

1
2

3
4

5

6

4
5

1

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

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

46 Verdu and Rosenberg (2011) developed a method for computing moments
 47 of admixture proportions in a model in which admixed population is formed
 48 as a mixture between multiple source populations, allowing for arbitrary
 49 gene-flow from the source populations over a number of generations (g).
 50 They establish recursions for the moments of the admixture fractions and
 51 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 in Verdu and Rosenberg (2011) to incorporate the effects of recombination and genetic drift by adding 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,

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

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

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

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

134 Let k_1 be the sample mean:

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

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

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

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

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

138 Its expectation is given by

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

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

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

141 Similarly,

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

142 Writing these two correlation functions as

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

143 we find that

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

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

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

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 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 $\frac{1}{2N}[\ell|\ell'] + (1 - \frac{1}{2N})$. 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 recombination and drift matrices are given by

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

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

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

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

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

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

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

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

179 For predicting the expected sample variance, the difference between these
180 two models is not large, as shown in figure 4. For the simulations and
181 inference in this paper, we will ignore crossover interference, and use the
182 Haldane map function. However, none of the mathematical results of this
183 paper will require this assumption of no crossover interference.

184 For computing higher-order correlation functions, we find a similar equa-
 185 tion

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

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

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

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

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

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

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

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

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

201 INFERENCE OF ADMIXTURE TIMES

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

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

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

211 Because cumulants are additive,

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

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

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

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

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

227 In figure 2, we have shown the admixture fractions for five replicate pop-
 228 ulations 5, 50, and 500 generations after an admixture pulse. The variance
 229 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 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 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

$$\begin{aligned}\hat{s} &= 0.0277 \\ \hat{g}_{start} &= 2 \\ \hat{g}_{stop} &= 11.\end{aligned}$$

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.

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

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

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

315 REFERENCES

- 316 David H Alexander, John Novembre, and Kenneth Lange. Fast model-based
317 estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):
318 1655–1664, 2009.
319 John Bennett. On the theory of random mating. *Annals of Eugenics*, 17(1):
320 311–317, 1952.

- 321 International HapMap 3 Consortium et al. Integrating common and rare
 322 genetic variation in diverse human populations. *Nature*, 467(7311):52–58,
 323 2010.
- 324 Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference
 325 of population structure using multilocus genotype data: linked loci and
 326 correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.
- 327 Simon Gravel. Population genetics models of local ancestry. *Genetics*, 191
 328 (2):607–619, 2012.
- 329 William G Hill. Disequilibrium among several linked neutral genes in fi-
 330 nite population i. mean changes in disequilibrium. *Theoretical Population*
 331 *Biology*, 5(3):366–392, 1974.
- 332 Mason Liang and Rasmus Nielsen. The lengths of admixture tracts. *Genet-*
 333 *ics*, pages genetics–114, 2014.
- 334 Brian K Maples, Simon Gravel, Eimear E Kenny, and Carlos D Bustamante.
 335 Rfmix: A discriminative modeling approach for rapid and robust local-
 336 ancestry inference. *The American Journal of Human Genetics*, 93(2):
 337 278–288, 2013.
- 338 Marilyn Menotti-Raymond, Victor A David, Solveig M Pflueger, Kerstin
 339 Lindblad-Toh, Claire M Wade, Stephen J O’Brien, and Warren E Johnson.
 340 Patterns of molecular genetic variation among cat breeds. *Genomics*, 91
 341 (1):1–11, 2008.
- 342 John E Pool and Rasmus Nielsen. Inference of historical changes in migration
 343 rate from the lengths of migrant tracts. *Genetics*, 181(2):711–719, 2009.
- 344 Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt,
 345 Nancy A Shadick, and David Reich. Principal components analysis cor-
 346 rects for stratification in genome-wide association studies. *Nature Genet-*
 347 *ics*, 38(8):904–909, 2006.
- 348 Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference
 349 of population structure using multilocus genotype data. *Genetics*, 155(2):
 350 945–959, 2000.
- 351 Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M
 352 Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman.
 353 Genetic structure of human populations. *Science*, 298(5602):2381–2385,
 354 2002.
- 355 Montgomery Slatkin. On treating the chromosome as the unit of selection.
 356 *Genetics*, 72(1):157–168, 1972.

- 357 Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual
358 admixture: analytical and study design considerations. *Genetic epidemi-*
359 *ology*, 28(4):289–301, 2005.
- 360 Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Recon-
361 structing genetic ancestry blocks in admixed individuals. *The American*
362 *Journal of Human Genetics*, 79(1):1–12, 2006.
- 363 Paul Verdu and Noah A Rosenberg. A general mechanistic model for admix-
364 ture histories of hybrid populations. *Genetics*, 189(4):1413–1426, 2011.
- 365 Baowei Zhang, Ming Li, Zejun Zhang, Benoît Goossens, Lifeng Zhu, Shan-
366 ning Zhang, Jinchu Hu, Michael W Bruford, and Fuwen Wei. Genetic
367 viability and population history of the giant panda, putting an end to
368 the “evolutionary dead end”? *Molecular biology and evolution*, 24(8):
369 1801–1810, 2007.

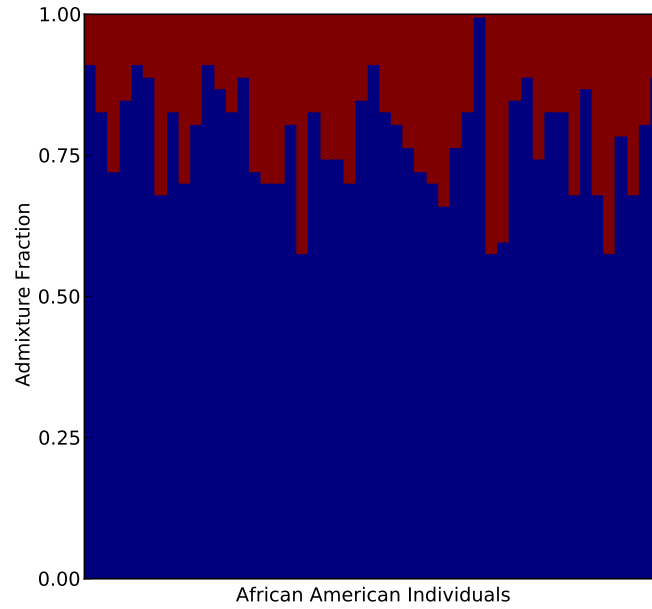


FIGURE 1. Admixture fractions for 49 African American individuals in the HapMap 3 data. Source population allele frequencies were estimated using 113 Yoruban and 111 European individuals.

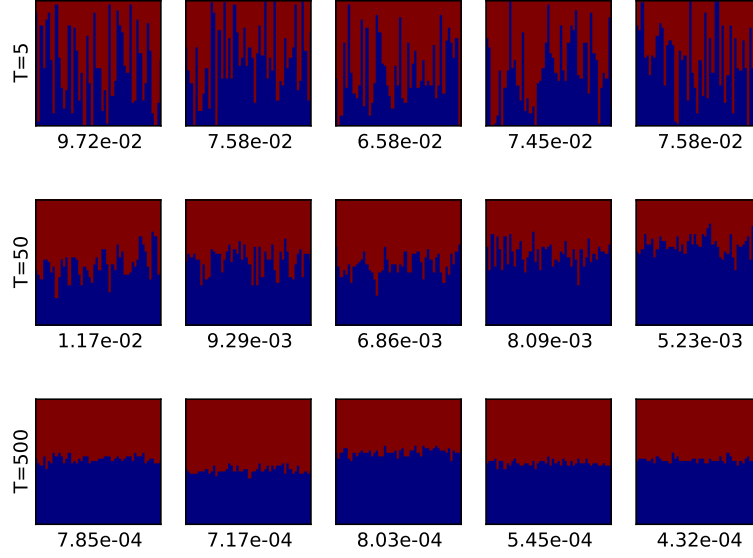


FIGURE 2. The admixture fractions of five replicate populations (each row) 5, 50, and 500 generations after an admixture pulse. The number underneath each structure plot is k_2 , the sample variance, for that replicate population. For more ancient admixture events, the variability within a replicate population decreases to 0, but some variability is still maintained across the populations.

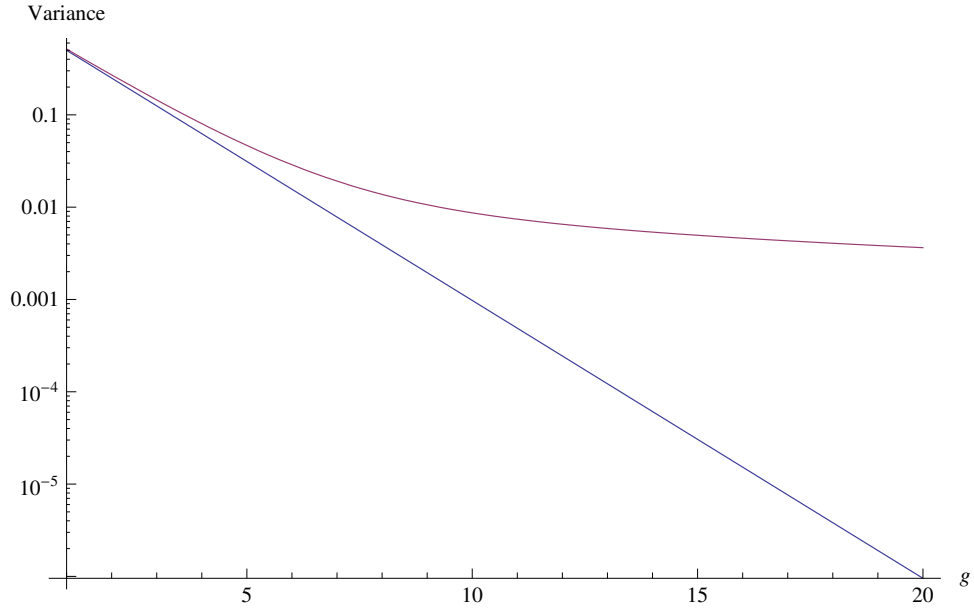


FIGURE 3. The variance predicted by Verdu and Rosenberg (2011) and equation 5, plotted on a logarithmic scale. The variance we predict (red) is always larger, but the two are very similar when g is small.

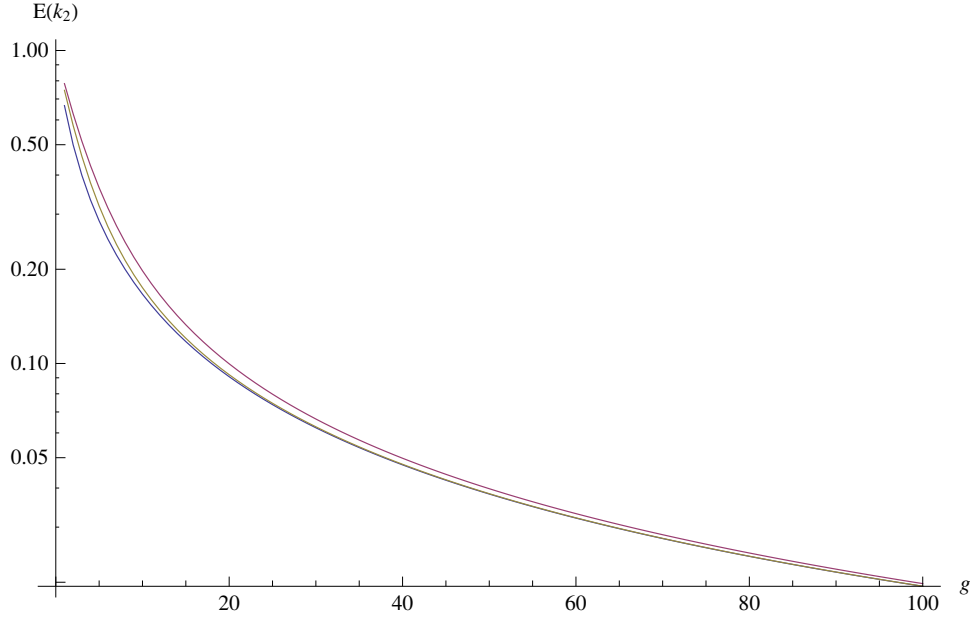


FIGURE 4. The expected sample variance given by equation 1 plotted on a logarithmic scale, for a three different map functions. We used a map distance of $L = 1$ Morgan and $N = 10^4$. The Haldane map function $(1/2 - e^{-2x}/2)$ is in red, the Kosambi map function $(\tanh(2x)/2)$ is in yellow, and the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

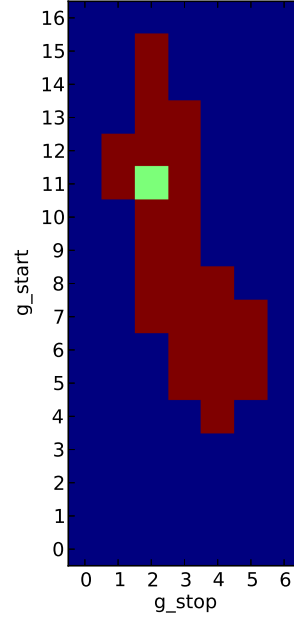


FIGURE 5. 95% confidence region for a model with constant admixture from generations g_{start} to g_{stop} . The point estimate of $g_{start} = 11$ and $g_{stop} = 2$ generations ago is colored green.

	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

370

APPENDIX

371 These are the matrices for computing $\mathbb{E}(k_3)$. The matrices for computing
 372 $\mathbb{E}(k_4)$ are 15×15 and not given here, but can be found in (Hill, 1974).

$$\begin{aligned}
 \mathbf{v}_{3(g)} &= \begin{pmatrix} \mathbb{P}\{A_{1(g)}(\ell) = A_{1(g)}(\ell') = A_{1(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{1(g)}(\ell') = A_{2(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{2(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{1(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{3(g)}(\ell'') = 1\} \end{pmatrix} \\
 \mathbf{U}_3 &= \begin{pmatrix} [\ell\ell'\ell''] & [\ell\ell'|\ell''] & [\ell|\ell'\ell''] & [\ell\ell''|\ell'] & 0 \\ 0 & [\ell\ell'] & 0 & 0 & [\ell|\ell'] \\ 0 & 0 & [\ell'\ell''] & 0 & [\ell|\ell''] \\ 0 & 0 & 0 & [\ell\ell''] & [\ell'|\ell''] \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
 \mathbf{L}_3 &= \frac{1}{4N^2} \begin{pmatrix} 4N^2 & 0 & 0 & 0 & 0 \\ 2N & 2N-1 & 0 & 0 & 0 \\ 2N & 0 & 2N-1 & 0 & 0 \\ 2N & 0 & 0 & 2N-1 & 0 \\ 1 & 2N-1 & 2N-1 & 2N-1 & (2N-1)(2N-2) \end{pmatrix} \\
 \mathbf{D}_{3(g)} &= \begin{pmatrix} 1-s_g & 0 & 0 & 0 & 0 \\ 0 & (1-s_g)^2 & 0 & 0 & 0 \\ 0 & 0 & (1-s_g)^2 & 0 & 0 \\ 0 & 0 & 0 & (1-s_g)^2 & 0 \\ 0 & 0 & 0 & 0 & (1-s_g)^3 \end{pmatrix}
 \end{aligned}$$

373 When there is migration from both source populations, the recursion re-
 374 lations for the i -point correlation functions will depend on $i-1$ -point, $i-2$ -
 375 point, \dots correlations functions as well. As an example, consider the case of
 376 $\mathbf{v}_{2(g)}$. Let the introgression probability from the second source population
 377 be given by t_g . The recursion equation for $\mathbf{v}_{2(g)}$ now also depends on $\mathbf{v}_{1(g)}$.

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

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