods as 'structure analyses'. This approach has proven highly useful for understanding genetic relationships in many different species, e.g. humans (Rosenberg et al., 2002), cats (Menotti-Raymond et al., 2008), or pandas (Zhang et al., 2007). Other analyses reconstruct admixture tracts for each genome in the sample, by inferring the local ancestry of every position, or window, in each sampled genome (Tang et al., 2006; Maples et al., 2013). In this context, the admixture fraction for a genome is the fraction of its total length that is inherited from a particular source population.

Although structure analyses are not tied to any particular mechanistic model of population history and demography, the admixture fractions and admixture tracts are commonly interpreted to be the result of past admixture events in which modern populations were formed by admixture (or introgression) between ancestral source populations. The distribution of admixture tract lengths has been related to specific mechanistic models of admixture (Falush et al., 2003; Tang et al., 2006; Pool and Nielsen, 2009), and has been used to estimate times of admixture (Gravel, 2012). However, the admixture proportions themselves also contain information regarding admixture times. Following an admixture event, the variance in admixture proportions within a population will be high, but will thereafter decrease, and will eventually converge to zero in the limit of large genomes. The variance in admixture fractions among individuals contains substantial information about the time since admixture that can be used in addition to the tract length distribution. In some cases, this may be more robust than inferences based on tract lengths, because the length distribution of tracts is often difficult to infer, and is often not modeled accurately by the hidden Markov model (HMM) methods used to infer tract lengths (Liang and Nielsen, 2014). Even in cases where tract lengths can be accurately inferred, studies aimed at estimating admixture times should benefit from using both variance in admixture proportions among individuals and overall admixture tract lengths distributions.

Verdu and Rosenberg (2011) developed a method for computing moments of admixture proportions in a model in which admixed population is formed as a mixture between multiple source populations, allowing for arbitrary gene-flow from the source populations over a number of generations (g). They establish recursions for the moments of the admixture fractions and use these equations to determine how the mean and the variance changes

through time in particular admixture scenarios. These moments are expectations for *single* individual's admixture fraction and are averaged over the possibile genealogical histories of the population. As a result, they can be difficult to relate to data because replicates from multiple identical populations rarely are available. In this paper, we consider a different problem, the problem of calculating sample moments for admixture proportions obtained from individuals in one population.

We extend the model in Verdu and Rosenberg (2011) to incorporate the effects of recombination and genetic drift by adding a a random union of zygotes component. Recombination is important because even if one half of a chromosome's ancestors are from the first source population, it is unlikely that exactly one half of that chromosome's genetic material is inherited from that population. Genetic drift is important because the individuals in a sample might share ancestors and, therefore, have more similar admixture fractions than expected by chance in a model without drift. The results developed in this paper should be directly applicable for quantifying the results of a structure analysis.

The General Mechanistic Model

We start by considering admixture fractions in haploid genomes. These haploid admixture fractions can later be paired up to create diploid admixture fractions. The admixture fraction of a (haploid) genome H_i, is the proportion of H_i that is inherited from a particular source population. For notational simplicity, we only consider gene-flow only from one population into another. We will later discuss how to extend this model to multiple admixing source populations. We use the same mechanistic admixture model of Verdu and Rosenberg (2011), and will use its notation where possible. Finally, we use the random union of zygotes model, with a diploid population size of N ($2N$ chromosomes), for genetic drift and recombination, and assume a sample size of n chromosomes from a single population.

In this model, a hybrid population of N diploid individuals forms in generation 1 from two previously isolated source populations. In this first generation, individuals in the hybrid population are from the first source population with probability s_0 or from the second source population with probability $1 - s_0$. In generation $g + 1$, each chromosome is, independently, from the first source population with introgression probability s_g, or from the hybrid population with probability $1 - s_g$. Chromosomes inherited from

the hybrid population are the product of the recombination of the two chromosomes of one individual (zygote), chosen uniformly at random. Finally, these $2N$ chromosomes are paired up to form the N individuals in generation $g+1$.

Finally, we let the stochastic process $A(\ell)$ represent the local ancestry along a chromosome as a function of ℓ, the physical position:

$$A(\ell) = \begin{cases} 0 & : \ \ell \text{ is descended from first source population} \\ 1 & : \ \ell \text{ is descended from second source population} \end{cases}.$$

The fraction of the chromosome descended from the second source population is given by

$$H = \frac{1}{L}\int_0^L A(\ell)d\ell,$$

where L is the total length of the chromosome.

Assume that g generations after the start of admixture we have randomly sampled n chromosomes from the hybrid population and determined their corresponding admixture fractions, $H_{1(g)}, H_{2(g)}, \ldots, H_{n(g)}$. We are interested in the joint distribution of these n random variables. When $n = 1$ and as $L \to \infty$, this is the admixture fraction considered by Verdu and Rosenberg (2011).

Because the n chromosomes have possibly overlapping geneologies, the admixture fractions are not independent. However, the joint distribution of the admixture fractions does not depend on their ordering, so they are exchangeable. As a result, they can be viewed as being identically and independently (*iid*) drawn from a random distribution $\mathcal{G}$. This random distribution can be interpreted as a function of the random genealogy of the entire hybrid population up to g generations in the past. When g is small, the genealogies of the n samples will be unlikely to differ from n non-overlapping binary trees, so $\mathcal{G}$ will be approximately constant. If g is large however, these genealogies are likely to overlap, and this will no longer be true.

Verdu and Rosenberg (2011) focus on moments of $H_{1(g)}$, in particular on the mean and variance. However, because the admixture fractions are not independent, even as $n \to \infty$, the sample mean and sample variance will converge to the mean and variance of $\mathcal{G}$, which are random quantities. For example,

$$\mathbb{E}(H_{1(g)}) \neq \mathbb{E}(H_{1(g)}|\mathcal{G}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H_{i(g)}$$

$$\text{var}\left(H_{1(g)}\right) \neq \text{var}(H_{1(g)}|\mathcal{G}) = \lim_{n\to\infty} \frac{1}{n-1}\sum_{i=1}^{n}\left(H_{i(g)} - \frac{1}{n}\sum_{j=0}^{n} H_{j(g)}\right)^2,$$

and similarly for higher-order moments. The moments of the admixture fractions have two components: randomness from sampling the population genealogy, and randomness from the sampling of chromosomes. The expressions to the left account for both, while the expressions to the right only account for the latter. In general, for k-statistics we have

$$\mathbb{E}[k_m|\mathcal{G}] = \kappa_m(H_{i(g)}|\mathcal{G}),$$

where κ_m is the m-th cumulant of $H_{i(g)}$. As we noted,

$$\mathbb{E}[k_m] = \mathbb{E}[\mathbb{E}[k_m(R_i)|\mathcal{G}]] = \mathbb{E}[\kappa_m(H_{i(g)}|\mathcal{G})] \neq \kappa_m(H_{i(g)}),$$

particularly,

$$\mathbb{E}[k_2] = \mathbb{E}[\mathbb{E}[k_2|\mathcal{G}]] = \mathbb{E}[var(H_{i(g)}|\mathcal{G})] = var(H_{i(g)}) - var(\mathbb{E}[H_{i(g)}|\mathcal{G}]).$$

Variances among individuals within one population correspond to $\text{var}(H_{1(g)}|\mathcal{G})$, while variances over replicate populations correspond to $\text{var}(H_{1(g)})$. This latter value will be larger than the expected sample variance calculated from multiple individuals sampled from the same population, and will rarely be useful for inference purposes.

In the following sections, we will show how expectations of k-statistics, can be derived for mechanistic models of introgression. By comparing these expectations to the observed admixture parameters from a sample, we will be able to construct a method of moments estimator for the parameters of the model.

Let k_1 be the sample mean:

$$k_1 \equiv \frac{1}{n}\sum_{i=1}^{n} H_{i(g)}.$$

We can express its expectation in terms of the 1-point correlation function of A:

$$\begin{aligned}\mathbb{E}(k_1) &= \mathbb{E}(H_{1(g)}) \\ &= \frac{1}{L}\int_0^L \mathbb{P}\{A_{1(g)}(\ell) = 1\} d\ell \\ &= \mathbb{P}\{A_{1(g)}(0) = 1\}.\end{aligned}$$

Similarly, let k_2 be the unbiased estimator of the sample variance:

$$k_2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(H_{i(g)} - k_1\right)^2.$$

Its expectation is given by

$$\begin{aligned}\mathbb{E}(k_2) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(H_{i,g}^2) - \frac{1}{n(n-1)}\sum_{i\neq j}\mathbb{E}(H_{i,g}H_{j,g}) \\ &= \mathbb{E}(H_{1,g}^2) - \mathbb{E}(H_{1,g}H_{2,g}).\end{aligned}$$

These expectations can be written in terms of two-point correlation functions of A:

$$\begin{aligned}\mathbb{E}(H_{1(g)}^2) &= \frac{1}{L^2}\mathbb{E}\left(\int_0^L A_{1(g)}(\ell)d\ell \int_0^L A_{1(g)}(\ell)d\ell\right) \\ &= \frac{1}{L^2}\int_0^L\int_0^L \mathbb{E}\left(A_{1(g)}(\ell)A_{1(g)}(\ell')\right) d\ell d\ell' \\ &= \frac{1}{L^2}\int_0^L\int_0^L \mathbb{P}\left\{A_{1(g)}(\ell) = 1, A_{1(g)}(\ell') = 1\right\} d\ell d\ell'.\end{aligned}$$

Similarly,

$$\mathbb{E}(H_{1(g)}H_{2(g)}) = \frac{1}{L^2}\int_0^L\int_0^L \mathbb{P}\left\{A_{1(g)}(\ell) = 1, A_{2(g)}(\ell') = 1\right\} d\ell d\ell'.$$

Writing these two correlation functions as

$$\mathbf{v}_{2(g)} = \begin{pmatrix} \mathbb{P}\left\{A_{1(g)}(\ell) = 1, A_{1(g)}(\ell') = 1\right\} \\ \mathbb{P}\left\{A_{1(g)}(\ell) = 1, A_{2(g)}(\ell') = 1\right\} \end{pmatrix},$$

we find that

$$\mathbb{E}(k_2) = \frac{1}{L^2}\int_0^L \int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{v}_{2(g)} d\ell d\ell'. \tag{1}$$

In general, the i^{th} k-statistic is an unbiased estimator of the i^{th} cumulant of $\mathcal{G}$, and its expectation can be written as an integral over $[0, L]^i$ of a linear combinations of i-point correlation functions. For example,

$$\mathbb{E}(k_3) = \frac{1}{L^3}\int_0^L \int_0^L \int_0^L \begin{pmatrix} 1 & -1 & -1 & -1 & 2 \end{pmatrix} \mathbf{v}_{3(g)} d\ell d\ell' d\ell''$$

$$\mathbb{E}(k_4) = \frac{1}{L^4}\int_{[0,L]^4} \Big(1 \underbrace{-1}_{4 \text{ times}} \underbrace{-1}_{3 \text{ times}} \underbrace{2}_{6 \text{ times}} 6\Big) \mathbf{v}_{4(g)} d\ell d\ell' d\ell'' d\ell'''$$

$\dots$

Remarkably, the linear combinations required to compute the expectations of the k-statistics correspond exactly to the higher-order disequilibria as defined by Bennett (1952). Furthermore, if instead the we choose to compute the expectations of the h-statistics, which estimate the central moments, the linear combinations would correspond to the higher-order disequilibria as defined by Slatkin (1972).

We next find the recurrence relations these correlation functions satisfy and solve them in the some special cases. In particular we will consider the case of a single admixture event g generations ago and the case of constant gene-flow starting g generations ago.

A Single Admixture Event. We start with a simple case, where introgression only occurs in the founding generation, i.e. $s_g = 0$ for $g > 0$. Using the random union of zygotes model, we can compute $\mathbf{v}_{2(g)}$ in terms of the probabilities from the previous generation:

If two sites at genetic distances ℓ and ℓ' are on the same chromosome in generation $g + 1$, then they were inherited from one chromosome from generation g with probability $[\ell\ell']$ and from two chromosomes from generation g with probability $[\ell|\ell']$. If they are on different chromosomes, then the probability that they are descended from one chromosome in generation g is $\frac{1}{2N}[\ell\ell']$ and the probability that they are descended from two chromosomes is $\frac{1}{2N}[\ell|\ell'] + \left(1 - \frac{1}{2N}\right)$ In matrix notation,

$$\mathbf{v}_{2(g+1)} = (\mathbf{L}_2\mathbf{U}_2)\,\mathbf{v}_{2(g)} = (\mathbf{L}_2\mathbf{U}_2)^g\,\mathbf{v}_{2(0)},$$

where the the recombination and drift matrices are given by

$$\mathbf{L}_2 = \begin{pmatrix} 1 & 0 \\ \frac{1}{2N} & 1 - \frac{1}{2N} \end{pmatrix}$$
$$\mathbf{U}_2 = \begin{pmatrix} [\ell\ell'] & [\ell|\ell'] \\ 0 & 1 \end{pmatrix}.$$

This is the the same matrix equation (Wright 1933 and Hill and Robertson 1966) derived for the decay of two-locus linkage disequilibrium. The 'alleles' we consider are the local ancestry at ℓ and ℓ'. To the extent possible, our notation will follow (Hill 1974), whose results for measures of multi-locus linkage disequilibria we use. The matrices $\mathbf{L}_2$ and $\mathbf{U}_2$ share $(1 \ -1)$ as a left-eigenvector, with corresponding eigenvalues $1-\frac{1}{2N}$ and $[\ell\ell']$. As a result,

$$\begin{aligned} \mathbb{E}(k_2) &= \frac{1}{L^2}\int_0^L\int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix} \cdot (\mathbf{L}_2\mathbf{U}_2)^g \, \mathbf{v}_{2(0)} d\ell d\ell' \\ &= \frac{1}{L^2}\left(1-\frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \int_0^L\int_0^L [\ell\ell']^g d\ell d\ell'. \end{aligned} \tag{2}$$

For a model using the Haldane map function, $[\ell|\ell'] = \frac{1-\exp(-2|\ell-\ell'|)}{2}$, this equation becomes

$$\begin{aligned} \mathbb{E}(k_2) &= \frac{1}{L^2}\left(1-\frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \int_0^L\int_0^L \left(\frac{1+\exp(-2|\ell-\ell'|)}{2}\right)^g d\ell d\ell' \\ &= \frac{2}{L^2}\left(1-\frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \int_0^L (L-\ell) \left(\frac{1+\exp(-2\ell)}{2}\right)^g d\ell d\ell', \end{aligned}$$

while for a model of complete crossover interference on a chromosome of length 1 Morgan, we can get a closed form solution:

$$\begin{aligned} \mathbb{E}(k_2) &= \left(1-\frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \int_0^1\int_0^1 \left(1-|\ell-\ell'|\right)^g d\ell d\ell' \\ &= \left(1-\frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \frac{2}{2+g}. \end{aligned}$$

For predicting the expected sample variance, the difference between these two models is not large, as shown in figure 6. For the simulations and inference in this paper, we will ignore crossover interference, and use the

Haldane map function. However, none of the mathematical results of this paper will require this assumption of no crossover interference.

For computing higher-order correlation functions, we find a similar equation

$$\mathbf{v}_{i(g)} = (\mathbf{L}_i \mathbf{U}_i)^g \, \mathbf{v}_{i(0)}. \tag{3}$$

Bennett's coefficients for higher-order linkage are left-eigenvectors of the recombination matrix $\mathbf{U}_i$. For $i = 3$, it is also a left-eigenvector of the drift matrix, so we immediately get that

$$\mathbb{E}(k_3) = \frac{s_0(1-s_0)(2-s_0)}{L^3} \left(1 - \frac{1}{2N}\right)^T \left(1 - \frac{2}{2N}\right)^T \int_{[0,L]^3} [\ell\ell'\ell'']^G d\ell d\ell' d\ell''.$$

For $i \geq 4$, this is no longer true, but the results of (Hill, 1974) can be used to compute $\mathbf{v}_i(g)$ without having to exponentiate the entire drift and recombination matrices. For example, for k_4, the drift and recombination matrices are 15×15, but using the technique in (Hill, 1974), we only need to exponentiate a 4×4 matrix to compute $\mathbb{E}(k_4)$.

Varying Migration. If $s_g > 0$ for $s \geq 1$, we obtain a modified version of Equation 3:

$$\mathbf{v}_{i(g)} = \mathbf{L}_i \mathbf{D}_{i(g)} \mathbf{U}_i \mathbf{v}_{i(g-1)}, \tag{4}$$

where the diagonal matrix $\mathbf{D}_{i(g)}$ has entries giving the probabilities the set of chromosomes, p, in a correlation function are all from the hybrid population in the previous generation:

$$d_{p,p(g)} = (1 - s_g)^{|p|}.$$

Note that if $s_{(g)}$ is fixed, then equation (4) is linear, and can be solved using a Laplace transform. Particularly, if the migration only lasted for g generations until present with a fixed migration matrix (a constant migration scenario), and

$$A_2 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \mathbf{L}_2 \mathbf{U}_2 \mathbf{D}_{2(g)},$$

then the expressions for the $\mathbf{v}_2^g$ components are

$$\mathbb{P}\left\{A_{1(g)}(\ell)=1, A_{1(g)}(\ell')=1\right\} = \frac{M(x_1^g+x_2^g)(1-s)+(x_1^g-x_2^g)(1-s)(d-a-2b(1-s))}{2M}$$

$$\mathbb{P}\left\{A_{1(g)}(\ell)=1, A_{2(g)}(\ell')=1\right\} = \frac{M(x_1^g+x_2^g)(1-s)^2+(x_1^g-x_2^g)(1-s)((a-d)(1-s)-2c)}{2M},$$

where $M^2=(a-d)^2+4bc$, and x_1, x_2 - the solutions of

$$\begin{vmatrix} a-x & b \\ c & d-x \end{vmatrix} = 0.$$

Hence, for the second moment we derive

$$\mathbb{E}(k_2) = \frac{1}{L^2}\int_0^L\int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix} \mathbf{v}_2^t dxdy$$

$$=\frac{1}{L^2}\int_0^L\int_0^L \frac{M(x_1^g+x_2^g)s+(x_1^g-x_2^g)(1-s)(2(d+c-a-b)+s(2b+a-d))}{2M} d\ell d\ell'.$$

Inference of admixture times. The equations in the previous section can be used to develop a method of moments-estimators for admixture parameters by numerically solving the admixture parameters in terms of the expectations for the k-statistics. Substituting in the observed values for the k-statistics gives estimates for the admixture parameter(s).

However, with real data, we only have estimates of the admixture fractions, so some of the variability seen in the distribution of admixture fractions will be due to estimation variability. To account for this, we assume that the estimations errors are additive and *iid*:

$$\hat{H}_{i(g)} = H_{i(g)} + \epsilon_i.$$

Because cumulants are additive,

$$\begin{aligned}\mathbb{E}(k_n) &= \mathbb{E}\left(\kappa_n(H_{i(g)}+\epsilon_i|\mathcal{G})\right) \\ &= \mathbb{E}\left(\kappa_n(H_{i(g)}|\mathcal{G})\right) + \kappa_n(\epsilon_i).\end{aligned}$$

The expectations we have computed are just the term of this sum. To correct for the variability in the estimates, we need to subtract off the second term. We use a block bootstrap to estimate these effects.

One additional complication arises in dealing with genotyping data. We have assumed that we have the ancestry fractions for each haplotype in the sample, but with genotyping data, we instead have their pairwise means:

$(H_{1(g)} + H_{2(g)})/2 \ldots$. This is results in a decrease in the expectations of the k-statitics. Conditional on the random distribution $\mathcal{G}$, $H_{1(g)}, H_{2(g)}, \ldots$ are *iid* drawn from $\mathcal{G}$. Cumulants are additive, so we find that

$$\begin{aligned}\kappa_i\left(\frac{H_{1(g)} + H_{2(g)}}{2}\middle|\mathcal{G}\right) &= \kappa_i\left(\frac{H_{1(g)}}{2}\middle|\mathcal{G}\right) + \kappa_i\left(\frac{H_{2(g)}}{2}\middle|\mathcal{G}\right) \\ &= 2^{-i+1}\kappa_i\left(H_{1(g)}|\mathcal{G}\right).\end{aligned}$$

Verification and comparison

Comparison to Verdu and Rosenberg. The recursion equations given by Verdu and Rosenberg (2011) are different from the ones we have derived. This is partly because we have accounted for the effects of genetic drift and recombination, but also because we are computing the moments of slightly different quantities.

In figure 2, we have shown the admixture fractions for five replicate populations 5, 50, and 500 generations after an admixture pulse. The variance that (Verdu and Rosenberg, 2011) compute variance over all the replicate populations, while the variance we have computed in this paper is the expectation of the variance within a single population. When g is small, these are similar, but when g is large, the variance within a population goes to zero, but the variance across the replicate populations does not. This effect is shown in Figure 1. Initially, both quantities decline exponentially in g, but after $2^g > nLg$, the variance we predict begins to decline linearly instead. This is because variance is inversely proportional to the number of genetic ancestors of the sample. When g is small, the number of genetic ancestors is approximately 2^g. However, the approximate number of recombination events in the sample is approximately bounded by nLg, so when this quantity is smaller than 2^g, it provides a better approximation for the number of genetic ancestors. In this regime, the variance will decline linearly in g.

It is also possible to compute the variance over all population replicates under our model, which allows a direct comparison to Verdu and Rosenberg (2011). In the case of one pulse of admixture, we can now solve equations 1 for $\mathbb{P}\{A_{1,g}(\ell) = 1, A_{1,g}(\ell') = 1\}$ to get

$$\begin{aligned}
\operatorname{var}(H_{1(g)}) &= \mathbb{E}(H_{1,g}^2) - s_0^2 \\
&= \frac{1}{L^2}\int_0^L \int_0^L \mathbb{P}\left\{A_{1,g}(\ell) = 1, A_{1,g}(\ell') = 1\right\} d\ell d\ell' - s_0^2 \\
&= \frac{1}{L^2}\left(s_0 - s_0^2\right)\int_0^L \int_0^L 1 - \left(1 - [\ell\ell']\right)\frac{1 - [\ell\ell']^g\left(1 - \frac{1}{2N}\right)^g}{1 - [\ell\ell']\left(1 - \frac{1}{2N}\right)} d\ell d\ell. \qquad (5)
\end{aligned}$$

This variance and the expectation of the second k-statistic have the same limit as $N \to \infty$, but for finite N, the variance is larger. This is because

$$\operatorname{var}(H_{1(g)}) = \operatorname{var}\left[\mathbb{E}(H_{1(g)}|\mathcal{G})\right] + \mathbb{E}\left[\operatorname{var}(H_{1(g)}|\mathcal{G})\right] = \operatorname{var}[k_1] + \mathbb{E}[k_2].$$

The first variance is small when N is large, but is always non-negative. The difference between this equation and equation 1 only becomes significant on a coalescent time scale. In the absence of genetic drift, the admixture fractions are approximately independent, because the samples do not share ancestors.

Genealogy influence. We noted that admixture proportions are not independent. A random genealogy determines their distribution, so different population replicates with the same admixture parameters produce different admixture proportion distributions. This effect is shown in Figure 2. The figure also demonstrates that the variance of admixture proportions consists of two terms: the first corresponds to variability within a population, and the second corresponds to variability across replicate populations. Admixture proportions within a replicate population are correlated, which implies that variability must be maintained across populations for any generation. However, for more ancient admixture events, the variability within a replicate population decreases to 0. The expectation of the second k-statistic of admixture proportions converges to the first term, and is thus informative about the timing of migration.

Accuracy of parameter estimates. We evaluated the accuracy of the AP-method using simulations (see Figure 3). The method demonstrated good performance in estimating recent admixture events. However, estimations of older migrations are less precise due to k-statistics given by higher values of T and T_d becoming very similar to each other (see Figure 4). To

obtain more precise parameter estimates in these cases, more accurate estimations of data moments are required, which can be achieved with a larger sample size.

Application to African American Data

We applied this method to a subset of the ASW, CEU, and YRI data from the HapMap 3 project (Consortium et al., 2010). After excluding children from trios, there were the genotypes for 49 ASW, 113 YRI, and 112 CEU individuals. We estimated the admixture fractions using the supervised learning mode of `Admixture`, with the CEU and YRI individuals assigned to separate clusters. The sampling distribution of the admixture fractions was estimated using the block bootstrap with 10^4 replicates and 2678 blocks, giving a block size of approximately 10 CM. The admixture fractions for the 49 ASW samples are shown in Figure 5 and the observed k-statistics are given in table 8.

We used a 3-parameter model of constant admixture in which the admixed population is entirely of African ancestry before generation g_start. For $g_{start} \leq g \leq g_{stop}$, the gene-flow probability from the European source population is $s_g = s$, with $s_g = 0$ otherwise. By matching the block-bootstrap corrected k_2 and k_3 to the predictions of equation 1 using a total recombination length of $R = 33$ Morgans, we obtained point estimates of

$$\hat{s} = 0.0277$$
$$\hat{g}_{start} = 2$$
$$\hat{g}_{stop} = 11.$$

We obtained confidence intervals, shown in Figure 7, by simulation. For each cell in the grid, we simulated 10^3 replicates under the corresponding g_{start} and g_{stop}, with $s = 1 - k_1^{1/(g_{stop} - g_{start} + 1)}$. For each replicate, we computed the k_2, k_3, and k_4 statistics. A cell was then included in the confidence interval if and only if the corrected k_2, k_3, and k_4 statistics from the HapMap data fall inside a centered interval containing 98.7% of the probability mass of the simulated distribution. This mass was chosen so that under the Bonferroni correction for three tests, there is at least a 95% chance of including the true parameter values in the confidence region.

The point estimates for g_{start} and g_{stop} correspond to the values for which the observed k-statistics are closest to their simulated medians.

Discussion

We have extended the mechanistic model of Verdu and Rosenberg (2011) to account for recombination and genetic drift. Doing so allows us to apply the predictions of this model to data. This mechanistic model allows for a large number of parameters. For the purposes of inference, it seems that imposing constraints, i.e. a small number of pulses or constant admixture, will be needed to narrow the search space.

In this paper, we have assumed that admixture only comes from one source population, this need not be the case. To account for admixture from multiple source populations, equation 1 must be modified to account for the probability that haplotypes trace their descent to multiple source populations. Algorithmically, this is feasible, but the notation is cumbersome. The resulting equations are given in the appendix, along with the equations for computing expectations of higher-order k-statistics.

Applications of the method to African-American HapMap data provides estimates of the time since admixture between people of European and African descent in America. Notice that the confidence set for the admixture parameters does not include values of $g_{stop} = 0$. We interpret this as evidence that admixture rates have declined the last few generations. The point estimate of time gene-flow stopped is $g_{stop} = 2$. This abrupt stop in gene-flow is a limitation of our model. A more realistic interpretation of this estimate would be that the rate of gene-flow has declined within the last 5 generations or so. Another possible explanation for this estimate is that individuals with more recent admixture tend to not self-identify, or were otherwise not included in the ASW population. The effects of this sampling bias would be the same as the decline in gene flow we observed. Also notice that admixture before 15 generations ago can be rejected. With a generation time of 25-30 years, this corresponds to 325-400 years, and is in good accordance with the historical record. The point estimate of the time of first admixture is 11 generations, or approx. 275-330 years ago.

Structure analyses have become one of the most commonly applied tools in population genomic analyses. The theory developed in this paper allows users of structure analyses to interpret their data in the context of a model of admixture between populations, and should find use in many studies aimed at understanding the history of populations.

References

David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9): 1655–1664, 2009.

John Bennett. On the theory of random mating. *Annals of Eugenics*, 17(1): 311–317, 1952.

International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191 (2):607–619, 2012.

William G Hill. Disequilibrium among several linked neutral genes in finite population i. mean changes in disequilibrium. *Theoretical Population Biology*, 5(3):366–392, 1974.

Mason Liang and Rasmus Nielsen. The lengths of admixture tracts. *Genetics*, pages genetics–114, 2014.

Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante. Rfmix: A discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, 93(2): 278–288, 2013.

Marilyn Menotti-Raymond, Victor A David, Solveig M Pflueger, Kerstin Lindblad-Toh, Claire M Wade, Stephen J O'Brien, and Warren E Johnson. Patterns of molecular genetic variation among cat breeds. *Genomics*, 91 (1):1–11, 2008.

John E Pool and Rasmus Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945–959, 2000.

Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.

Montgomery Slatkin. On treating the chromosome as the unit of selection. *Genetics*, 72(1):157–168, 1972.

Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, 28(4):289–301, 2005.

Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006.

Paul Verdu and Noah A Rosenberg. A general mechanistic model for admixture histories of hybrid populations. *Genetics*, 189(4):1413–1426, 2011.

Baowei Zhang, Ming Li, Zejun Zhang, Benoît Goossens, Lifeng Zhu, Shanning Zhang, Jinchu Hu, Michael W Bruford, and Fuwen Wei. Genetic viability and population history of the giant panda, putting an end to the "evolutionary dead end"? *Molecular biology and evolution*, 24(8): 1801–1810, 2007.

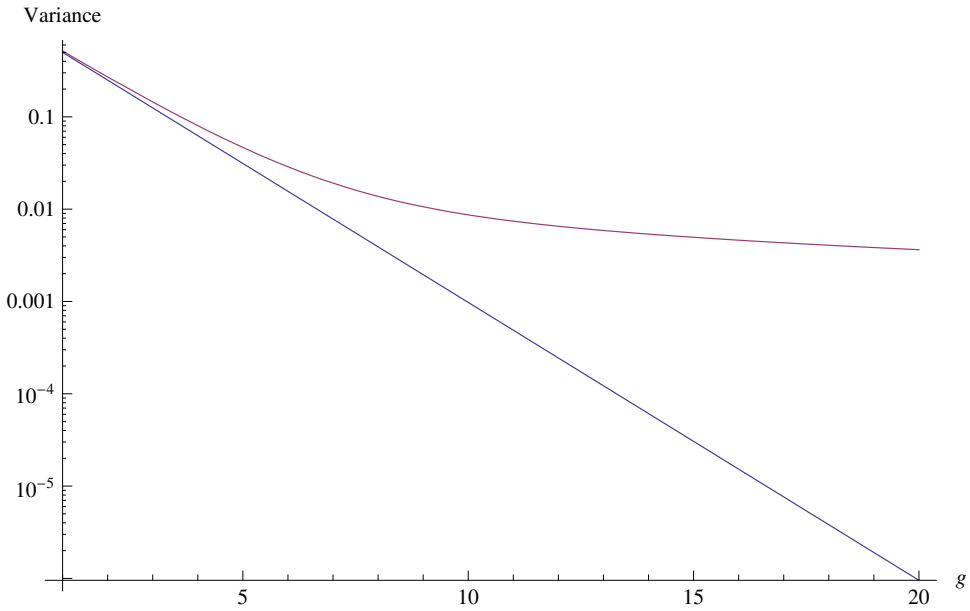

FIGURE 1. The variance predicted by Verdu and Rosenberg (2011) and equation 5, plotted on a logarithmic scale. The variance we predict (red) is always larger, but the two are very similar when g is small.

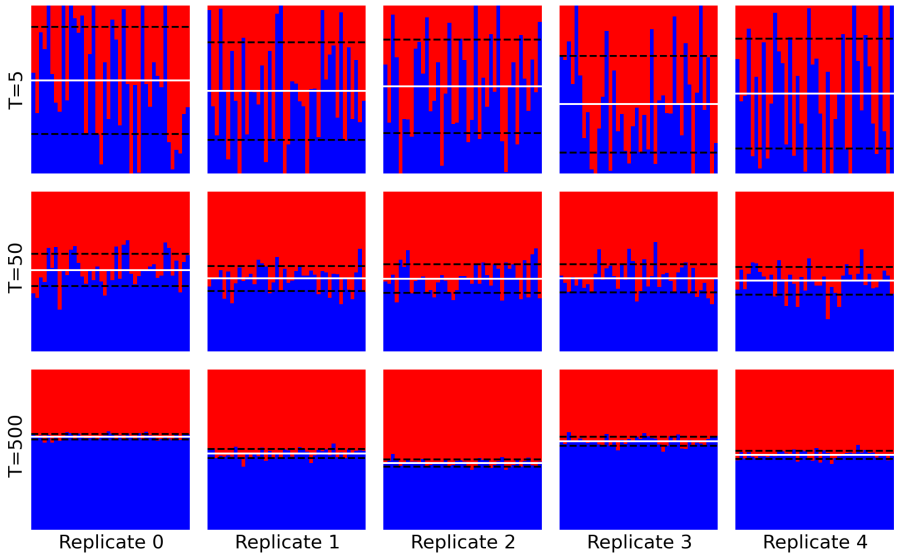

FIGURE 2. The admixture fractions of five replicate populations (each row) 5, 50, and 500 generations after an admixture pulse. Black dashed lines depict a standard deviation, white solid lines represent a sampled mean. For more ancient admixture events, the variability within a replicate population decreases to 0, but some variability is still maintained across the populations.

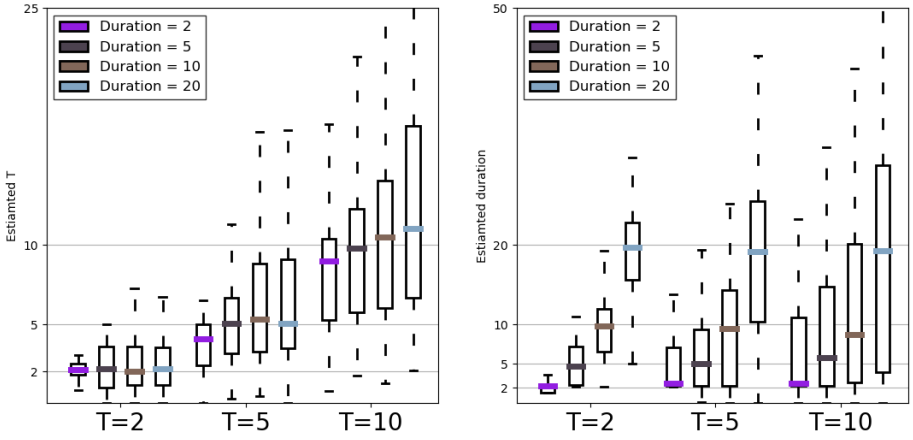

FIGURE 3. **Accuracy of estimates of time $g_{\mathbf{start}}$ to the migration ended and duration of migration T_d as a function of other parameters.** Twelve admixture scenarios, $T \in \{2, 5, 10\}$ and $T_d \in \{2, 5, 10, 20\}$, were simulated 100 times each. The total admixture proportion was fixed at $s = 0.36$. The error in the estimates increases with the increasing of time to migration or migration duration.

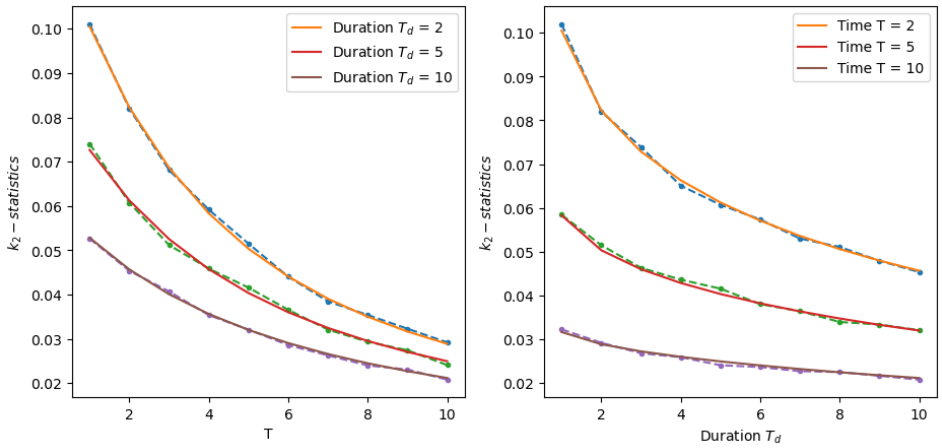

FIGURE 4. **k_2-statistics given by a constant migration as function of T and T_d.** Theoretical values of expected k_2 depicted as solid lines, dashed lines represent values estimated from simulations. For recent migrations the difference in values of k_2 is considerable, however it decreases for older admixtures, making them hardly distinguishable.

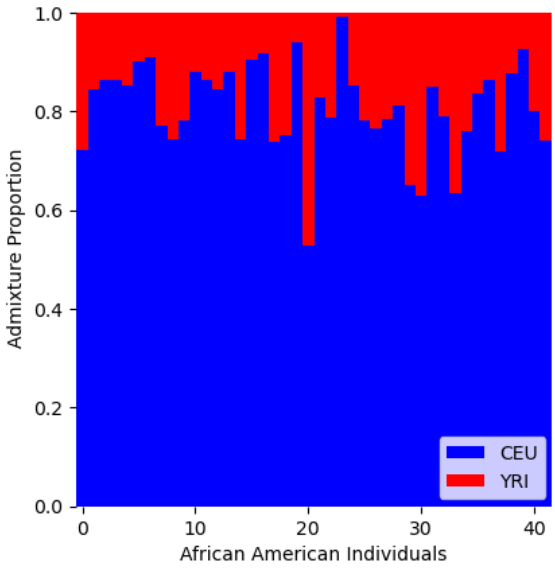

FIGURE 5. Admixture fractions for 49 African American individuals in the HapMap 3 data. Source population allele frequencies were estimated using 113 Yoruban and 111 European individuals.

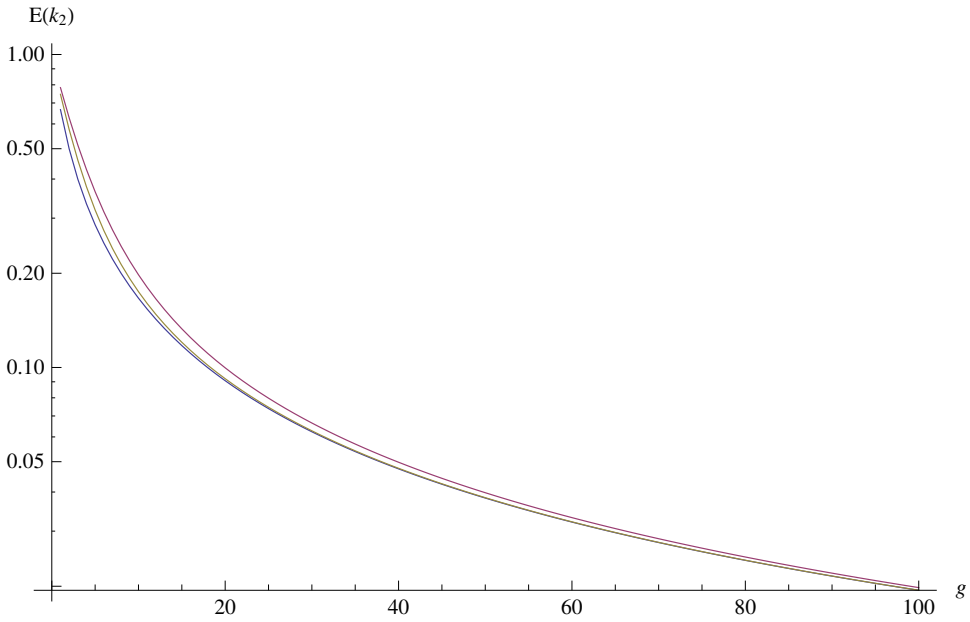

FIGURE 6. The expected sample variance given by equation 1 plotted on a logarithmic scale, for a three different map functions. We used a map distance of $L = 1$ Morgan and $N = 10^4$. The Haldane map function $(1/2 - e^{-2x}/2)$ is in red, the Kosambi map function $(\tanh(2x)/2)$ is in yellow, and the complete interence map function (x) is in blue. For all values of g, the expectations are ordered in the same order as the map functions, but the difference between the three disappears by $g = 100$.

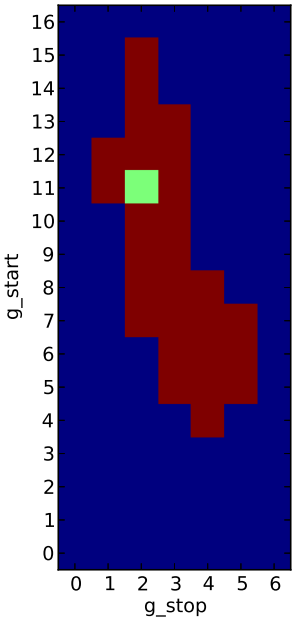

FIGURE 7. 95% confidence region for a model with constant admixture from generations g_{start} to g_{stop}. The point estimate of $g_{start} = 11$ and $g_{stop} = 2$ generations ago is colored green.

	Observed	Bootstrap	Corrected
k_1	0.777	-2.22×10^{-15}	0.777
k_2	9.00×10^{-3}	2.59×10^{-4}	8.75×10^{-3}
k_3	2.98×10^{-4}	1.60×10^{-5}	2.82×10^{-4}
k_4	-3.99×10^{-5}	-1.41×10^{-6}	-3.85×10^{-5}

FIGURE 8. k-statistics

APPENDIX

These are the matrices for computing $\mathbb{E}(k_3)$. The matrices for computing $\mathbb{E}(k_4)$ are 15×15 and not given here, but can be found in (Hill, 1974).

$$\mathbf{v}_{3(g)} = \begin{pmatrix} \mathbb{P}\{A_{1(g)}(\ell) = A_{1(g)}(\ell') = A_{1(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{1(g)}(\ell') = A_{2(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{2(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{1(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{3(g)}(\ell'') = 1\} \end{pmatrix}$$

$$\mathbf{U}_3 = \begin{pmatrix} [\ell\ell'\ell''] & [\ell\ell'|\ell''] & [\ell|\ell'\ell''] & [\ell\ell''|\ell'] & 0 \\ 0 & [\ell\ell'] & 0 & 0 & [\ell|\ell'] \\ 0 & 0 & [\ell'\ell''] & 0 & [\ell|\ell''] \\ 0 & 0 & 0 & [\ell\ell''] & [\ell'|\ell''] \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{L}_3 = \frac{1}{4N^2} \begin{pmatrix} 4N^2 & 0 & 0 & 0 & 0 \\ 2N & 2N-1 & 0 & 0 & 0 \\ 2N & 0 & 2N-1 & 0 & 0 \\ 2N & 0 & 0 & 2N-1 & 0 \\ 1 & 2N-1 & 2N-1 & 2N-1 & (2N-1)(2N-2) \end{pmatrix}$$

$$\mathbf{D}_{3(g)} = \begin{pmatrix} 1-s_g & 0 & 0 & 0 & 0 \\ 0 & (1-s_g)^2 & 0 & 0 & 0 \\ 0 & 0 & (1-s_g)^2 & 0 & 0 \\ 0 & 0 & 0 & (1-s_g)^2 & 0 \\ 0 & 0 & 0 & 0 & (1-s_g)^3 \end{pmatrix}$$

When there is migration from both source populations, the recursion relations for the i-point correlation functions will depend on $i-1$-point, $i-2$-point, ... correlations functions as well. As as example, consider the case of $\mathbf{v}_{2(g)}$. Let the introgression probability from the second source population be given by t_g. The recursion equation for $\mathbf{v}_{2(g)}$ now also depends on $\mathbf{v}_{1(g)}$.

$$\begin{aligned}\mathbf{v}_{2(g+1)} &= \mathbf{L}_2 \begin{pmatrix} 1-s_g-t_g & 0 \\ 0 & (1-s_g-t_g)^2 \end{pmatrix} \mathbf{U}_2 \mathbf{v}_{2(g)} + \begin{pmatrix} t_g \\ t_g^2 + 2t_g \mathbb{P}\{A_{1(g)}(\ell) = 1\} \end{pmatrix} \\ &= \mathbf{L}_2 \begin{pmatrix} 1-s_g-t_g & 0 \\ 0 & (1-s_g-t_g)^2 \end{pmatrix} \mathbf{U}_2 \mathbf{v}_{2(g)} + \begin{pmatrix} t_g \\ t_g^2 + 2t_g \mathbf{v}_{1(g)} \end{pmatrix}.\end{aligned}$$

Similarly, the recursion equation for $\mathbf{v}_{3(g)}$ depends on $\mathbf{v}_{2(g)}$ and $\mathbf{v}_{1(g)}$.