

1 UNDERSTANDING ADMIXTURE FRACTIONS

2 MASON LIANG 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 generations until 2 generations ago gives a better fit.

3 INTRODUCTION

4 It is common in population genetic analyses to consider individuals as
5 belonging fractionally to two or more discrete source populations. The pro-
6 portion of an individual's genome that belongs to a population is called
7 that individual's 'admixture fraction' or 'admixture proportion'. Programs
8 such as *Structure* (Pritchard et al., 2000), *Eigenstrat* (Price et al., 2006),
9 *Frappe* (Tang et al., 2005), or *Admixture* (Alexander et al., 2009) can jointly
10 estimate these admixture fractions for multiple individuals in a sample, along
11 with the corresponding allele frequencies in each of the source populations.
12 These admixture fractions are often presented in a 'structure plot,' an ex-
13 ample of which is shown in Figure 1. We will henceforth refer to these
14 methods as 'structure analyses'. This approach has proven highly useful for
15 understanding genetic relationships in many different species, e.g. humans

Date: July 3, 2014.

(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 possible 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 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, $H_{1(g)}, H_{2(g)}, \dots, H_{n(g)}$. We are interested in the joint distribution of these n random variables. When $n = 1$ and as $L \rightarrow \infty$, this is the admixture fraction considered by Verdu and Rosenberg (2011).

Because the n chromosomes have possibly overlapping genealogies, 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 $n \rightarrow \infty$, the sample mean and sample variance will converge to the mean and variance of \mathcal{G} , which are random quantities. For example,

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

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

117 and similarly for higher-order moments. The moments of the admixture
118 factions have two components: randomness from sampling the population
119 genealogy, and randomness from the sampling of chromosomes. The ex-
120 pressions to left account for both, while the expressions to the right only
121 account for the latter. Variances among individuals within one popula-
122 tion correspond to $\text{var}(H_{1(g)}|\mathcal{G})$, while variances over replicate populations
123 correspond to $\text{var}(H_{1(g)})$. This latter value will be larger than the expected
124 sample variance calculated from multiple individuals sampled from the same
125 population, and will rarely be useful for inference purposes.

126 In the following sections, we will show how the constants on the left-hand
127 side, as well as expectations of the random variables on the right-hand side,
128 can be derived for mechanistic models of introgression. By comparing these
129 expectations to the observed admixture parameters from a sample, we will
130 be able to construct a method of moments estimator for the parameters of
131 the model.

132 Let k_1 be the sample mean:

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

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

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

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

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

136 Its expectation is given by

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

137 These expectations can be written in terms of two-point correlation func-
138 tions of A :

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

139 Similarly,

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

140 Writing these two correlation functions as

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

141 we find that

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

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

$$\begin{aligned}\mathbb{E}(k_3) &= \frac{1}{L^3} \int_0^L \int_0^L \int_0^L \begin{pmatrix} 1 & -1 & -1 & -1 & 2 \end{pmatrix} \mathbf{v}_{3(g)} d\ell d\ell' d\ell'' \\ \mathbb{E}(k_4) &= \frac{1}{L^4} \int_{[0,L]^4} \begin{pmatrix} 1 & \underbrace{-1}_{4 \text{ times}} & \underbrace{-1}_{3 \text{ times}} & \underbrace{2}_{6 \text{ times}} & 6 \end{pmatrix} \mathbf{v}_{4(g)} d\ell d\ell' d\ell'' d\ell''' \\ &\dots\end{aligned}$$

145 Remarkably, the linear combinations required to compute the expecta-
146 tions of the k -statistics correspond exactly to the higher-order disequilibria
147 as defined by Bennett (1952). Furthermore, if instead the we choose to
148 compute the expectations of the h -statistics, which estimate the central
149 moments, the linear combinations would correspond to the higher-order dis-
150 equilibria as defined by Slatkin (1972).

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

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

159 If two sites at ℓ and ℓ' are on the same chromosome in generation $g + 1$,
160 then they were inherited from one chromosome from generation g with prob-
161 ability $[\ell\ell']$ and from two chromosomes from generation g with probability
162 $[\ell|\ell']$. If they are on different chromosomes, then the probability that they
163 are descended from one chromosome in generation g is $\frac{1}{2N}[\ell\ell']$ and the prob-
164 ability that they are descended from two chromosomes is $\frac{1}{2N}[\ell|\ell'] + (1 - \frac{1}{2N})$
165 In matrix notation,

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

166 where the the recombination and drift matrices are given by

$$\mathbf{L}_2 = \begin{pmatrix} 1 & 0 \\ \frac{1}{2N} & 1 - \frac{1}{2N} \end{pmatrix}$$

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

167 This is the the same matrix equation (Wright 1933 and Hill and Robertson
168 1966) derived for the decay of two-locus linkage disequilibrium. The ‘alleles’
169 we consider are the local ancestry at ℓ and ℓ' . To the extent possible, our
170 notation will follow (Hill 1974), whose results for measures of multi-locus
171 linkage disequilibria we use. The matrices \mathbf{L}_2 and \mathbf{U}_2 share $(1 \ -1)$ as a
172 left-eigenvector, with corresponding eigenvalues $1 - \frac{1}{2N}$ and $[\ell\ell']$. As a result,

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

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

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

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

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

177 For predicting the expected sample variance, the difference between these
178 two models is not large, as shown in figure 4. For the simulations and
179 inference in this paper, we will ignore crossover interference, and use the
180 Haldane map function. However, none of the mathematical results of this
181 paper will require this assumption.

182 For computing higher-order correlation functions, we find a similar equa-
183 tion

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

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

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

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

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

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

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

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

197 Note that if $s_{(g)}$ is fixed, then equation (4) is linear, and can be solved
198 using a Laplace transform.

199 INFERENCE OF ADMIXTURE TIMES

200 The equations in the previous section can be used to develop a method
201 of moments-estimators for admixture parameters by numerically solving the
202 admixture parameters in terms of the expectations for the k -statistics. Sub-
203 stituting in the observed values for the k -statistics gives estimates for the
204 admixture parameter(s).

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

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

209 Because cumulants are additive,

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

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

213 One additional complication arises in dealing with genotyping data. We
214 have assumed that we have the ancestry fractions for each haplotype in the
215 sample, but with genotyping data, we instead have their pairwise means:
216 $(H_{1(g)} + H_{2(g)})/2 \dots$. This results in a decrease in the expectations of
217 the k -statistics. Conditional on the random distribution \mathcal{G} , $H_{1(g)}, H_{2(g)}, \dots$
218 are *iid* drawn from \mathcal{G} . Cumulants are additive, so we use the law of total
219 expectation to find that

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

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

225 In figure 2, we have shown the admixture fractions for five replicate pop-
226 ulations 5, 50, and 500 generations after an admixture pulse. The variance
227 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 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 3. 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

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

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

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

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.

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 1 and the observed k -statistics are given in table 6.

We assumed a 3-parameter model of constant admixture. For $g_{start} \leq g \leq g_{stop}$, $s_g = s$ with $s_g = 0$ elsewhere. By matching the block-bootstrap corrected k_2 and k_3 to the predictions of equation 1, we obtained a point estimates of

$$\hat{s} = 0.0277$$

$$\hat{g}_{start} = 2$$

$$\hat{g}_{stop} = 11.$$

We obtained confidence intervals, shown in Figure 5, 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 Europe 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 probably reflects a more gradual reduction in gene-flow within the last 5 generations or so, rather than a discrete stop in gene-flow 2 generations ago. The discreteness is enforced by the model. 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.

- 322 William G Hill. Disequilibrium among several linked neutral genes in fi-
323 nite population i. mean changes in disequilibrium. *Theoretical Population*
324 *Biology*, 5(3):366–392, 1974.
- 325 Mason Liang and Rasmus Nielsen. The lengths of admixture tracts. *Genet-*
326 *ics*, pages genetics–114, 2014.
- 327 Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante.
328 Rfmix: A discriminative modeling approach for rapid and robust local-
329 ancestry inference. *The American Journal of Human Genetics*, 93(2):
330 278–288, 2013.
- 331 Marilyn Menotti-Raymond, Victor A David, Solveig M Pflueger, Kerstin
332 Lindblad-Toh, Claire M Wade, Stephen J OBrien, and Warren E Johnson.
333 Patterns of molecular genetic variation among cat breeds. *Genomics*, 91
334 (1):1–11, 2008.
- 335 John E Pool and Rasmus Nielsen. Inference of historical changes in migration
336 rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.
- 337 Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt,
338 Nancy A Shadick, and David Reich. Principal components analysis cor-
339 rectes for stratification in genome-wide association studies. *Nature Genet-*
340 *ics*, 38(8):904–909, 2006.
- 341 Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference
342 of population structure using multilocus genotype data. *Genetics*, 155(2):
343 945–959, 2000.
- 344 Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M
345 Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman.
346 Genetic structure of human populations. *Science*, 298(5602):2381–2385,
347 2002.
- 348 Montgomery Slatkin. On treating the chromosome as the unit of selection.
349 *Genetics*, 72(1):157–168, 1972.
- 350 Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual
351 admixture: analytical and study design considerations. *Genetic epidemi-*
352 *ology*, 28(4):289–301, 2005.
- 353 Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Recon-
354 structing genetic ancestry blocks in admixed individuals. *The American*
355 *Journal of Human Genetics*, 79(1):1–12, 2006.
- 356 Paul Verdu and Noah A Rosenberg. A general mechanistic model for admix-
357 ture histories of hybrid populations. *Genetics*, 189(4):1413–1426, 2011.

358 Baowei Zhang, Ming Li, Zejun Zhang, Benoît Goossens, Lifeng Zhu, Shan-
359 ning Zhang, Jinchu Hu, Michael W Bruford, and Fuwen Wei. Genetic
360 viability and population history of the giant panda, putting an end to the
361 evolutionary dead end? *Molecular biology and evolution*, 24(8):1801–1810,
362 2007.

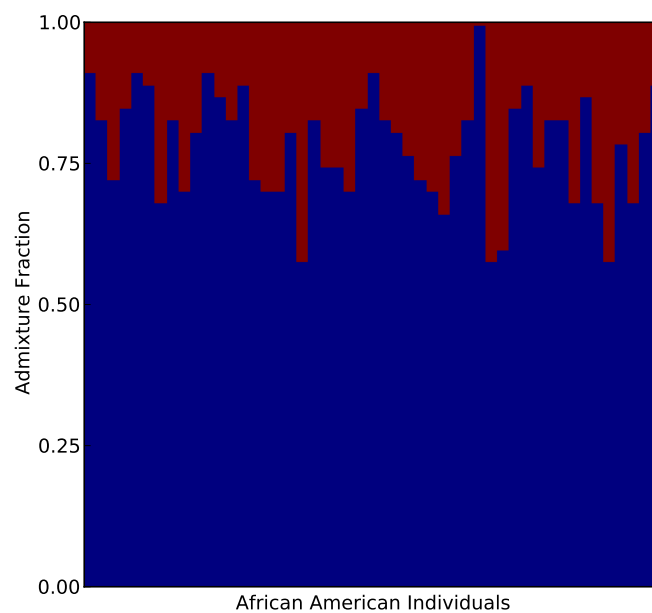


FIGURE 1. Admixture fractions for 49 African American individuals in the HapMap 3 data. Source population allele frequencies were estimated using 113 Yoruban and 111 European individuals.

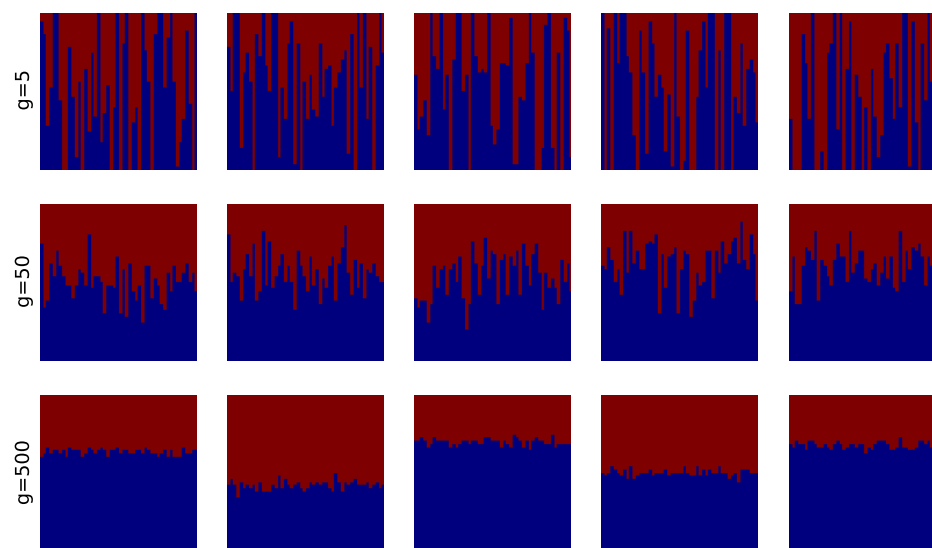


FIGURE 2. The admixture fractions of five replicate populations (each column) 5, 50, and 500 generations after an admixture pulse. As the admixture event grows more ancient, the variability within a replicate population decreases, but some variability is still maintained across the populations.

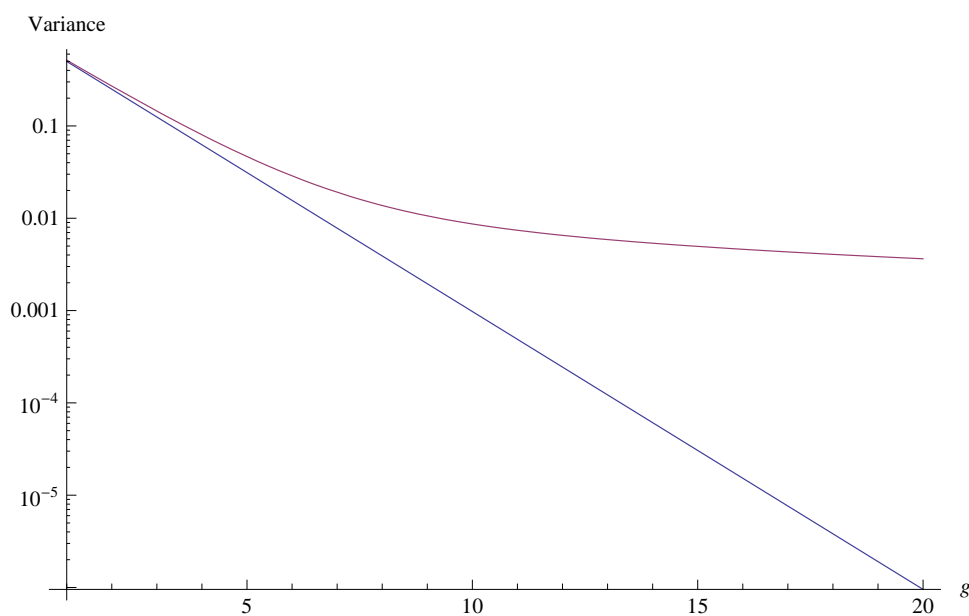


FIGURE 3. The variance predicted by Verdu and Rosenberg (2011) and equation 5, plotted on a logarithmic scale. The variance we predict (red) is always larger, but the two are very similar when g is small.

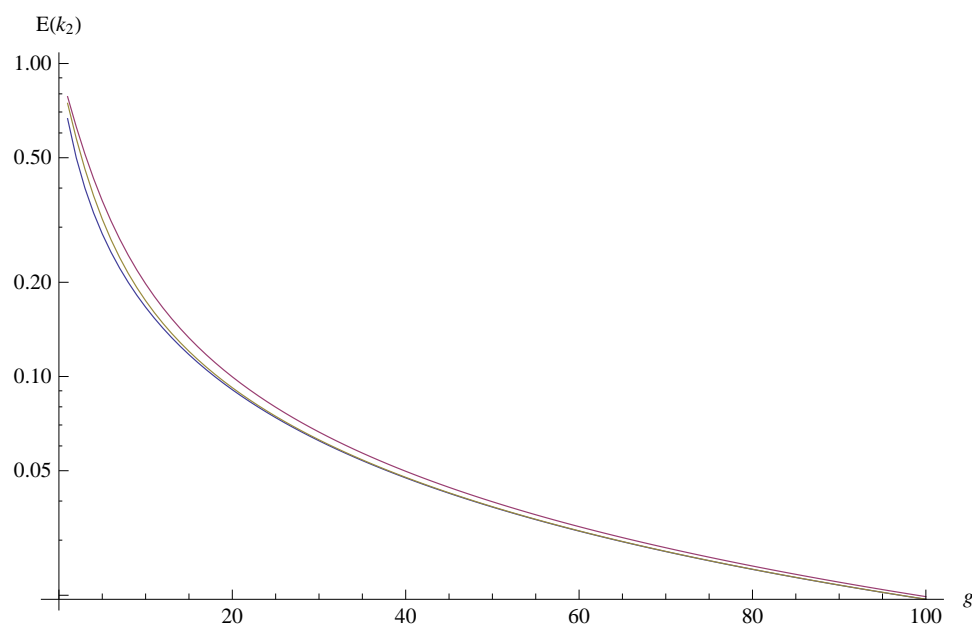


FIGURE 4. The expected sample variance given by equation 1 plotted on a logarithmic scale, for a three different map functions. We used a map distance of $L = 1$ Morgan and $N = 10^4$. The Haldane map function $(1/2 - e^{-2x}/2)$ is in red, the Kosambi map function $(\tanh(2x)/2)$ is in yellow, and the complete intercence 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 $g = 100$.

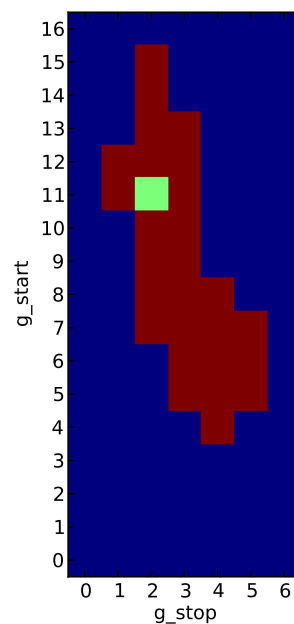


FIGURE 5. 95% confidence region for a model with constant admixture from generations g_{start} to g_{stop} . The point estimate of $g_{start} = 11$ and $g_{stop} = 2$ generations ago is colored green.

	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 6. k -statistics

363

APPENDIX

364 These are the matrices for computing $\mathbb{E}(k_3)$. The matrices for computing
365 $\mathbb{E}(k_4)$ are 15×15 and not given here, but can be found in (Hill, 1974).

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

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

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

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

366 When there is migration from both source populations, the recursion re-
367 lations for the i -point correlation functions will depend on $i-1$ -point, $i-2$ -
368 point, \dots correlations functions as well. As an example, consider the case of
369 $\mathbf{v}_{2(g)}$. Let the introgression probability from the second source population
370 be given by t_g . The recursion equation for $\mathbf{v}_{2(g)}$ now also depends on $\mathbf{v}_{1(g)}$.

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

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