Title
On the Distribution of Tract Lengths During Adaptive Introgression.

Permalink
https://escholarship.org/uc/item/2sw2d3rn

Journal
G3 (Bethesda, Md.), 10(10)

ISSN
2160-1836

Authors
Shchur, Vladimir
Svedberg, Jesper
Medina, Paloma
et al.

Publication Date
2020-10-05

DOI
10.1534/g3.120.401616

Peer reviewed
On the Distribution of Tract Lengths During Adaptive Introgression

Vladimir Shchur,*†,‡,1 Jesper Svedberg,† Paloma Medina,§ Russell Corbett-Detig,†‡ and Rasmus Nielsen*†,‡,§

*Department of Integrative Biology and Department of Statistics, UC Berkeley, Berkeley, CA, ‡National Research University Higher School of Economics, Russian Federation, †Department of Biomolecular Engineering and Genomics Institute, UC Santa Cruz, California 95064, and §Center for GeoGenetics, Globe Institute, University of Copenhagen, 2100 Denmark

ORCID IDs: 0000-0002-1431-5260 (V.S.); 0000-0001-6760-3525 (P.M.); 0000-0001-6535-2478 (R.C.-D.); 0000-0003-0513-6591 (R.N.)

ABSTRACT Admixture is increasingly being recognized as an important factor in evolutionary genetics. The distribution of genomic admixture tracts, and the resulting effects on admixture linkage disequilibrium, can be used to date the timing of admixture between species or populations. However, the theory used for such prediction assumes selective neutrality despite the fact that many famous examples of admixture involve natural selection acting for or against admixture. In this paper, we investigate the effects of positive selection on the distribution of tract lengths. We develop a theoretical framework that relies on approximating the trajectory of the selected allele using a logistic function. By numerically calculating the expected allele trajectory, we also show that the approach can be extended to cases where the logistic approximation is poor due to the effects of genetic drift. Using simulations, we show that the model is highly accurate under most scenarios. We use the model to show that positive selection on average will tend to increase the admixture tract length. However, perhaps counter-intuitively, conditional on the allele frequency at the time of sampling, positive selection will actually produce shorter expected tract lengths. We discuss the consequences of our results in interpreting the timing of the introgression of EPAS1 from Denisovans into the ancestors of Tibetans.

Admixture—a process wherein genetically distinct populations hybridize and exchange alleles—is increasingly recognized as a dominant force in evolutionary genomics. In particular, admixture has the potential to introduce adaptive alleles from a donor population to a recipient, particularly if the donor population has already adapted to e.g., unique environmental conditions. For example, when the ancestors of modern ethnic Tibetans first colonized the extremely high elevation Tibetan Plateau their success in adapting to these harsh conditions stemmed in part from genes acquired from a Denisovan-like ancestral population (Huerta-Sánchez et al. 2014). More generally, as genome sequence data continues to accumulate, it is increasingly clear that introgression has played a major role in shaping adaptive outcomes across a wide range of species, e.g., (Hufford et al. 2013; Martin et al. 2013). However, the specific genetic signature of adaptive introgression and its impact on patterns of genetic variation remains less explored (but see e.g., (Setter et al. 2019)). There is, therefore, a clear need to develop a theoretical framework to investigate the impacts of adaptive introgression on the genomes of natural populations.

When two ancestral populations hybridize, the genome of each individuals is a mixture of the ancestral populations, and the ancestry at each site in the genome can be traced back to a single population. Because alleles from a given ancestral population are initially on the same chromosomes, and therefore strongly linked within the admixed population, this process can dramatically shape linkage disequilibrium within admixed populations (commonly termed admixture LD (Stephens et al. 1994; Falush et al. 2003)). Admixture LD, and the haplotypic analog, ancestry tracts, which are unbroken stretches of sites from one ancestral population, have become the key elements of genomic inference in studying admixed populations. These two features of genetic variation are the primary units of demographic model inference within admixed populations. For example, the decay of LD (Loh et al. 2013) and the ancestry tract length distribution (Gravel 2012; Corbett-Detig and Nielsen 2017) (Skov et al. 2018) are commonly used for timing the onset of admixture assuming a neutral model during admixture.
(Hill 1974; Liang and Nielsen, 2014b, 2016) showed recombination, genetic drift, and migration processes act as linear operators on the 2-locus and 3-locus linkage disequilibrium coefficients (as defined in (Lewontin and ichi Kojima 1960) and (Slatkin 1972) respectively). In contrast, natural selection is not a linear operator on the LD coefficients, and these approaches can, therefore, not be extended to accommodate the impacts of natural selection. Instead, other methods are required to analyze the impact of selection. A recent study (Sachdeva and Barton 2018) derives the ancestry tract length distribution under an infinitesimal selection model. Such a model assumes that there are many different sites and each of them is under weak selection. However, another potent force that affects admixed populations is adaptive introgression, where one or a few adaptive genes is acquired via hybridization and each has potentially strong effects on fitness. The analysis of adaptive introgression and its impacts on linkage disequilibrium and the ancestry tract length distribution is a largely unexplored topic and requires a novel theoretical framework.

The distribution of tract lengths during adaptive introgression is closely related to the study of the strength of linkage disequilibrium. Two main models that address this question are 2-locus and 3-locus linkage models, which model the joint behavior of 2 and 3 linked loci respectively. Substantial theory and numerical simulations are known from earlier works, e.g., (Kimura 1956; Maynard and Haigh 2007; Thomson 1977; Lewontin and ichi Kojima 1960; Ewens 1968; Karlin and Feldman 1970; Feldman et al. 1974; Karlin and Carmelli 1975; Karlin 1975; Lewontin, 1964a, b; Franklin and Lewontin 1970)). There are two main differences between these works and our approach. The first is the problem itself: in the aforementioned works the object of interest is selected site. The second difference is that aforementioned works consider allele frequencies at all the sites affect the strength of LD. In this paper sites are considered to be segregating in a single ancestral population, and primarily the strength of LD in different selection scenarios, so all the main differences between these works and our approach. The Feldman 1970; Feldman 1977; Lewontin and ichi Kojima 1960; Ewens 1968; Karlin and Carmelli 1975; Karlin 1975; Lewontin, 1964a, b; Franklin and Lewontin 1970).

The expected trajectory of an allele under selection can be well approximated using a logistic function (see e.g., (Kaplan et al. 1989)) under the following conditions: First, selection should be strong ($Nes > 1$, where $N$ is effective population size, and $s$ is selection coefficient such that an individual with two selected alleles has fitness $1+s$ and a heterozygous individual has fitness $1+s/2$). Second, the frequency of the selected allele is above a critical threshold, so in our model, the admixture fraction must not be too small. Finally, the time since adaptive introgression is not too large compared to the selection coefficient, so that the frequency of $A$ is not too close to 1 at the time of sampling (Smith 1971; Messer and Neher 2012).

Approximating the allele frequency trajectory of the adaptive allele using a deterministic logistic trajectory, as in (Kaplan et al. 1989), allows us to avoid integration over the stochasticity caused by genetic drift, which makes our approach similar to the mean field theory widely used in physics (Weiss 1907) and epidemiology (Kermack William Ogilvy and Thomas 1927). We show that this approximation is highly accurate, in a certain range of parameters, for estimating the expected tract lengths. In this range, the expected tract length can be considered invariant under the order of integration over genetic drift and over the tract length distribution.

Outside this range, when the proportion of introgression is very small, which means that drift is strong relative to selection, the logistic function will yield inaccurate predictions. This observation matches that of (Durrett and Schweinsberg 2004). In this case the expected
trajectory can be estimated by generating a large random sample of stochastic trajectories. This can be done efficiently, because, in practice, the selection coefficient $s \ll 1$ is small, and the diploid population can be approximated by a haploid population under genic selection. At each generation the selected allele frequency is simulated from a binomial distribution with probability weighted by the total fitness contributed of the selected allele. We average allele frequencies at each generation over the simulated trajectories (conditioning on observing the selected allele at the time of sampling). We will show that such an approach vastly extends the applicability of our method to cases that require greater stochasticity. In particular, we use it to analyze the Denisovan introgression in the ancestry of Tibetans, where the proportion of introgression is estimated to be as small as 0.06% (Huerta-Sánchez et al. 2014), and therefore is outside the range of parameters where the logistic approximation to the allele frequency trajectory yields accurate results.

### 3-locus model during adaptive introgression

Let $\beta$ and $\gamma$ be two neutral loci near $\alpha$ (the locus under selection) so that there are 3 loci on the chromosome: $\alpha$, $\beta$, and $\gamma$, following each other sequentially in this order on the chromosome. Let $r_1$ be the distance (in Morgans) between $\alpha$ and $\beta$, $r_2$ is the distance between $\beta$ and $\gamma$. We are following the ancestry of $\beta$ and $\gamma$. In other words, we track the lineage leading to each locus to the time of introgression, while simultaneously tracking how recombination and coalescence act on the segment. If at the time of introgression the ancestral locus is on a haplotype carrying allele $\alpha$, then, by definition, its ancestry is from the introgressed population (denoted ancestry type 1). If, in contrast, the ancestral locus is linked to allele $a$, then it comes from the recipient population (denoted ancestry type 0). Ancestry type 0 corresponds to the recipient population and ancestry type 1 corresponds to the donor population.

The importance of this 3-locus model is that we can calculate transition rates between ancestry types, and hence, we can numerically calculate the distribution of tract lengths of ancestry type 1 (or type 0) near the locus $\alpha$.

### Model derivation

Under the coalescent with recombination, the model can be considered Markovian backward in time, when conditioned on the allele frequency path. To describe the dynamics of the adaptive introgression 3-locus model, we need to enumerate the possible states, which describes the ancestry configurations across the three loci, and we need to find the transition rates between them.

The model has 6 possible states. Each state represents an ancestral configuration for an observed chromosome with three loci: $\alpha$, $\beta$, and $\gamma$. At $\alpha$ we track the allelic state, $\alpha$ or $a$, which also indicates ancestry. In $\beta$ and $\gamma$ we only need to know if the given chromosome is ancestral to the observed chromosome or not. We use the notation $\beta^n$ and $\gamma^n$ to indicate DNA in $\beta$ or $\gamma$ that is ancestral to the observed chromosome. $\beta^n$ and $\gamma^n$ is used to indicate DNA that is not ancestral to the observed chromosome. The six states are then:

- $(a - \beta^n - \gamma^n)$: ancestral material at $\beta$ and $\gamma$ on a chromosome carrying allele $A$ at the selected locus $\alpha$.
- $(a - \beta^n - \gamma^n, A - \beta^n - \gamma^n)$: ancestral material at $\beta$ and $\gamma$ on different chromosomes, both carrying allele $A$.
- $(a - \beta^n - \gamma^n, a - \beta^n - \gamma^n)$: ancestral material at $\beta$ and $\gamma$ on different chromosomes, the chromosome ancestral to the observed chromosome at $\beta$ is carrying allele $A$, while the chromosome ancestral to the observed chromosome at $\gamma$ is carrying allele $a$.
- $(a - \beta^n - \gamma^n, A - \beta^n - \gamma^n)$: ancestral material at $\beta$ and $\gamma$ on different chromosomes, and the ancestral chromosome at $\beta$ is carrying allele $a$, while the ancestral chromosome at $\gamma$ is carrying allele $A$.
- $(a - \beta^n - \gamma^n, a - \beta^n - \gamma^n)$: the ancestral material of $\beta$ and $\gamma$ are on different chromosomes, both carrying allele $a$.
- $(a - \beta^n - \gamma^n)$: ancestral material at both $\beta$ and $\gamma$ are on the same chromosome carrying allele $a$.

We denote the frequency of the selected allele at time $t$ by $\omega(t)$. As we indicated previously, we assume that $\omega(t)$ deterministically follows a logistic function:

$$\omega(t) = 1 - \frac{1}{1 + e^{-\alpha t^2}} = \frac{1}{1 + e^{\alpha t^2}},$$

because we are working in backward time. This is a reflection of a logistic function relative to $t = 0$. The time of introgression corresponds to a point, $t_1$, on the deterministic allele frequency trajectory such that $\omega(t_1)$ equals the admixture proportion $\omega_0$. Notice, that the time of sampling $t_0 = t_1 - T$, where $T$ is the time since introgression, does not necessarily equal $t = 0$.

Recombination acts at a rate proportional to the recombination distances between loci. We assume that recombination between loci $\alpha$ and $\beta$ occurs at rate $r_1$ and recombination occurs between loci $\beta$ and $\gamma$ at rate $r_2$. In other words, we measure distance between the loci in Morgans.

Consider for example the ancestry configuration state $(A - \beta^n - \gamma^n)$ at time $t$. If recombination occurs between loci $\alpha$ and $\beta$, then ancestral material at loci $\beta$ and $\gamma$ remain on the same ancestral chromosome. The allele at the site $\alpha$ on this ancestral chromosome will be $A$ with probability $\omega(t)$ (state does not change) and $a$ with the probability $1 - \omega(t)$. Hence, the transition rate from state $(A - \beta^n - \gamma^n)$ to the state $(a - \beta^n - \gamma^n)$ is $r_1(1 - \omega(t))$ at time $t$. If a recombination event occurs on a chromosome with configuration $(A - \beta^n - \gamma^n)$ between loci $\beta$ and $\gamma$, then it is split into two chromosomes, and there are two different possible ancestry configurations. In both configurations the first ancestral chromosome is $A - \beta^n - \gamma^n$. The second chromosome is either $A - \beta^n - \gamma^n$ with probability $\omega(t)$, or $a - \beta^n - \gamma^n$ with probability $1 - \omega(t)$. So, recombinations are the first type of transition in our model. The other type of transition is coalescence, which is possible between chromosomes that carry the same allele ($A$ or $a$) at the selected locus $\alpha$ at the rate $A/\omega(t)$ for $A$ and $A/\omega(t)$ for $a$, where $A = 1/2N_i$ and $2N_i$ is the haploid effective population size. The full transition rate matrix is

$$\mathbb{P}(t) = \mathbb{P}(t) \mathbb{M}(t).$$

The initial condition for this differential equation is

$$\mathbb{P}(t_0) = (1, 0, 0, 0, 0, 0)$$

for the dynamics of introgressed tracts (hence, carrying allele $A$), and

$$\mathbb{P}(t_0) = (0, 0, 0, 0, 0, 1)$$

for the tracts from the recipient population (individuals with allele $a$).

Equation 1 cannot be solved analytically, because eigenvalues of the matrix $\mathbb{M}(t)$ cannot be derived analytically in general. However, it
and this position has ancestry type 0. Moreover, all the loci between left of this position has ancestry type 1 and a loci ($0$ is equal to the probability that the ancestral chromosome carries given model the probability that a locus is of ancestry type 1 or type set of ancestry tracts with ancestry type 1 at the selected site. In the \end{quote}

As previously discussed, we are interested in transitions of ancestry at a locus while moving along the chromosome away from the selected site. The process is only approximate because the real ancestry at a locus while moving along the chromosome away from

The transition rate is $$\tau_{10}(r) = \lim_{dr \to 0} \frac{P(S(r + dr) = 0|S(r) = 1)}{dr} = \lim_{dr \to 0} \frac{1}{dr} \frac{P(S(r + dr) = 0, S(r) = 1)}{P(S(r) = 1)}.$$ (2)

The numerator $P(S(r + dr) = 0, S(r) = 1)$ is the probability of being at the third state ($A - \beta^a - \gamma^a, a - \beta^a - \gamma^a$), and the denominator $P(S(r) = 1)$ is the marginal probability of ancestry type 1, hence it is the sum of probabilities of the first three ancestral configurations ($A - \beta^a - \gamma^a, (A - \beta^a - \gamma^a, A - \beta^a - \gamma^a)$ and $(A - \beta^a - \gamma^a, a - \beta^a - \gamma^a$). Expression 2 is easy to evaluate numerically by considering sufficiently small values of $r_2$.

\textbf{Model validation}

In order to estimate how accurate our deterministic approximation is, we compared our results with the average tract length estimated from simulations using the forward-in-time simulation framework SELAM (Corbett-Detig and Jones 2016). Briefly, we used the software to simulate a Wright-Fisher population of constant size 5000 hermaphroditic diploid individuals. We simulated a single chromosome of length one Morgan with a single selected site at position 0.5 Morgans. At each sampling point, we extracted 50 individuals (100 chromosomes) at random from within the population and output the ancestry across each chromosome. We performed 10 to 100 thousand replicates of each combination of selective coefficient, times since admixture and admixture proportion.

\textbf{Data availability}

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. The code is available at the GitHub repository associated with this project https://github.com/vlshchur/DAIM.

\textbf{RESULTS}

\textbf{Accuracy of deterministic approximation}

We simulate both relatively weak and strong selection, and we allow selection to act in different periods of time. We mostly explore the accuracy through the comparison of expected tract length and its variance. We also present the distribution of break-points of tracts containing selected allele. The tract length is the sum of two independent random variables, one representing the distance to the left of the selected mutation and the other representing the distance to the right of the selected mutation.

As the waiting time to a recombination event along the length of the genome is exponentially distributed, one might be inclined to think that tracts lengths also are exponentially distributed. However, this is not the case as illustrated in Figure 1. The figure shows the distribution of the distance from the selected locus to one end of the tract, estimated from simulations, from the deterministic
approximation, and an exponential distribution with the rate equal to the inverse of the mean of the simulated distribution. In this scenario, the approximation provides an accurate estimate of the empirical distribution. The exponential distribution provides a much worse fit. The reason why the exponential distribution fits so poorly is that it does not incorporate the possibility of back-recombination and coalescence. In the presence of these processes, the tract length distribution is not exponential (see also (Liang and Nielsen, 2014a) for more discussion of this).

In all of the scenarios that we explored, the relative error does not exceed 5%, and in more than half of all cases it is within 2% (Tables 1 and 2). This level of precision should be sufficient for most applications given that the uncertainty in the real data analysis is usually much greater than the error of the approximation model.

As previously discussed, the logistic function is a good approximation for the allele frequency trajectory only when selection is strong. Still, we wondered if the deterministic approach would give consistent results for a case of neutral admixture without adaptive introgression. Following Liang, Nielsen (Liang and Nielsen, 2014a) the expected tract length under a neutral admixture model can be calculated as

$$\frac{2}{2N_e(1-\omega)}(1-e^{-T/2N_e})$$

under the SMC’ model (Marjoram and Wall 2006). Notice that there is a factor of two in the numerator due to the fact that we condition on observing the introgressed allele on a haplotype. We find that in the limiting case of no selection, the approximation closely approximates the analytical result from (Liang and Nielsen, 2014a) (Table 3).

Another case where the logistic function is not expected to be a good approximation of the mean trajectory, is when the admixture proportion is close to zero. To explore this issue, we chose a set of parameters, which are similar to Neanderthal introgression into non-Africans (2.5%), and to the Denisovan introgression into Tibetans (0.06%). We found that the logistic approximation performs poorly for the Denisovan-like scenario, and is not very accurate (up to 10% relative error) for the Neanderthal-like introgression with week

---

**Table 1** The accuracy of the deterministic approximation for the expected tract length under adaptive introgression compared to estimates from simulations. For every set of parameters (introgression fraction, selection coefficient, and time of introgression), we performed $10^5$ replicate simulations. The haploid effective population size was 10,000 chromosomes, with 100 chromosomes sampled from each population. The relative error was calculated by comparing the simulated expected tract length to the prediction given by the deterministic model.

| Introgression parameters | Expected tract length | Simulations | Deterministic approximation | Relative error |
|--------------------------|-----------------------|-------------|-----------------------------|---------------|
| Proportion Selection Time |                        |             |                             |               |
| 0.01 0.001 50            | 0.0417436             | 0.0401907   | 3.7%                        |               |
|                         | 0.0197123             | 0.0200985   | 2.0%                        |               |
|                         | 0.00407317            | 0.00402525  | 1.1%                        |               |
|                         | 0.00206252            | 0.00201644  | 2.2%                        |               |
| 0.01 0.01 50            | 0.0415686             | 0.0402287   | 3.3%                        |               |
|                         | 0.0202129             | 0.02014     | 0.4%                        |               |
|                         | 0.00423004            | 0.00411598  | 2.8%                        |               |
|                         | 0.00237283            | 0.00227902  | 4.0%                        |               |
| 0.05 0.001 50           | 0.0427474             | 0.0419025   | 2.0%                        |               |
|                         | 0.0210069             | 0.0209646   | 0.2%                        |               |
|                         | 0.00428813            | 0.00421551  | 1.7%                        |               |
|                         | 0.00217073            | 0.00212356  | 2.2%                        |               |
| 0.05 0.01 50            | 0.0428456             | 0.0420981   | 1.7%                        |               |
|                         | 0.021438              | 0.0211766   | 1.2%                        |               |
|                         | 0.00470731            | 0.00462923  | 1.7%                        |               |
|                         | 0.00290033            | 0.00286404  | 1.3%                        |               |
selection ($s = 0.001$); in particular, the predicted allele frequency is strongly underestimated. Presumably these discrepancies reflect the increased impacts of genetic drift for small admixture proportions. To address this issue, we estimated the expected trajectory by forward-in-time stochastic simulations of a large sample of possible allele frequency trajectories (see subsection Selected allele trajectory for details). The stochastically estimated expected trajectory is then used instead of the logistic function in Equation 1 to model the tract length distribution. The deterministic model (with numerically pre-computed expected trajectories) gives tract length estimates quite similar to those observed in the simulated data sets (Table 2).

### Variance

The variance in the tract length distribution predicted by our method is an accurate approximation of the across population variance, that is, the variance among all the tracts obtained in many replicates with the same parameters (see Table 4). This is equivalent to the variance in tract length among tracts within a population if the tracts introgressed into the population independently of each other and have not coalesced or recombined with each other. Of course, in reality, two tracts might coalesce before the time of introgression, so their lengths are not independent of each other. Hence, we expect that the within-population variance is smaller than across-population variance. On the other hand, if introgression is recent enough, then the chance of coalescence is small, and the correlation between tract lengths within population will be small.

Indeed, for most of the scenarios with 50 and 100 generations since the time of introgression (Table 4) the variances of the tract lengths are similar. For larger times, the across-populations and within-population variance are different by up to 20%, reflecting the increased probability of coalescence among selected haplotypes that share a recombination event within the admixed population.

### Run time

Calculation of the expected tract length for 100 introgression scenarios takes just 106 sec for a Python implementation executed on MacBook Pro (2.9 GHz Intel Core i5) when using the logistic function to approximate the allele frequency trajectory. Transition rates are

---

**Table 2** The accuracy of the deterministic approximation for the expected tract length under adaptive introgression compared to the estimates from simulations for numerically estimated trajectory. For scenarios with relatively small admixture fractions ($\omega_1 = \text{Proportion} = 0.0006$), the logistic function does not accurately describe the allele frequency trajectory, so we numerically estimated the mean trajectory using stochastic simulations. The relative error is for the deterministic approximation with numerically estimated trajectories relatively to the simulation estimates.

| Proportion | Selection | Time | Simulations | Deterministic Approximation (numerical) | Deterministic Approximation (logistic) | Relative error |
|------------|-----------|------|-------------|--------------------------------------|---------------------------------------|---------------|
| 0.0006     | 0.01      | 1500 | 0.00176     | 0.00173                               | 0.00142                               | 1.7%          |
|            |           | 1750 | 0.00158     | 0.00160                               | 0.00129                               | 1.3%          |
|            |           | 2000 | 0.00150     | 0.001500                              | 0.00120                               | 0.2%          |
|            |           | 2250 | 0.00141     | 0.00142                               | 0.00114                               | 0.7%          |
| 0.0006     | 0.02      | 1500 | 0.00234     | 0.00230                               | 0.00198                               | 1.7%          |
|            |           | 1750 | 0.00222     | 0.00215                               | 0.00186                               | 3.2%          |
|            |           | 2000 | 0.00210     | 0.00203                               | 0.00175                               | 3.3%          |
|            |           | 2250 | 0.00205     | 0.00193                               | 0.00167                               | 5.9%          |
| 0.025      | 0.001     | 1500 | 0.001415    | 0.001436                              | 0.00138                               | 1.5%          |
|            |           | 1750 | 0.001227    | 0.001244                              | 0.001189                              | 1.4%          |
|            |           | 2000 | 0.001136    | 0.0011101                             | 0.001043                              | 3.1%          |
|            |           | 2250 | 0.001020    | 0.0009990                             | 0.000930                              | 2.9%          |
|            |           | 2500 | 0.000883    | 0.000902                              | 0.000839                              | 2.2%          |
| 0.025      | 0.005     | 1500 | 0.00164     | 0.00158                               | 0.00155                               | 4.9%          |
|            |           | 1750 | 0.00146     | 0.00142                               | 0.00139                               | 2.7%          |
|            |           | 2000 | 0.00133     | 0.00130                               | 0.00128                               | 2.3%          |
|            |           | 2250 | 0.00126     | 0.00121                               | 0.00120                               | 4.0%          |
|            |           | 2500 | 0.00118     | 0.00114                               | 0.00114                               | 3.4%          |

**Table 3** Deterministic prediction for the expected tract length in a neutral model compared to the theoretical expectation under SMC’ model.

| Proportion | Time | Expected tract length (deterministic approximation) | Expected tract length (theoretical) | Relative error |
|------------|------|-----------------------------------------------------|-------------------------------------|---------------|
| 0.01       | 10   | 0.20194                                             | 0.200923                            | 0.5%          |
|            | 100  | 0.0201956                                           | 0.0200946                           | 0.5%          |
|            | 1000 | 0.000202122                                         | 0.000201177                         | 0.5%          |
| 0.05       | 10   | 0.210491                                            | 0.209387                            | 0.5%          |
|            | 100  | 0.021054                                            | 0.0209442                           | 0.5%          |
|            | 1000 | 0.000211024                                         | 0.00020999                          | 0.5%          |
| 0.1        | 10   | 0.222252                                            | 0.222278                            | 0.01%         |
|            | 100  | 0.022233                                            | 0.022278                            | 0.2%          |
|            | 1000 | 0.000223105                                         | 0.000227824                         | 2.1%          |
Table 4 Simulated across and within population variance and the across population variance estimated from the deterministic model. The percents show the relative error of the deterministic approximation compared to the simulations both for across and within population values. The introgression with 0.0006 (or, 0.06%) admixture proportion was chosen to approximate Neanderthal introgression into non-Africans. The introgression with 0.025 (or, 2.5%) admixture proportion was chosen to approximate Denisovan introgression into Tibetans.

| Proportion | Selection | Time | Deterministic approximation | Simulated tract length (across populations) | Simulated tract length (within population all replicates) |
|------------|-----------|------|-----------------------------|---------------------------------------------|--------------------------------------------------------|
| 0.01       | 0.001     | 50   | 0.028513                    | 0.029397 3.01% 0.027405 4.04%               |                                                        |
| 0.01       | 0.01      | 50   | 0.028534                    | 0.028719 0.64% 0.027571 3.49%               |                                                        |
| 0.05       | 0.001     | 50   | 0.029728                    | 0.030492 2.51% 0.029590 0.47%               |                                                        |
| 0.0006     | 0.01      | 1500 | 0.001125                    | 0.001173 4.09% 0.008692 16.94%              |                                                        |
| 0.0006     | 0.02      | 1500 | 0.001350                    | 0.001423 5.13% 0.001241 8.78%               |                                                        |
| 0.025      | 0.001     | 1500 | 0.001014                    | 0.001052 3.52% 0.001042 3.49%               |                                                        |
| 0.025      | 0.005     | 1500 | 0.001082                    | 0.001136 4.75% 0.000991 9.18%               |                                                        |

Dependence of mean tract length on different parameters

The deterministic approach facilitates the calculation of expected tract lengths for a wide array of parameters. In Figure 2 we illustrate the dependency of the mean tract length on the admixture fraction given the time of introgression and selection coefficient. As expected, a larger admixture fraction results in longer expected tract lengths. In Figure 3 we illustrate the dependency on the selection coefficient given the time of introgression and admixture proportion. Again, stronger selection results in longer expected ancestry tract lengths. The strength of selection has a stronger effect on the expected tract length than the admixture proportion. For example, consider an adaptive introgression model with time $T = 1000$, selection coefficient $s = 0.01$ and admixture proportion 0.025. The expected tract length is 0.0025 Morgans under this model. If we double the admixture proportion to 0.05, the expected tract length increases to 0.0029 Morgans, or approximately 13%. If on the other hand we increase the strength of selection by a factor of two, then the expected tract length increases by 41% and becomes 0.0036 Morgans.

The new approximation facilitates investigation of another important dependency, which would otherwise require prohibitively large numbers of simulations, namely the dependency of the expected tract length on the observed allele frequency. In Figure 4 we demonstrate that conditioning on the allele frequency at the time of sampling, stronger selection actually reduces the expected tract length. This is explained by the fact that stronger selection decreases (going backward in time) the allele frequency faster than weaker selection. Hence, the probability of recombining back to a haplotype that contains the selected allele following a recombination event is smaller for scenarios with stronger selection. More formally, assume that we compare two scenarios with selection coefficients $s_1$ and $s_2$ (all other parameters are the same), and that $s_1 < s_2$. Let $\omega^{(1)}(t)$ be the logistic trajectory for selection coefficient $s_1$. We condition on the observed allele frequencies at time $t_0$, hence $\omega^{(1)}(t_0) = \omega^{(2)}(t_0)$. Ancestral recombination occurs on a chromosome independent of the strength of selection. Assume that recombination occurred at time $t_r > t_0$. The expected allele frequency at time $t_r$, which is defined by the logistic function, is larger for the smaller selection coefficient:
Denisovan introgression into Tibetans

A well-known example of adaptive introgression is Denisovan introgression into Tibetans which facilitated altitude adaptation by introducing an adaptive allele of EPAS1 affecting red blood cell production into the ancestors of modern Tibetans (Yi et al. 2010; Huerta-Sánchez et al. 2014). We wondered how selection might affect estimates of the time of introgression, conditioning on the present-day allele frequency of EPAS1 of 85% (Huerta-Sánchez et al. 2014) and an overall genomic admixture proportion of $s_1 = 0.06\%$ (Sankararaman et al. 2016).

The admixture proportion is very small, and the logistic function will therefore be quite inaccurate. For a range of values of selection coefficient ($s = 0.005, 0.006, \ldots, 0.02$), we numerically estimated the expected trajectories (using the stochastic trajectory simulator) for the given admixture proportion. Given the frequency of the adapted allele in the modern population, the time since introgression is estimated as the time needed for the allele to reach this observed allele frequency following the expected trajectory.

We summarize the dependency of the introgression time on the strength of selection in Table 5 for the mean value $s_1 = 0.06\%$. It is clear from the table that dating the time of introgression from the tract length under the hypothesis of neutrality, would cause underestimation of the time of introgression by about 20–25%. For example, if the estimated tract length is 0.00117 Morgans, the corresponding introgression time without selection affecting the allele frequency is 1791 generations, or about 45000 years. However, taking selection into account, this tract length corresponds to selection coefficient $s = 0.006$ and an introgression time of 2282 generations, or 57000 years ago. Therefore, analyses that seek to understand the timing of introgression events of adaptively introgressed alleles will be greatly facilitated by incorporating the dependency on the strength of selection on the expected tract length distribution. The best option for providing bounds for the time of introgression of single tracts, if these tracts cannot be assumed to be part of a larger introgression pulse that included other fragments, would be to jointly estimate the introgression time, the selection coefficient, and the time of introgression, using tract length, allele frequency, and other information such as local haplotype homozygosity and/or local genealogical information.

DISCUSSION

Adaptive introgression is an important and common phenomenon in evolutionary genetics (Hedrick 2013). In this paper we derived an accurate approximation model for the ancestry tract length distribution under adaptive introgression, near a selected site, with a single admixture pulse. This framework facilitated calculations of the distribution of tract lengths under a range of plausible adaptive introgression scenarios. However, we emphasize that future work should consider additional demographic models including continuous migration or multiple pulses. We demonstrated, perhaps counter-intuitively, that conditioned on the observed allele frequency, stronger selection produces shorter admixture tracts. Furthermore, as an example, we showed that selection should be taken into account when dating the well-known case of altitude adaptation in Tibetans through the introgression of EPAS1 gene from Denisovans. When ignoring the impacts of selection, the time since introgression might be underestimated by as much as 20–25% if conditioning on the allele frequency. If the allele frequency is allowed to be random, the effect would be opposite.
Our results illustrate that selection should be carefully considered and incorporated when studying adaptive introgression events. More generally, this framework opens new possibilities for understanding the timing of admixture and the strength of adaptive introgression across a wide range of populations. Our work therefore lays the groundwork for the development of new inference frameworks that...
can detect adaptive introgression using tract lengths and allele frequencies. Particularly, in light of the fact that adaptive introgression is common historically, and may be accelerated in the future by anthropogenic impacts such as climate change, e.g., (Muhlfeld et al. 2014), developing a strong theoretical basis for understanding the impacts of adaptive introgression on patterns of genetic variation is crucial for interpreting the genomic consequences of admixture.

**ACKNOWLEDGMENTS**

This research was supported by NIH grant R01GM116044-01 (RN, VS) and a grant from the Danish DFF (RN). PM was supported by NIH R35GM128932 (RC-D). This work was also supported by NIH training grant T32 HG008345. RN, RCD were funded within the framework of the HSE University Basic Research Program. VS was supported by grant RFBR 19-07-00515.

**LITERATURE CITED**

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. 72: 123–133. https://doi.org/10.1017/S0016672398003462

Corbett-Detig, R., and M. Jones, 2016 SELAM: simulation of epistasis and local adaptation during admixture with mate choice. Bioinformatics 32: 3035–3037. https://doi.org/10.1093/bioinformatics/btw365

Corbett-Detig, R., and R. Nielsen, 2017 A hidden markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. PLoS Genet. 13: e1006529. https://doi.org/10.1371/journal.pgen.1006529

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66: 129–138. https://doi.org/10.1016/j.tpb.2004.04.002

Ewens, W. J., 1968 A genetic model having complex linkage behaviour. Theor. Appl. Genet. 38: 140–143. https://doi.org/10.1007/BF00338080

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Feldman, M. W., I. Franklin, and G. J. Thomson, 1974 Selection in complex genetic systems i. the symmetric equilibria of the three-locus symmetric viability model. Genetics 76: 135–162.

Franklin, L., and R. C. Lewontin, 1970 Is the gene the unit of selection? Genetics 65: 707–734.

Gravel, S., 2012 Population genetics models of local ancestry. Genetics 191: 607–619. https://doi.org/10.1534/genetics.112.139808

Hedrick, P. W., 2013 Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol. Ecol. 22: 4606–4618. https://doi.org/10.1111/mec.12415

Hill, W. G., 1974 Disequilibrium among several linked neutral genes in finite population: II. variances and covariances of disequilibria. Theor. Popul. Biol. 6: 184–198. https://doi.org/10.1016/0040-5809(74)90023-9

Hindmarsh, A. C., 1983 ODEPACK, A Systematized Collection of ODE Solvers, pp. 55–64 in Scientific Computing, edited by R. S. Stepleman et al. IMACS Transactions on Scientific Computation, North-Holland, Amsterdam.

Huerta-Sánchez, E., X. Jin, Z. Asan, Z. Biaoba, and B. M. Peter 2014 Altitude adaptation in tibetans caused by introgression of denisovan-like dna. Nature 512: 194–197. https://doi.org/10.1038/nature13408

Hufford, M. B., P. Lubinsky, T. Pyhrvi, M. T. Devyengenzo, N. C. Ellstrand et al., 2013 The genomic signature of crop-wild introgression in maize. PLoS Genet. 9: e1003477. https://doi.org/10.1371/annotation/2eef7b5b-29b2-412f-8472-8fd7f9bd65ab

Kaplan, N. L., R. R. Hudson, and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics 123: 887–899.

Karlin, S., 1975 General two-locus selection models: Some objectives, results and interpretations. Theor. Popul. Biol. 7: 364–398. https://doi.org/10.1016/0040-5809(75)90025-8

Karlin, S., and D. Carmelli, 1975 Numerical studies on two-loci selection models with general viabilities. Theor. Popul. Biol. 7: 399–421. https://doi.org/10.1016/0040-5809(75)90026-X

Karlin, S., and M. W. Feldman, 1970 Linkage and selection: Two locus symmetric viability model. Theor. Popul. Biol. 1: 39–71. https://doi.org/10.1016/0040-5809(70)90041-9

Kermack William Ogilvy, M. A. G., and W. G. Thomas, 1927 A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond., A. Contain. Pap. Math. Phys. Character 115: 700–721. https://doi.org/10.1098/rspa.1927.0118

Kimura, M., 1958 A model of a genetic system which leads to closer linkage by natural selection. Evolution 10: 278–287. https://doi.org/10.1111/j.1558-5646.1956.tb02852.x

Lewontin, R. C., 1964a The interaction of selection and linkage. i. general considerations; heterotic models. Genetics 49: 49–67.
