Title

Estimating the Timing of Multiple Admixture Pulses During Local Ancestry Inference

Journal

Genetics, 210(3)

ISSN

0016-6731

Authors

Medina, Paloma

Thornlow, Bryan

Nielsen, Rasmus

et al.

Publication Date

2018-11-01

DOI

10.1534/genetics.118.301411

Estimating the Timing of Multiple Admixture Pulses During Local Ancestry Inference

Paloma Medina,* Bryan Thornlow,* Rasmus Nielsen,† and Russell Corbett-Detig*,1
*Department of Biomolecular Engineering, Genomics Institute, UC Santa Cruz, California 95064 and †Departments of Integrative Biology and Statistics and Museum of Natural History, University of Copenhagen, 2100 Denmark

ABSTRACT Admixture, the mixing of genetically distinct populations, is increasingly recognized as a fundamental biological process. One major goal of admixture analyses is to estimate the timing of admixture events. Whereas most methods today can only detect the most recent admixture event, here, we present coalescent theory and associated software that can be used to estimate the timing of multiple admixture events in an admixed population. We extensively validate this approach and evaluate the conditions under which it can successfully distinguish one- from two-pulse admixture models. We apply our approach to real and simulated data of *Drosophila melanogaster*. We find evidence of a single very recent pulse of cosmopolitan ancestry contributing to African populations, as well as evidence for more ancient admixture among genetically differentiated populations in sub-Saharan Africa. These results suggest our method can quantify complex admixture histories involving genetic material introduced by multiple discrete admixture pulses. The new method facilitates the exploration of admixture and its contribution to adaptation, ecological divergence, and speciation.

KEYWORDS admixture; *Drosophila melanogaster*; local ancestry inference

THERE is an increasing appreciation for the importance of admixture, an evolutionary process wherein genetically divergent populations encounter each other and hybridize. Admixture has shaped genetic variation within natural plant, animal, and human populations (Rieseberg *et al.* 2003; Pool *et al.* 2012; Hufford *et al.* 2013; Hellenthal *et al.* 2014; Sankararaman *et al.* 2014). If an admixture event has occurred relatively recently, we can use local ancestry inference methods (LAI) to trace the ancestry of discrete genomic segments, called "ancestry tracts," back to the ancestral populations from which they are derived (Pool and Nielsen 2009; Price *et al.* 2009; Sankararaman *et al.* 2012; Maples *et al.* 2013; Corbett-Detig and Nielsen 2017). Due to ongoing recombination within admixed populations, the lengths of these ancestry tracts are expected to be inversely related to the timing of admixture. Therefore, it is possible to estimate the timing of admixture events by inferring ancestry tract lengths (Pool and Nielsen 2009; Gravel 2012), or by evaluating the rate of decay of linkage disequilibrium (LD) among ancestry informative alleles (Moorjani *et al.* 2011; Loh *et al.* 2013).

The latter approach is based on modeling expected decay of LD among alleles that are differentiated between admixed populations. Briefly, even if there is little LD in the ancestral populations themselves, admixture will create admixture LD (ALD) within the admixed population among alleles whose frequencies are differentiated between ancestral populations (Chakraborty and Weiss 1988). The decay of ALD is expected to be approximately exponential, with a rate parameter that is proportional to the timing of admixture. Two popular methods using this approach, ROLLOFF (Moorjani *et al.* 2011) and ALDER (Loh *et al.* 2013), model the decay of two-locus ALD to estimate single-pulse admixture models. Admixture histories may be more complex, including multiple distinct admixture events (Gravel *et al.* 2013) and multiple ancestral populations (Paşaniuc *et al.* 2009), suggesting that some of these methods may not be suitable for deeply characterizing the admixture history of many admixed populations (Figure 1), although recent work suggests it is possible to detect multiple admixture pulses by modeling LD decay as a mixture of two exponential distributions (Pickrell *et al.* 2014). Additionally, it may be possible to extend LD decay approaches using three-point LD to estimate the timing of two admixture pulses in populations with more

doi: https://doi.org/10.1534/genetics.118.301411
Manuscript received July 24, 2018; accepted for publication August 31, 2018; published Early Online September 11, 2018.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7070108.
[1] Corresponding author: UC Santa Cruz, 1156 High St., PBSci 316, Santa Cruz, CA 95064. E-mail: rucorbet@ucsc.edu

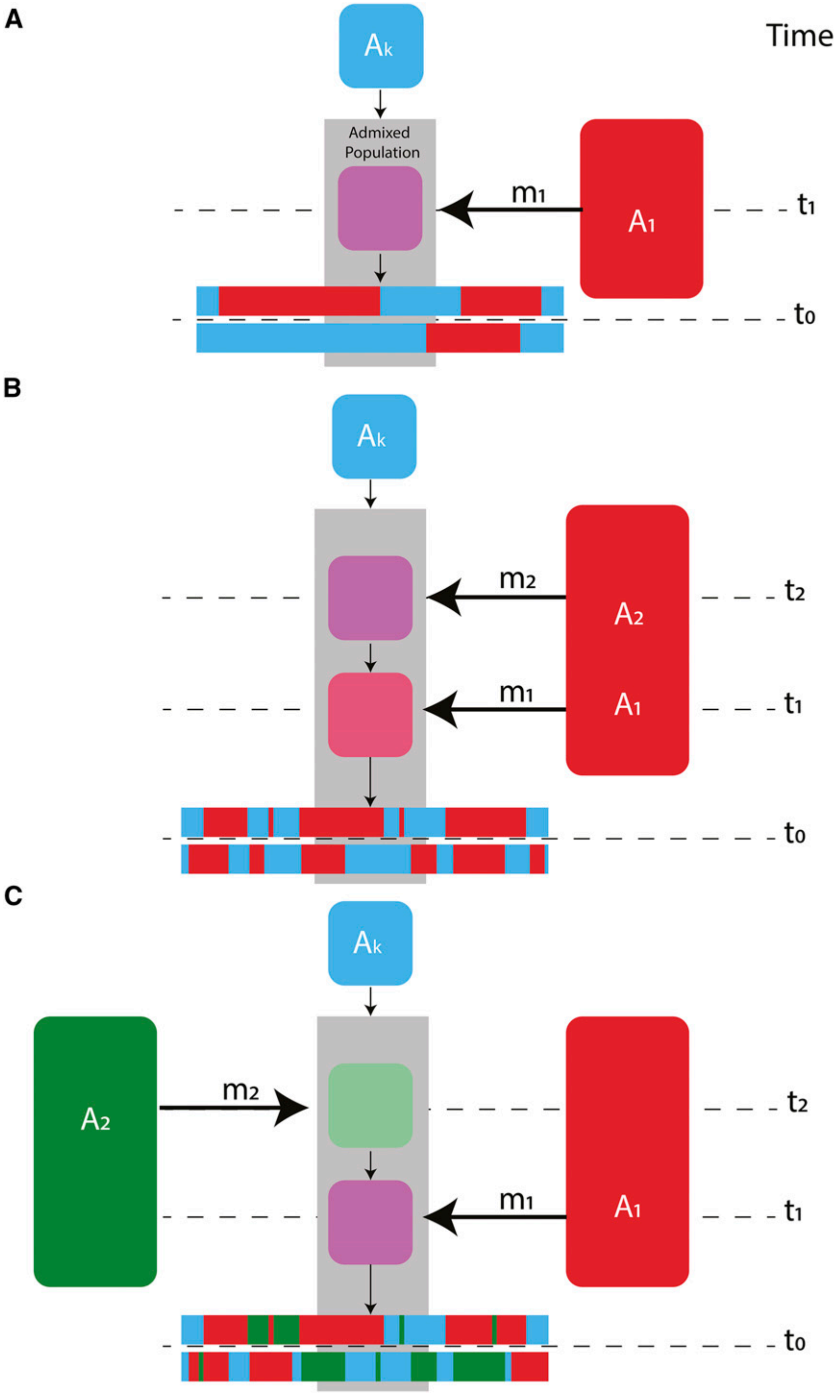

Figure 1 A schematic of (A) a single-pulse model with two ancestry types, (B) a two-pulse model with two ancestral populations, and (C) a two-pulse model with three ancestral populations. A_k is considered the ancestry type of the resident population and would be A_2 in (A) and A_3 in (B) and (C). Note that A_1 and A_2 may come from the same ancestral population, but are modeled as independent states. The gray shaded region draws attention to the admixed population(s). Time since admixture pulses are measured in generations and are denoted as t_1 and t_2, where t_1 occurs more recently than t_2. The time of sampling is represented by t_0, where $t_0 = 0$ if sampling occurred in the present. The proportion of ancestry in the admixed population that entered during an admixture pulse is denoted as m_1 and m_2. Colors represent genetically distinct ancestry types. Local ancestry across a chromosome after admixture is represented by horizontal bars at the bottom of each subplot.

complex admixture histories (Liang and Nielsen 2016). Finally, wavelet-based techniques can be used to infer the relationship between SNPs along a chromosome to estimate the time of admixture and admixture proportions (Sanderson *et al.* 2015; Pugach *et al.* 2016).

A second set of approaches uses the lengths of LA tracts across the genomes of admixed individuals to estimate the timing of admixture events (Pool and Nielsen 2009; Gravel 2012; Gravel *et al.* 2013; Ni *et al.* 2018). Here, the complexity of admixture models that can be accommodated depends on the accuracy of the estimated ancestry tract length distributions within an admixed population. However, these methods require that LAI be performed prior to estimating population admixture models. LAI necessitates *a priori* assumptions about the admixture model itself, as such assumptions are a prerequisite—implicitly or explicitly—for most LAI frameworks that have been developed to date. These assumptions have the potential to bias the outcomes of LAI and can affect admixture model selection (Sankararaman *et al.* 2008; Pool *et al.* 2012; Corbett-Detig and Nielsen 2017). Additionally, LAI methods require accurately phased chromosomes—an unlikely prospect for species with low levels of LD (Bukowicki *et al.* 2016) and for relatively ancient admixture events where the rate of phasing switch errors would be similar to the rate of transitions between ancestry types. Therefore, it is often preferable to estimate the timing of admixture events and the LA of a sample simultaneously (Corbett-Detig and Nielsen 2017). Finally, modern sequencing techniques often sequence to light coverage, thus necessitating a tool that can accommodate read pileup data rather than genotypes.

Prior to this study, we developed a framework for simultaneously estimating admixture times and LA across the genomes of admixed populations (Corbett-Detig and Nielsen 2017). Our method can estimate LA and admixture times in a hidden Markov model (HMM) framework assuming a single admixture pulse model. Here, we extend this method to admixed populations that have experienced multiple ancestry pulses in their recent evolutionary past (Figure 1). We evaluate the performance of this approach using extensive forward-in-time admixture simulations. Finally, we apply this method to study admixed *Drosophila melanogaster* populations in sub-Saharan Africa, and we find evidence consistent with a recent single pulse of cosmopolitan admixture into African populations as well as evidence for more ancient admixture among genetically differentiated populations in sub-Saharan Africa.

Methods

Coalescent simulations

We simulated ancestral population genetic variation using the coalescent simulation software package MACS (Chen *et al.* 2009). Briefly, to consider a simplified ancestral process, we consider a model where a population, initially of size 2N, subdivided into three daughter populations of identical size, 2N. To assay a range of genetic divergence among the daughter populations, we performed replicate simulations where we allowed populations to diverge (D in the command line below) for 0.05, 0.1, 0.25, 0.5, and 1 N_e generations. Note that the functionally relevant parameter for LAI using this approach is probably related to allele frequency changes between the ancestral populations, and therefore to the density of ancestry informative markers along the genome, rather than to the timing of divergence. For example, a population bottleneck would result in rapid and dramatic allele frequency differences despite a short divergence time. Note also that alternative divergence models will also impact the distribution of LD among markers, underscoring the importance of considering each unique model when applying this approach for admixed model estimation. We therefore used the following command line for all simulated populations:

```
$ macs 600 100000000 -i 1 -h 100 -t 0.001 -r 0.001 -I 3 200
    200 200 0 -ej D 2 1 -ej D 3 1.
```

This will simulate a single chromosome of length 100 Mb with a per-site theta of 0.001 and equal recombination rate. Therefore, these simulations could resemble a reasonably sized mammalian population. This will output 200 chromosomes per ancestral population, and for all admixture simulations, we used the first 100 as the reference panel, and the second 100 to simulate genetic variation across admixed chromosomes.

Admixture simulations

We simulated admixed populations using the forward-in-time admixture simulation program, SELAM (Corbett-Detig and Jones 2016). We first explored the levels of LD pruning that are necessary for producing unbiased estimates of admixture times using two-population, single-pulse simulations. All simulations for LD pruning optimizations were run with ancestry proportions 0.5 and 0.5, and with an admixed population size of 5000 males and 5000 females.

We then simulated admixture models with three ancestral population models where the resident population contributed 1/2 of the total ancestry, and each admixture pulse contributed 1/4th of the total ancestry to the final admixed population. Therefore the first pulse, in forward time, contributed 1/3 of the admixed population and then had 1/4th of that ancestry replaced by the subsequent pulse. All simulations were performed across the range of divergence times that we simulated for each ancestral population.

Read simulation

We simulated short read data for each admixed individual following the approach of Corbett-Detig and Nielsen (2017). Briefly, reads are drawn binomially from each samples' genotype and read depths from a Poisson distribution with mean equal to 2. That is, we simulated 2× sequencing depths to represent a likely use case of this software where individuals are sequenced to relatively light coverage.

Drosophila simulation

We have previously used this framework to simulate data consistent with the ancestral populations of *D. melanogaster*, following the coalescent simulation approach of Corbett-Detig and Nielsen (2017). All other features of the simulated *Drosophila* dataset are similar to those above, except that we simulated genotypes rather than short read pileup data. We did this because the dataset that we used is sequenced to sufficiently high depths so as to preclude most uncertainty in short read data (Lack *et al.* 2015, 2016).

Robustness to smaller sample sizes, incorrect recombination rates, and ancestral population divergence

We simulated varying sample sizes, erroneous recombination maps and reference population divergences to characterize our method's ability to correctly estimate the timing of admixture. We simulated single admixture pulses 20, 40, 60, and 80 generations ago using sample sizes of 10, 25, and 50.

We performed 20 replicate simulations. For all robustness checks, we used simulated data where the ancestral populations were 0.25 N_e generations divergent from one another.

To test for the effect of erroneous recombination maps, we modified the recombination map every 5 Mb using the following approach. For each window, we do a draw of D from uniform $(1-d,\ 1+d)$ where d is 0.25, 0.5, 0.75, or 1. We multiply D by the actual recombination rate at that window to obtain our modified recombination rate. For example, where d is 1, D is chosen from a uniform distribution of range (0,2). Where d is 0.25, D is chosen from a uniform distribution of range (0.75,1.25). Thus, the larger the value of d, the more the recombination map will be perturbed.

We also tested our approach under scenarios where one of the reference populations used to estimate the timing of admixture is not the actual source population contributing to the admixed population. We estimated the timing of single and double pulse admixture events using a reference population that was 0.005, 0.01, 0.02, 0.05, and 0.1 N_e generations diverged from the real source population. We simulated single pulse data where the timing of admixture occurred 20, 40, 60, and 80 generations ago. We fit both single and double pulse admixture models to these data, and computed summary statistics describing the fit of single pulse model to single pulse data. In addition, we simulated double pulse data where the timing of the most recent admixture pulse occurred 20, 40, 60, and 80 generations ago, and where the most distant admixture pulse always occurred 100 generations ago. We computed summary statistics describing the fit of two pulse models to these data.

Comparison to ALDER

We estimated the timing of single- and double-pulse admixtures using ALDER v1.02 (Loh *et al.* 2013) and MALDER v1.0 (Pickrell *et al.* 2014), respectively. We evaluated the results of our approach, ALDER, and MALDER by performing 20 replicates of each admixture scenario, and computing the NRMSE for estimated admixture time and admixture proportion. We simulated single pulse admixture data using two reference populations of 50 individuals each and an admixed population of 50 individuals. Similarly, we simulated double pulse admixture data by adding a third reference population of 50 individuals. The timings of the single- and double-pulses are identical to those outlined above, and can be found explicitly in Supplemental Material, Tables S13 and S14. ALDER and MALDER were run with a maxdis of 0.1. We note that the default settings of ALDER and MALDER could not detect admixture events occurring >100 generations ago. Thus, we extended the minimum distance (mindis) to 0 for any scenario with an admixture event occurring >100 generations ago. Extending the minimum distance to 0 may confound estimated admixture times because ancestral LD will have a greater impact on short scales, but appeared to produce better admixture time estimates for the scenarios we considered. We note that there may be other ways to improve ALDER and MALDER's performance in detecting ancient admixture events.

Drosophila sample application

We obtained *D. melanogaster* genotype data for natural isolates from the *Drosophila* Genome Nexus (Lack *et al.* 2015, 2016), which curates data from >1000 natural isolates of this species. We used the recommended masking packages supplied for that site for all samples in reference populations. For SD, we used the ZI population, for RG we used a set of Central and West African samples (as in Corbett-Detig and Nielsen 2017), and for EA we used the EF population. All cosmopolitan pulses were modeled using FR as the cosmopolitan reference population.

Application to admixed samples

We used our model to assay the admixture histories of Rwandan, South African, and Ethiopian samples. We fit both a single-pulse model and double-pulse model to genotype data from each population, running 100 bootstrap replicates using a block size of 5000 SNPs.

```
- single pulse.
./ancestry_hmm -i RG.auto.panel -s RG.ploidy.txt -a 2 0.1
   0.9 -p 1 100000 0.9 -p 0 -100 0.1 -g -e 1e-4 -b 100 5000.
./ancestry_hmm -i SD.auto.panel -s SD.ploidy.txt -a 2 0.17
   0.83 -p 1 100000 0.83 -p 0 -100 0.17 -g -e 1e-4 -b
   100 5000.
./ancestry_hmm -i EA.auto.panel -s EA.ploidy.fixed.txt -a
   2 0.34 0.66 -p 1 100000 0.66 -p 0 -100 0.34 -g -e 1e-4
   -b 100 5000.
- double pulse.
./ancestry_hmm -i RG.auto.panel -s RG.ploidy.txt -a 2 0.1
   0.9 -p 1 100000 0.9 -p 0 -100 -0.05 -p 0 -100 -0.05 -e
   1e-4 -g -b 100 5000.
./ancestry_hmm -i SD.auto.panel -s SD.ploidy.txt -a 2 0.17
   0.83 -p 1 100000 0.83 -p 0 -100 -0.085 -p 0 -100 -0.085
   -e 1e-4 -g -b 100 5000.
./ancestry_hmm -i EA.auto.panel -s EA.ploidy.txt -a 2 0.34
   0.66 -p 1 100000 0.66 -p 0 -100 -0.17 -p 0 -100 -0.17 -e
   1e-4 -g -b 100 5000.
```

Application to three population mixture from Gambella, Ethiopia

We used our model to assay the admixture histories of Gambella, Ethiopia. We fit a double pulse model to genotype data from Gambella, running 1000 bootstrap replicates each using a block size of 5000 SNPs. Moreover, since the native ancestry type for Gambella is unknown, we modeled West African (AF), Ethiopian (EA), and French (FR) as the ancestral population.

```
- EA native ancestry.
./ancestry_hmm -i three_pop.auto.EA.panel -s three_pop.
   auto.EA.ploidy -a 3 0.4 0.31 0.29 -p 0 -1000 0.4 -p
   1 100000 0.31 -p 2 -1000 0.29 -g -b 1000 5000.
```

Data availability

We implemented the following model into our software package called *Ancestry_HMM* (www.github.com/russcd/

Ancestry_HMM). Below, wherever possible, we give the script name and line number responsible for a computation, denoted as *header* : *line_number*. All code is in the *src*/ directory within the *Ancestry_HMM* repository. All line numbers and results reported here are based on version 0.94. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7070108.

Model

Overview

In our previous work, we developed an HMM approach to simultaneously estimate local ancestry and admixture times using next generation sequencing data in samples of arbitrary sample ploidy. No phasing is necessary, and, unlike most LAI methods, our models read pileups, rather than genotypes, making this approach appropriate for low-coverage sequencing data (Skotte *et al.* 2013). Additionally, although not addressed in this work, our method accommodates samples of arbitrary ploidy, making it ideal for poolseq applications or for populations with unusual ploidy, *e.g.*, tetraploids. Our previous work assumed a single exponential tract length distribution, and is therefore limited to accommodating only a single admixture event between two populations (Corbett-Detig and Nielsen 2017). Here, we seek to extend this framework to accommodate additional admixture pulses either from distinct ancestry types or multiple pulses from the same type. Wherever possible, we have kept our notation identical to our previous work to facilitate comparisons between the models. Please refer to Table S1 for descriptions of the statistics we used to describe admixture model and LAI accuracy.

State space

Our model incorporates the ancestry of samples with arbitrary ploidy of n chromosomes and with k distinct ancestry types resulting from $k-1$ admixture pulses. Therefore the state space S, is defined as the set of all possible k-tuples of non-negative integers, $H=(l_1,\ldots,l_k)$, such that $\sum_{i=1}^{k} l_i = n$, where l_j is the number of chromosomes in the sample from admixture pulse j.

Emission probabilities

We model the probability of read pileup data (or alternatively genotypes) of each sample as a function of the allele counts in each ancestral population and assume a uniform prior on the allele frequencies in each source population. Note that where the same ancestral population contributes multiple pulses, the emission probabilities for each site at these states are identical and computed based on the sum of the number of chromosomes of each ancestry type. To accommodate up to k distinct ancestry types, we use multinomial read sampling probabilities. Specifically, if the representation in the read data are exactly equal (in expectation) for each chromosome, the probability of sampling a given read from chromosomes of ancestry pulse k is l_k/n. Therefore, the probability of any given vector of read counts, $R=(r_1,\ldots,r_k)$, sampled from a site with depth r and across the chromosomes in a given hidden state $H \in S$, assuming no mapping or sequencing biases is ($read\_emissions.h$: 31)

$$R|H,n,r \sim Mult\left(r, \pi = \left(\frac{l_1}{n},\ldots,\frac{l_k}{n}\right)\right) \tag{1}$$

Conditional on the read count vector, R, the number of reads carrying the A allele (assuming an A/a di-allelic locus) is independent among ancestry pulses.

For each reference population, the allele count is binomially distributed given the (unknown) true allele frequencies, f_j, $j=1\ldots k$. Let C_j represent the total allele count for reference population j and let C_{jA} represent the total number of A alleles for reference population j. Then ($read\_emissions.h$: 51),

$$C_{jA}|C_j,f_j \sim Bin(C_j,f_j) \tag{2}$$

While we assume genotypic data for the reference populations, we assume short-read pileup data for the admixed population. The (unobserved) allele counts in the admixed chromosomes, stratified by admixture pulse origin, $C_{M1A}, C_{M2A},\ldots C_{MkA}$ and $C_{M1a}, C_{M2a},\ldots C_{Mka}$ are also binomially distributed, *i.e.*, for each ancestry type ($read\_emissions.h$: 47),

$$C_{MjA}|H,f_j \sim Bin(l_j,f_j) \tag{3}$$

Then, assuming a symmetrical and identical error rate among alleles, ϵ, the probability of obtaining r_j reads of allele A from within the r_j reads derived from chromosomes of type j is ($read\_emissions.h$: 47),

$$r_{j_A}|H,r_j,C_{MjA},\epsilon \sim Bin\left(r_j,(1-\epsilon)\frac{C_{MjA}}{l_j}+\epsilon\left(1-\frac{C_{MjA}}{l_j}\right)\right) \tag{4}$$

By integrating across all possible allele frequencies in the ancestral population assuming a uniform $(0,1)$ prior, we obtain the probability of a given number of reads of allele A, r_{jA}

$$\begin{aligned} Pr(r_{jA}|r_j,H,\epsilon) = \sum_{k=0}^{l_k} Pr(r_{jA}|H,r_j,C_{MjA}=k,\epsilon)\int_0^1 Pr(C_{MjA} \\ = k|H,f_j)df_j \end{aligned} \tag{5}$$

We then find all possible ways of arranging r_A reads of allele A across the read vector R, $R_A=\{(r_{1A},\ldots,r_{kA})\}$ Here, for each ancestry type, we require that $0 \le r_{jA} \le r_j$ and $\sum_{j=1}^{k} r_{jA} = r_A$. The probability of a given configuration is ($distribute\_alleles.h$: 31),

$$Pr(R_A = R_A^*|R) = \frac{\prod_{i=j}^{k}\binom{r_j^*}{r_{jA}^*}}{\sum_{R_A\in\Omega_R}\prod_{i=j}^{k}\binom{r_j}{r_{jA}}} \tag{6}$$

where Ω_R is the set of all configurations of R_A compatible with R. For each configuration of reads, R and R_A, these above probabilities for each ancestral population combine multiplicatively across all ancestral populations, and we are then able to obtain the emissions probabilities for the hidden state, $H = i$, as

$$\begin{aligned} &Pr(X, C_{1A}, \ldots, C_{kA}|H, \epsilon, n, C_1, \ldots, C_k) \\ &= \left(\prod_{j=1}^{k} Pr(C_{jA}|C_j)\right) \sum_{\{R,R_A\}\in\Omega_x} Pr(R_A|R)Pr(R|H,n,r) \\ &\times \prod_{j=1}^{k} Pr(r_{jA}|r_j, n, H, \epsilon) \end{aligned} \tag{7}$$

where X is all the read data for the admixed population for the site and Ω_x indicates the set of all values of $\{R_A, R\}$ compatible with X.

Transition probabilities

Transition probabilities must also be significantly expanded to accommodate the more complicated ancestry models investigated in this work.

Modeling multiple pulses into a single recipient population

First, we consider a scenario where two ancestry pulses enter the same admixed populations at two distinct times (Figure 1c). Here, we will refer to the time of each ancestry pulse as t_k, where $t_0 = 0$ and refers to the present and t_1, for example, refers to the most recent admixture pulse in backward time. During each pulse, a proportion of the resident population, m_k, is replaced. In this model, there are two distinct epochs during which a recombination event may occur along a chromosome. The last epoch, in backward time, occurs during the time interval between t_1 and t_2. During this time, there are two ancestry types present, and the transition rate between them is identical to that in our previous work (Corbett-Detig and Nielsen 2017). Specifically, the transition rate between the resident ancestry type A_3 and ancestry type A_2 is

$$2N\left(1 - e^{\frac{t_1 - t_2}{2N}}\right)m_2 \tag{8}$$

in units of Morgans per segment (Liang and Nielsen 2014). A nearly identical relationship holds for the transition rate from ancestry type A_2 to A_3,

$$2N\left(1 - e^{\frac{t_1 - t_2}{2N}}\right)(1 - m_2) \tag{9}$$

However, after the second ancestry pulse in forward time, additional transitions between ancestry types A_3 and A_2 will occur. During this interval, the transition rate from ancestry type A_3 to ancestry type A_2 is

$$2N\left(1 - e^{\frac{-t_1}{2N}}\right)e^{\frac{t_1 - t_2}{2N}}(1 - m_1)m_2 \tag{10}$$

This transition rate reflects the chance of a recombination event occurring between t_1 and the time of sampling, with no back coalescence to the previous marginal genealogy in either time epoch. Finally, $1 - m_1$ is the probability that this recombination event does not choose a lineage that entered the population during the second ancestry pulse in forward time, and m_2 is the probability of recombining with a lineage from the first ancestry pulse in forward time. Hence, the total rate of transition between ancestry type A_3 and ancestry type A_2 across both epochs, is

$$\lambda_{32} = 2N\left(1 - e^{\frac{t_1 - t_2}{2N}}\right)m_2 + 2N\left(1 - e^{\frac{-t_1}{2N}}\right)e^{\frac{t_1 - t_2}{2N}}(1 - m_1)m_2 \tag{11}$$

Again, a similar rate holds for transitions between ancestry types A_2 and A_3. Transition rates associated with the second ancestry pulse are simpler, and closely resemble those for a single pulse model. Specifically, the rate of transition from ancestry type A_3 to ancestry type A_1 is

$$\lambda_{31} = 2N\left(1 - e^{\frac{-t_1}{2N}}\right)m_1 \tag{12}$$

Here, we include the probability that a recombination event occurs within the second epoch and that the lineage selects ancestry type A_1. Note that if the lineage selects ancestry type A_1, we need not consider the probability of back coalescence in the time interval between t_1 and t_2 as it must exit the admixed population during the first ancestry pulse. Similarly, the transition rate from ancestry type A_1 to ancestry type A_3 is

$$\lambda_{13} = 2N\left(1 - e^{\frac{-t_1}{2N}}\right)(1 - m_1)(1 - m_2) \tag{13}$$

where this equation reflects the probability that a recombination event occurs during the second epoch and then the lineage selects ancestry type A_3. Note that no consideration is given to back coalescence to the previous marginal genealogy during the epoch between t_2 and t_1 because there is no ancestry type A_1 present in the population during this time. Here again, similar rates hold for transitions from ancestry type A_1 to ancestry type A_2, and from ancestry type A_2 to ancestry type A_1 following similar logic presented above.

Transition rates can be generalized to include an arbitrarily large number of distinct ancestry pulses. Briefly, transitions between ancestry types may occur during any epoch in which they are both present within the admixed population. For example, for the first pulse in forward time, transitions between ancestry type A_k and A_{k-1} may occur anytime between t_{k-1} and the present. Therefore, all epochs will be necessarily included the transition calculations between ancestry type A_k

and A_{k-1}. More generally, the transition rate between ancestry pulses, i and j, where $1 \leq i < j \leq k$ is,

$$\begin{aligned}\lambda_{ij} = 2Nm_j\left(\prod_{l=i}^{j-1} 1-m_l\right)\Bigg(1-e^{\frac{t_{i-1}-t_i}{2N}}+\sum_{q=1}^{i-1}\Bigg(\bigg(1-e^{\frac{t_{q-1}-t_q}{2N}}\bigg)\\ \times\prod_{p=q+1}^{i} e^{\frac{t_{p-1}-t_p}{2N}}(1-m_{p-1})\Bigg)\Bigg)\end{aligned} \tag{14}$$

Conversely, if $1 \leq j < i \leq k$, the transition rate between ancestry of pulse i to pulse j is very similar, *i.e.*, transitions may occur during the same epochs (t_j through the present). Therefore,

$$\begin{aligned}\lambda_{ij} = 2Nm_j\Bigg(1-e^{\frac{t_{j-1}-t_j}{2N}}+\sum_{q=1}^{j-1}\Bigg(\bigg(1-e^{\frac{t_{q-1}-t_q}{2N}}\bigg)\\ \times\prod_{p=q+1}^{j} e^{\frac{t_{p-1}-t_p}{2N}}(1-m_{p+1})\Bigg)\Bigg)\end{aligned} \tag{15}$$

Note that it is not necessary for each ancestry pulse to be derived from distinct ancestral populations. Indeed, we conceived of this approach as a means of fitting multiple pulse ancestry models. Therefore any number of pulses may be contributed by as few as two ancestral populations. However, in order to model full transition rates, it is necessary to estimate the proportion contributed by each pulse even when they are from the same ancestry type, potentially making multiple pulse ancestry models more challenging to fit than models where each pulse is contributed by a separate subpopulation. More broadly, while this model is quite generalizable, there are limits to what is practical to infer using real datasets (see below).

Transition rates per basepair

Equation 14 and Equation 15 model transitions in Morgans per segment between ancestry states. We must therefore convert these expressions into a transition rate per basepair. To do this, we multiply the recombination rate by Morgans/bp using an estimate of the local recombination rate within that segment of the genome, r_{bp}. Therefore, the single chromosome transition matrix, $P(l) = P_{ij}(l), i,j \in S$, for a two pulse population model for two markers at distance l bp from one another would be as follows (*create_transition_rates.h* 69).

$$P(l) = \begin{bmatrix} 1-(\lambda_{12}+\lambda_{13})r_{bp} & \lambda_{12}r_{bp} & \lambda_{13}r_{bp} \\ \lambda_{21}r_{bp} & 1-(\lambda_{21}+\lambda_{23})r_{bp} & \lambda_{23}r_{bp} \\ \lambda_{31}r_{bp} & \lambda_{32}r_{bp} & 1-(\lambda_{31}+\lambda_{32})r_{bp} \end{bmatrix}^{l} \tag{16}$$

Modeling sample ploidy

The above model describes the ancestry transitions along a single chromosome. However, many datasets contain samples that are diploid or pooled rather than haploid, or equivalently completely inbred. For simplicity, we model each sample of ploidy n as the union of n independent admixed chromosomes. To approximate the transition probability from state i to state j in a sample of n chromosomes, we assume the ancestry proportion, m, contributed by ancestry type A are known. Let the current hidden state ancestry vector be $s = \{s_1, s_2, \ldots, s_k\}$ where k is the maximal number of ancestry types and s_i indicates the number of chromosomes with ancestry component i. Then $t = \sum_{i=1}^{k} s_i$ is the total number of different chromosomes in the pool. Furthermore, let $P_{ij}(d)$ be the d-step transition probability from ancestry i to j of the previously defined 1-chromosome process. Define q_{ij} as $q_{ij} = s_i P_{ij}(d)$. Also, let $a = \{a_1, a_2, \ldots, a_k\}$, $t = \sum_{i=1}^{k} a_i$, be the ancestry vector d sites downstream from the location of the locus with ancestry vector s. Then, we will approximate the transition probability from, s to a, as

$$p(a|s) = \sum_{z \in Z} \prod_{i=1}^{k} \binom{s_i}{z_{i1}, z_{i2}, \ldots, z_{ik}} \prod_j q_{ij}^{z_{ij}}$$

where $z = \{z_{ij}\}$ is an $k \times k$ matrix of non-negative integers, and Z is the set of all such matrices for which $\sum_{i=1}^{k} z_{ij} = a_j \forall a_j \in a$ and $\sum_{i=1}^{k} z_{ji} = s_i \forall s_i \in s$. The sum over all $z \in Z$ can be large, and increases exponentially in k.

The true ancestral recombination graph is potentially more complex than our simple approximation which assumes all ancestry transitions to be independent. Therefore, caution is warranted when using our model on samples with higher ploidy (Corbett-Detig and Nielsen 2017) (*transition_information.h* : 21).

Model optimization

Because the search space for admixture models is potentially quite complex, we have implemented optimization using the Nelder-Mead direct search simplex algorithm (Nelder and Mead 1965) to optimize the HMM using the forward-equation to compute model likelihoods. As with all direct search algorithms, there is no guarantee that the optimum discovered is a globally optimal solution. We therefore include random restarts to insure that the globally optimal solution can be recovered consistently.

Assumptions

Perhaps the central assumption of this approach is that admixture occurs in discrete, distinct "pulses." Whereas this is violated in a wide array of true admixture events with either ongoing or periodic admixture (Pool and Nielsen 2009; Gravel 2012), the pulse model is tractable for estimating admixture histories (Gravel 2012; Loh *et al.* 2013; Corbett-Detig and Nielsen 2017). We have previously shown that LAI using our method is robust to a wide array of perturbations including continuous migration and natural selection (Corbett-Detig and Nielsen 2017). Nonetheless, all results, particularly those that hinge heavily on the precise timing of admixture events, should be interpreted cautiously. See also our discussion below.

Validation

Confirmation of the ancestry tract length approximation

We first confirmed that our sequential Markov coalescent (SMC') approximation for the ancestry tract length distribution is correct. Specifically, we simulated tract length distributions from the forward-in-time simulation program, SELAM (Corbett-Detig and Jones 2016), and compared those with the expected tract length distribution under our model. In comparing the two, we found that the model provides an excellent approximation for the ancestry tract length distribution (Figure 2). This therefore indicates that our SMC' tract length approximation is likely to be sufficient for our purposes. Moreover, we also confirmed that this framework can also accurately accommodate models involving more ancestral populations (Figure S1 and Table S2).

Simulating ancestral populations

We used the approach of Corbett-Detig and Nielsen (2017) to validate our HMM software implementation and to test the performance of this expanded model. Briefly, we simulated ancestral genotype data using the coalescent simulation framework MACS (Chen *et al.* 2009), we then simulated ancestry tract length distributions in forward-time using our software package SELAM (Corbett-Detig and Jones 2016). Genotype data for admixed individuals was then drawn from the ancestral data simulated using MACS where genotypes were drawn from each population and assigned to each tract based on that tracts' ancestry type. To explore a wide range of genetic divergences among ancestral populations, we simulated genotype data for ancestral populations at varying levels of genetic divergence from one another. Specifically, we considered populations that are 0.05, 0.1, 0.25, 0.5, and 1 N_e generations divergent from one another.

LD pruning

LD among sites may inflate transition rates between ancestry states. In order to mitigate this, we first pruned sites in strong LD in each reference panel by computing LD among all pairs of markers within 0.1 cM of each other. We then discarded one site from each pair with increasingly stringent pruning until we found that admixture time estimates were approximately unbiased in two pulse admixture models. We note that the LD pruning necessary for two pulse models appears to be sufficient for fitting more complex models of admixture (see below). This suggests that simple, single-pulse, admixture simulations could be used successfully to determine the necessary levels of LD pruning. Also, we found that substantially more stringent LD pruning is necessary to produce unbiased admixture time estimates for ancestral populations that are minimally divergent from one another (Figure S2).

Three ancestral populations

We next sought to evaluate the accuracy of this method when two ancestry pulses are contributed from distinct ancestral populations. For all levels of population divergence considered, we find that our method recovers the correct times of admixture events with reasonably high accuracy for recent admixture times (Figure 3). Notably, the error in time estimates decrease substantially with increasingly divergent ancestral populations, indicating that this method will produce the most accurate results when ancestral populations are highly genetically divergent.

Whereas our method accurately estimates admixture times when admixture events occur at intermediate times prior to sampling, we find that the accuracy suffers somewhat when estimating the timings of two more ancient admixture events. Indeed, for modestly divergent ancestral populations, *e.g.*, $0.1N_e$ generations divergent, we find that our approach consistently underestimates the times since admixture events, particularly for the more ancient ancestry pulse (Figure 3). This is true despite a nearly linear relationship for estimated admixture times and actual admixture times in a single-pulse two-ancestral-population model (Figure S2), indicating that there is a significant cost for accuracy of model estimation when estimating additional ancestry pulse parameters. However, for more divergent ancestral populations, it is still feasible to obtain accurate admixture time estimates (Figure 3). We note that our good performance in estimating old admixture times may be a function of more informative sites and less LD present in *D. melanogaster* compared to humans, for example. Additionally, because there is still a monotonic relationship between the time estimates and true admixture times, it might be feasible to correct for this bias and obtain accurate admixture time estimates even when ancestral populations are not particularly divergent.

Two-pulse admixture model

It is significantly more challenging to distinguish between models with a single ancestry pulse and models with two distinct ancestry pulses from the same ancestral population. As we described above, one of the key difficulties is that, in addition to estimating admixture times, our model must also estimate the proportion of the total ancestry contributed by each pulse. Furthermore, it is essential that this approach includes a mechanism for distinguishing between single- and double-pulse models. We therefore began with a single-pulse admixture simulation and fit both single- and two-pulse admixture models to these data.

We find that traditional likelihood ratio tests (LRT) universally favor two-pulse *vs.* single-pulse models even when data are simulated under a single-pulse model for all scenarios we considered (Figure S3). Additionally, we note that two-pulse models may become degenerate, either because a pulse's ancestry proportion may be nearly zero, or because

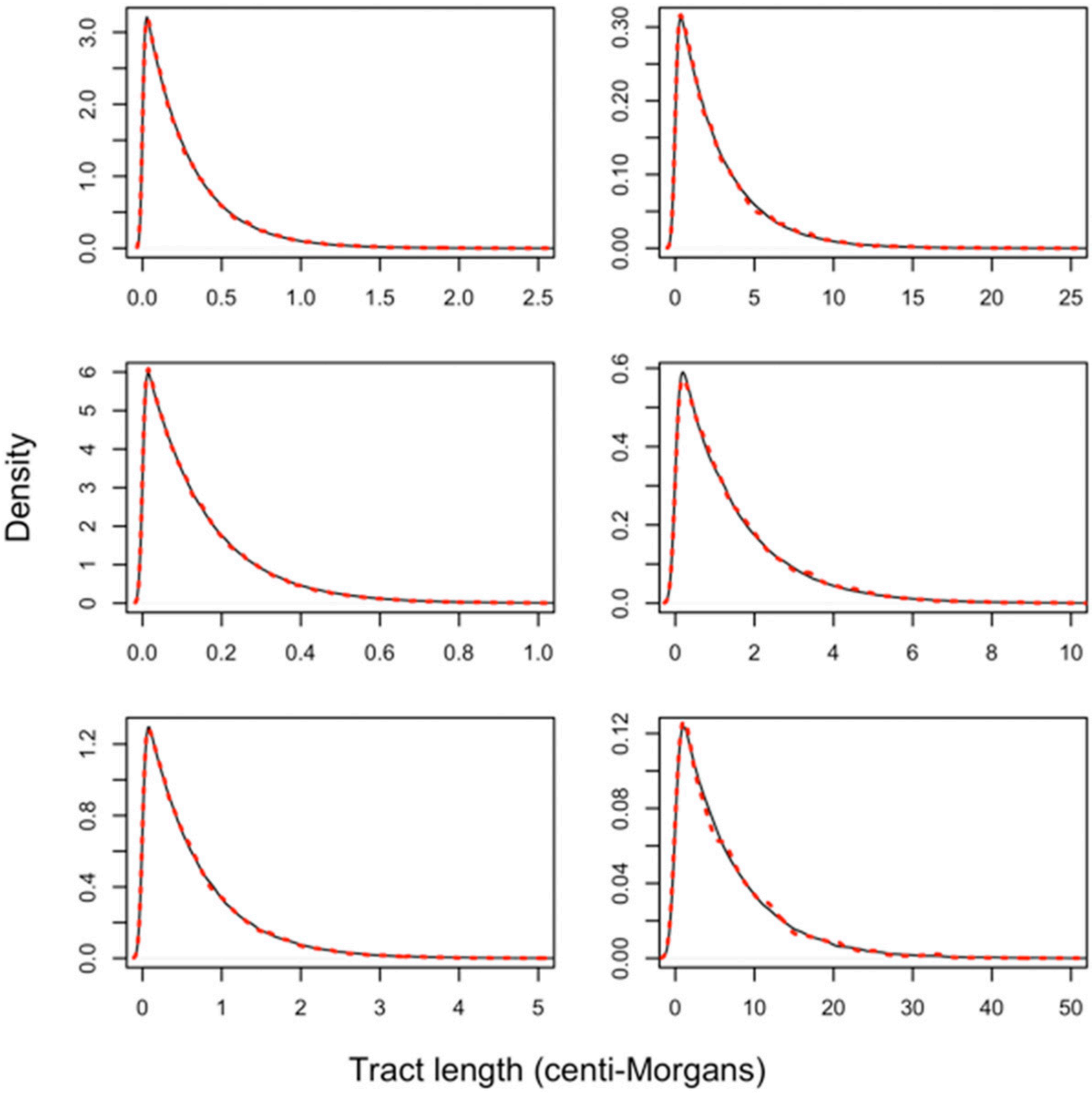

Figure 2 Tract length distributions obtained using our tract length model approximation (solid black) and forward-in-time simulation (dashed red). A model like Figure 1c was considered. An initial pulse of ancestry type A_2 entered a resident population of ancestry type A_3 and replaced 1/3 of the resident population. Then, a second ancestry pulse in forward time, from ancestry type A_1, replaced 1/4 of the resident population. Each simulation had a diploid population size of size 10,000, and we aggregated data from 50 sampled individuals across 100 simulations to produce the full tract length distribution. From top to bottom, respective ancestry tract length distributions correspond to ancestry type A_3, A_1, and A_2. We investigated two admixture models. The first model has a first and second pulse occurring 200 and 1000 generations, respectively, before the present (left). The second model has a first and second pulse occurring 20 and 100 generations, respectively, prior to the present (right). Note that axes differ between figure panels.

two pulses may occur at virtually identical times. We therefore caution that these factors could make traditional LRTs challenging to interpret. We therefore suggest that, when using this method to analyze admixed populations, it will usually be preferable to choose the simplest admixture model that is consistent with the data. More complex models are usually favorable by standard statistical comparisons, but may overfit the data. This consideration is reminiscent of concerns for the STRUCTURE model (Pritchard *et al.* 2000; Evanno *et al.* 2005). In general, this means the model that contains the fewest distinct ancestry pulses should be selected.

Spurious admixture models can be identified by interrogating the timing and proportion of ancestry contributed by each admixture pulse. First, many of the resulting two-pulse admixture models contain ancestry pulses with similar times (Figure 4 and Tables S3–S5), suggesting that these could be recognized as representing a single admixture event. Second, of the models containing two pulses, many contain one pulse near the correct admixture time and another distant pulse that introduced a very small proportion, *i.e.*, ≤1% of the total ancestry in the sampled admixed population (Figure 4 and Tables S3–S5). Often the pulse introducing a small

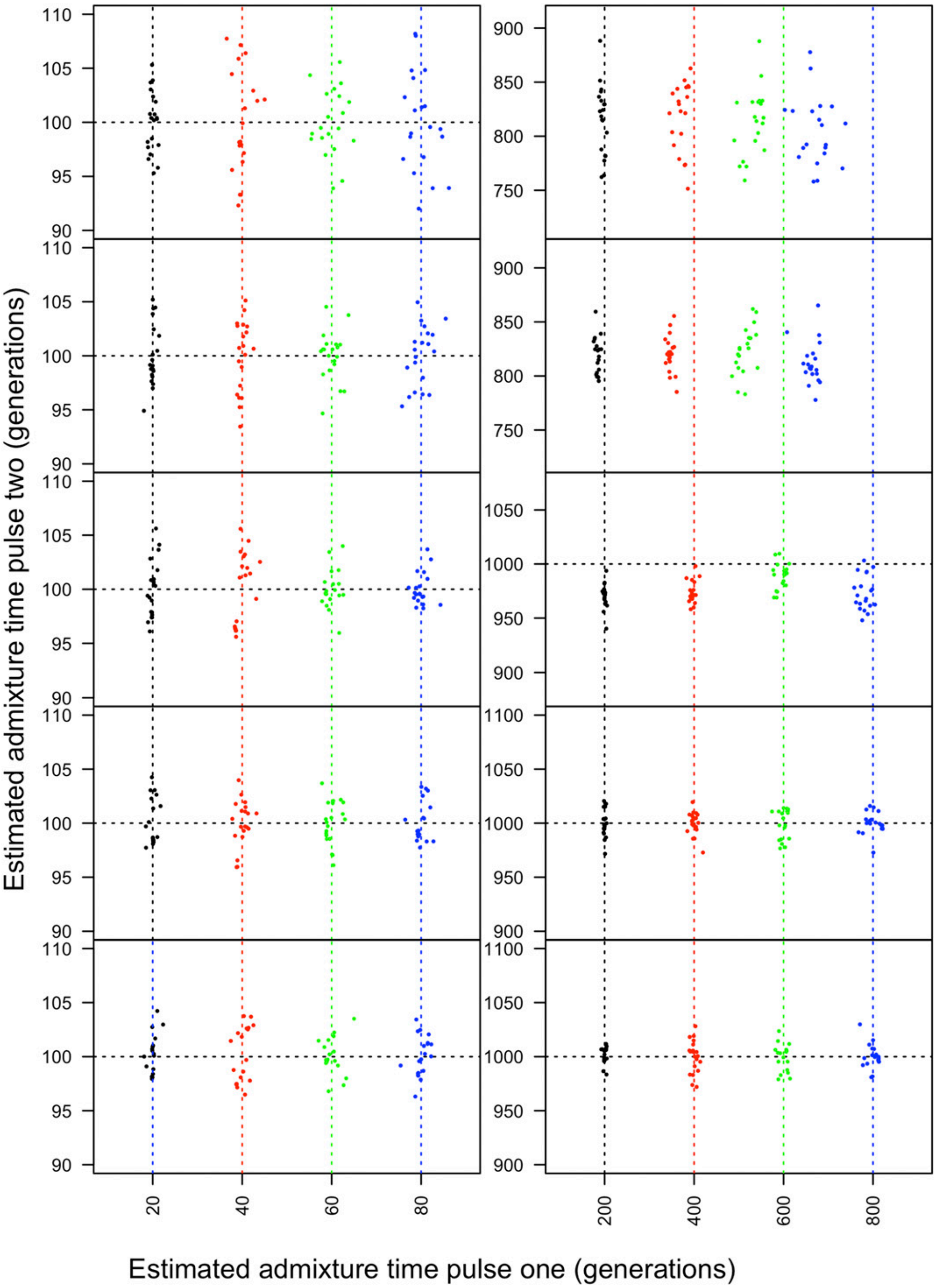

Figure 3 Admixture time estimates for two-pulse population models with three distinct ancestral populations. From top to bottom, panels include the divergence time between the ancestral populations is 0.05, 0.1, 0.25, 0.5, and 1 N_e generations. True admixture times are indicated by the dashed lines. On the left, for all admixture models considered, the second pulse occurred 100 generations before the present. The first pulse occurred 20 (black), 40 (red), 60 (green), and 80 (blue) generations prior to sampling. On the right, the second pulse occurred 1000 generations before the present, and the first pulse occurred at 200 (black), 400 (red), 600 (green) and 800 (blue) generations prior to sampling. Note that, due to underestimation of admixture times, the figure axes differ between panels for plots of more ancient admixture events.

proportion of ancestry in the sampled population occurred in the distant past, where admixture time is harder to estimate. Therefore, admixture pulses that (1) occur in the distant past and contribute relatively small proportions of the total ancestry or (2) pulse at similar times, may indicate a spurious admixture model and should be disregarded in favor of simpler admixture scenarios.

When we simulated admixed populations whose histories contained two distinct ancestry pulses, we found two-pulse models could sometimes be fit accurately. Specifically, when the two ancestry pulses occurred at fairly different times (*e.g.*, 20 and 100, and 200 and 1000 generations), our approach identified models that correspond closely to the parameters under which the admixed population was simulated. However, when the two ancestry pulses occurred at closer time intervals, we were less frequently able to reliably recover the correct admixture model (Figure 5 and Table S4). Additionally, we note that the accuracy of two-pulse models depends

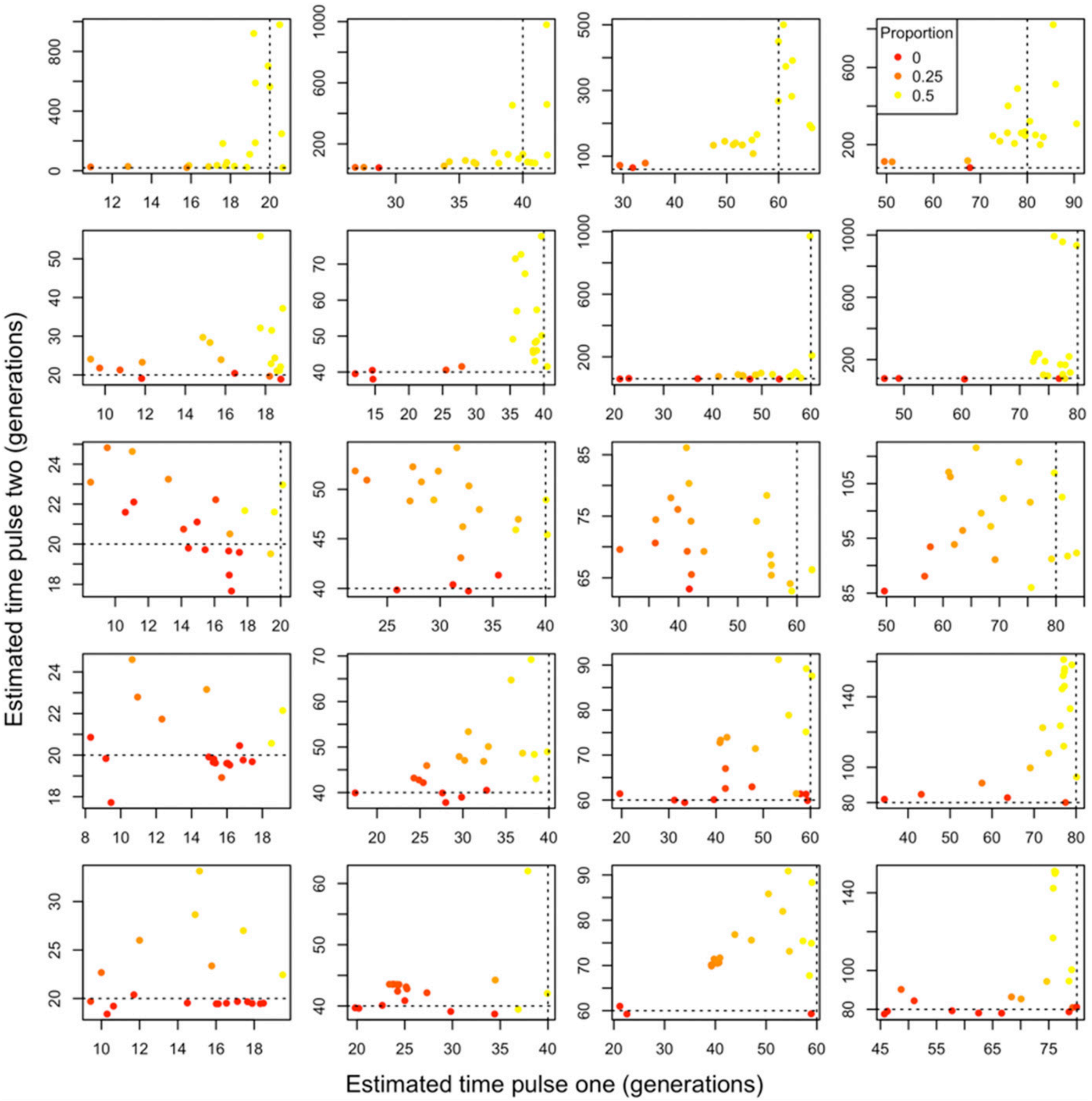

Figure 4 Two-pulse admixture models fitted using our framework to data generated under a single-pulse admixture model. We considered varying levels of population divergence. From top to bottom, the ancestral populations are 0.05, 0.1, 0.25, 0.5, and 1 N_e generation divergent from one another. From left to right, the single admixture pulse occurred 20, 40, 60, and 80 generations prior to sampling and replaced one half of the ancestry. Point colors correspond to the proportion of ancestry that is attributable to the second pulse, with a gradient running from 0 (red) to the maximum possible, 0.5 (yellow). Dashed lines reflect the true admixture time. Note that axes differ between figure panels.

on the genetic distance between ancestral populations (Figure 5), where it is generally more straightforward to fit admixture models for genetically distant ancestral populations.

Collectively, our results suggest that it might not be possible to distinguish between single-pulse and two-pulse admixture models when ancestry pulses occurred at similar times. However, it is feasible to distinguish single pulse models from admixture models with both relatively ancient and recent admixture. Therefore, we anticipate that this approach will be valuable for investigating a range of hypotheses with dramatically different admixture times. Furthermore, as we consider here intermediate-sized population samples, *i.e.*, 50 individuals, it may be feasible to distinguish more fine-grained admixture models using larger samples sizes.

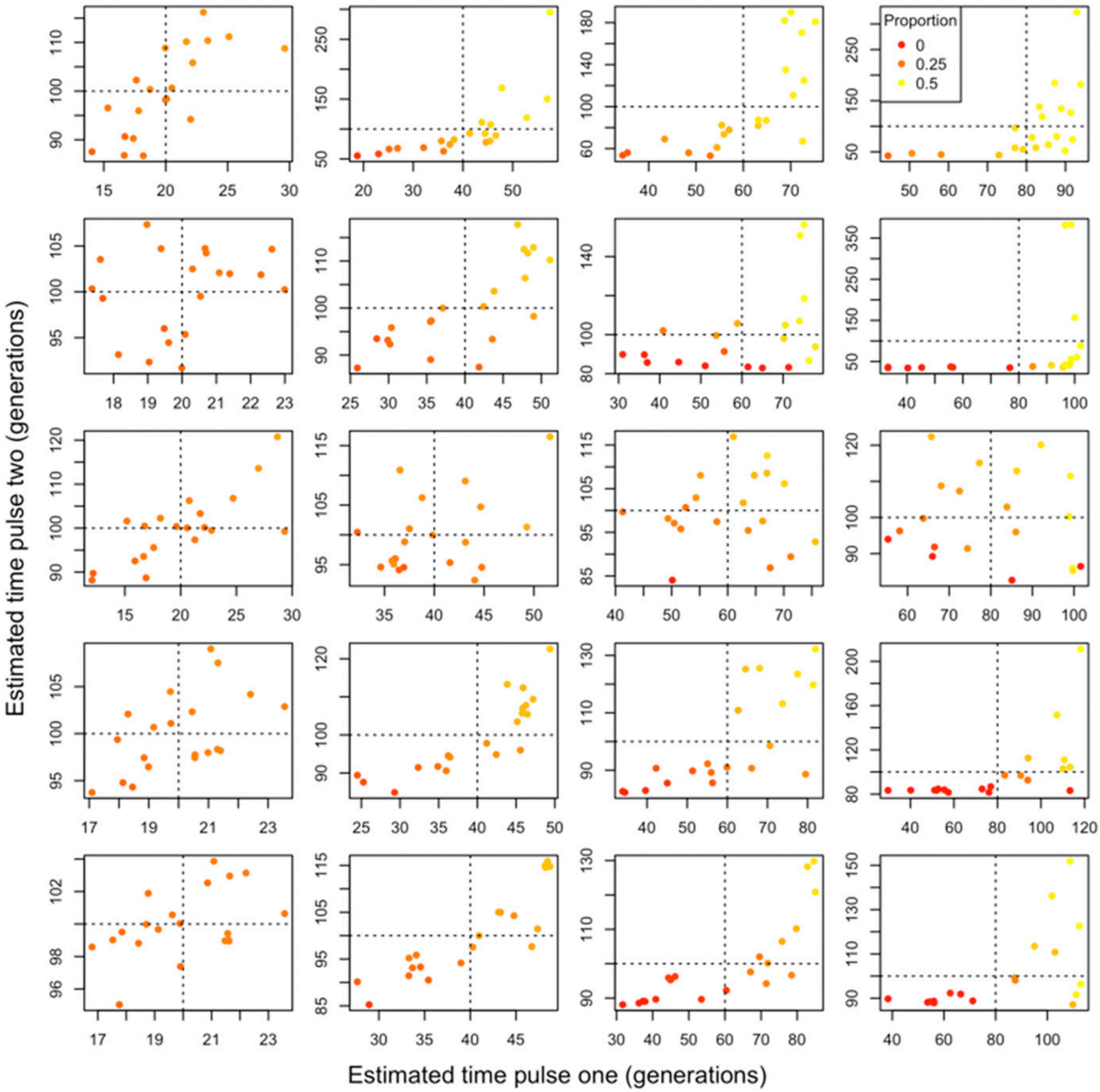

Figure 5 Two-pulse admixture models fitted using our framework to data generated under a two-pulse admixture model. We considered varying levels of population divergence. From top to bottom, the ancestral populations are 0.05, 0.1, 0.25, 0.5, and 1 N_e generation divergent from one another. For all models, the second admixture pulse occurred 100 generations prior to sampling. From left to right, the first pulse occurred 20, 40, 60, and 80 generations prior to sampling. The second pulse replaced 1/3 of the resident population and the first pulse replaced 1/4 of the resident population. Therefore, each ancestral population contributed one half of the ancestry at the time of sampling. Point colors correspond to the proportion of total ancestry at the time of sampling that is attributable to the first pulse with a gradient running from 0 (red) to the maximum, 0.5 (yellow). Dashed lines reflect the true admixture times for the first and second pulse. Note that axes may differ between subplots.

LAI accuracy

We note that while two pulse models do improve the accuracy of LAI, the improvement observed tends to be slight (Tables S3–S5). This suggests that for studies aimed at evaluating patterns of LA across the genome, there may be no need to extensively optimize admixture models to obtain reasonable estimates of the LA landscape in the admixed population. Single-pulse models may be sufficient to accurately characterize LA across the genome for many admixed populations.

Robustness of our approach to smaller sample sizes, divergent reference populations, and inaccuracies in the genetic map

We anticipate our approach will be used to estimate the timing of admixture in model and nonmodel systems. It is likely that the underlying assumptions will be violated in some applications. To characterize our model's performance and robustness in light of these challenges, we evaluated our method's ability to correctly estimate the timing of admixture under varying sample sizes, reference population divergences, and erroneous recombination maps. See File S1 for definitions of summary statistics used to quantify accuracy of admixture models and LAI.

Accurate fine-scale recombination maps are necessary for estimating transition rates between adjacent ancestry states in our model. However, we anticipate that our method will be applied to nonmodel systems where a fine-scale recombination map might not be available. Therefore, we tested the resilience of our method against perturbations of the recombination map using populations that are 0.25 N_e generations divergent (*Methods*). The normalized root mean squared error (NRMSE) in estimating the timing of a single pulse admixture event using our approach is $<12\%$ for all admixture times 20, 40, 60, and 80 even when the recombination map is substantially distorted (Figure S6 and Tables S6 and S7). Moreover, our approach is unlikely to give false support for a two pulse model to truly single pulse data (Figure S7). When fitting a two-pulse model to two pulse data, our method consistently overestimates the timing of admixture events when the true recombination map is distorted relative to the assumed map. However, for all but the most strongly distorted recombination maps, our results imply that two pulse models can be reliably identified, albeit with overestimated admixture times (Figure S8). Therefore, for single-pulse admixture models, a high-density recombination map does not appear necessary; however, for accurate two-pulse admixture models, a high density genetic map may be required to obtain accurate admixture time estimates. See Table S7 for detailed error reports as a result of perturbed rates of recombination.

Additionally, the source population contributing to the admixture history of a population may not always be available for genetic sequencing, or may be incorrectly identified. To test the sensitivity of our approach when a related, but distinct, population from the true admixture source is used as a reference, we simulated admixture events using reference populations of varying divergence to the actual source population and estimated the timing of admixture (see *Methods*). We found our approach consistently overestimated the timing of admixture in both single and double pulse models (Figures S9 and S10). In general, as reference populations are more divergent from the true admixture source, our method increasingly overestimated the timing of admixture events. Nonetheless, it is feasible to identify a two-pulse model, despite overestimating the timing of admixture (Figure S10). Moreover, we fitted a double-pulse model to single-pulse data and found no false support for a double-pulse model (Figure S11). We advise that, when using our approach, it will be necessary to carefully identify the closest possible reference population, as there are limitations of how divergent the reference can be to the actual source population in order to obtain accurate estimates of admixture time (Tables S8 and S9).

Lastly, we evaluated our method's performance under varying sample sizes of the admixed population. Because we considered only a single chromosome in all analysis, we note that our sample sizes considered here, down to 10 individuals, are very modest relative to most sequencing applications with respect to total data considered. Regardless, we do not obtain false support for a two-pulse model when applied to admixed populations that experienced only a single pulse (Figure S12). Estimating the timing of admixture becomes more accurate with large sample sizes. However, a sample size of just 10 individuals at a single chromosome still has a NRMSE $<6\%$ for all admixture times we considered (Tables S10–S12). Therefore, we conclude that our approach is still useful in cases with limited sampling.

Comparison to ALDER

We compared the accuracy of our approach in estimating the timing of admixture to that of ALDER (Loh *et al.* 2013). Briefly, ALDER and its extension, MALDER (implemented in Pickrell *et al.* (2014)), is based on modeling the decay of admixture LD within admixed populations. In all single- and double-pulse models that we considered (See *Methods*), our approach more accurately estimates the timing of admixture (Tables S13 and S14). Notably, our method significantly outperforms ALDER in models involving pulses occurring more distantly in the past (*i.e.*, >80 generations ago). This discrepancy might be attributable to our method's ability to model short tracts of LA, which get disregarded in ALDER by default (Loh *et al.* 2013), potentially impacting the ability to identify ancient admixture events (Figure S13). We note that estimating ancient admixture pulses is not the intended use of ALDER, but was included in this analysis to fully compare the utility of our software in estimating both recent and ancient admixture pulses. Considering a relatively recent two-pulse model where $t_1 = 20$ and $t_2 = 100$, MALDER correctly fits a two-pulse model 6 out of 20 of the simulations, but selects a single-pulse model for the 14 others (Table S14). Our approach fit a two-pulse model to the model in all 20 simulations and produced a NRMSE of 0.0307 and 0.0214 for t_1 and t_2, respectively, compared to MALDER's 0.1207 and 0.1856. We used default parameters for ALDER/MALDER simulations and further parameterization might improve its performance (see *Methods*).

When is this method a good choice?

We recommend using our software over ALDER/MALDER, especially when estimating ancient admixture pulses. Our software performs best with reference populations that are minimally divergent to the source population and an accurate fine-scale recombination map. Our results suggest that this approach is robust to varying levels of data availability and to errors in the fine-scale genetic map. Given that this is the only approach for LAI and admixture time estimation that can accommodate nondiploid samples, short-read pileup data, and relatively ancient admixture pulses, it will likely be useful in a variety of systems.

Nonetheless, we advise carefully considering, *e.g.*, through simulation, how this approach will perform with available data, particularly when ancestral LD is extensive or populations are only weakly genetically differentiated, such as what is found in some human admixed human populations. Particularly for these situations, we suggest to instead apply LAI software that explicitly models the haplotype structure of admixed samples (Price *et al.* 2009; Maples *et al.* 2013; Dias-Alves *et al.* 2018) and use the resulting tract length distribution to estimate admixture time (Gravel 2012; Ni *et al.* 2018). Moreover, if there is concern that the assumed reference population is divergent from the true reference population, other methods such as ALDER might be more robust in estimating the timing of admixture (Loh *et al.* 2013).

Application

Admixture in D. melanogaster

Biogeographical evidence (Lachaise *et al.* 1988) and patterns of genetic variation (Begun and Aquadro 1993; Caracristi and Schlötterer 2003; Thornton and Andolfatto 2006) suggest that *D. melanogaster* originated in sub-Saharan Africa, and went on to colonize much of the rest of the world relatively recently (Duchen *et al.* 2013). During this expansion, the population that left sub-Saharan Africa experienced a dramatic bottleneck that reshaped patterns of genetic diversity across much of the genome (Caracristi and Schlötterer 2003; Ometto *et al.* 2005; Duchen *et al.* 2013). The resulting "cosmopolitan" haplotypes are distinguishable from those of the ancestral sub-Saharan populations based solely on patterns of genetic variation (Pool *et al.* 2012), and these two ancestral populations have since encountered each other and admixed in numerous geographic locations within sub-Saharan Africa and worldwide (Caracristi and Schlötterer 2003; Pool *et al.* 2012).

Thus, *D. melanogaster* has emerged as an important model for understanding the genomic and phenotypic consequences of admixture in natural populations. There are important premating behavioral isolation barriers between cosmopolitan and African individuals (Ting *et al.* 2001), as well as substantial phenotypic consequences that operate postmating within admixed individuals (Lachance and True 2010; Kao *et al.* 2015). Additionally, recent work has investigated the genomic consequences of admixture both demographically (Duchen *et al.* 2013; Bergland *et al.* 2016) and with the goal of evaluating the impact of natural selection in shaping genome-wide patterns of variation (Pool 2015). The genomic consequences of admixture have only been studied using relatively simple demographic models [*i.e.*, a single event (Pool *et al.* 2012)], leaving open the possibility that more complex admixture dynamics are common in natural populations of this species. Accurately characterizing the admixture histories and the patterns of local ancestry across the genome is essential to further our understanding of the demographic history of this species and build a complete null model for studying natural selection on ancestry during admixture.

Evaluating possible applications to D. melanogaster admixed populations

In agreement with our general conclusions from admixture simulations, we find that our method is accurate, and single-pulse admixture models are distinguishable from two-pulse models for temporally distinct admixture events (*i.e.*, wherein two pulses occurred at significantly different times, Figure 6 and Figure S5). However, also consistent with our results above, we find that when admixture pulses occurred at relatively similar times (*e.g.*, $t_1 = 200$ and $t_2 = 250$ generations prior to sampling), a single ancestry pulse contributes the vast majority of admixed ancestry, and this scenario is therefore ultimately most similar to a single-pulse admixture model. These data therefore suggest that it is possible to distinguish single- from two-pulse admixture models in data consistent with *D. melanogaster* ancestral populations when the admixture pulses occurred at dramatically different times.

Perhaps owing to the higher marker density, the greater accuracy of genotyped markers along the genome, and/or the inbred genomes in our *Drosophila* samples, we find that our method is slightly more accurate for population histories consistent with those of *D. melanogaster* than for the simulated populations considered above (Tables S3–S5). Additionally, admixture times can be accurately estimated even when admixture events occurred in the distant past (*e.g.*, 5000 generations prior to sampling, Figure 6 and Figure S5).

Application to admixed D. melanogaster samples

To demonstrate the performance of our method on real datasets, we applied our approach to *D. melanogaster* variation data from sub-Saharan African populations (Figure 7). We studied populations from Rwanda (RG), South Africa (SD), and Gambella, Ethiopia (EA) and estimated the time of cosmopolitan admixture using both double and single pulse models.

In two-pulse models, cosmopolitan ancestry pulsed twice (Figure 1b), and resulted in the most recent pulse contributing ~99 and 97.5% (for RG and SD, respectively) of the total cosmopolitan ancestry present in the population at the time of sampling. The second, more ancient, pulse tends toward the maximum time allowed in our inference method, 10,000 generations, and contributes 0.01 and 2.5% of the total cosmopolitan ancestry present in the population at the time of sampling RG and SD (Figure 7). As described above, a distant admixture pulse contributing small amounts of ancestry likely indicates a spurious admixture model. Though it is

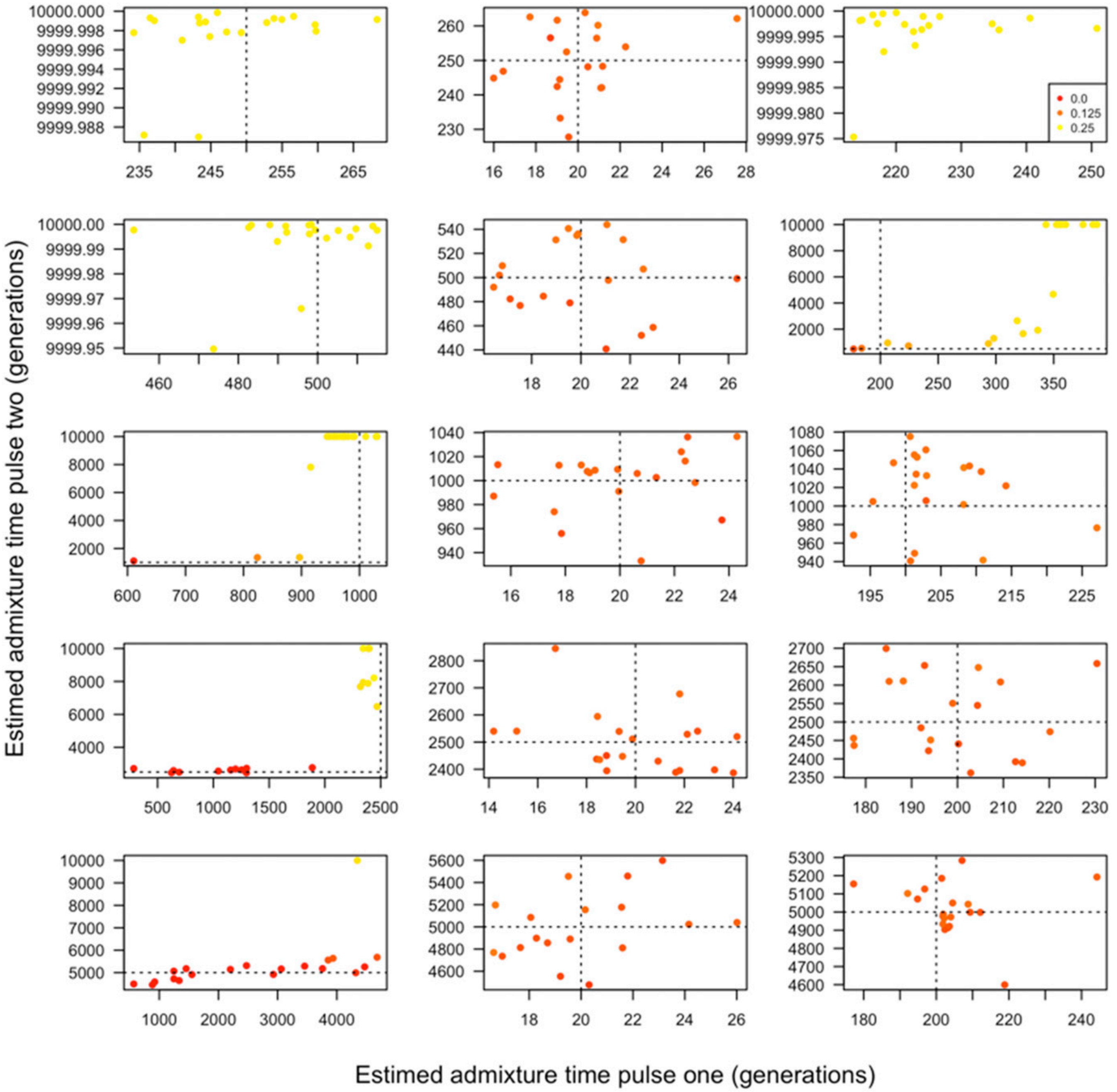

Figure 6 Admixture model fitted for data consistent with admixed *Drosophila* populations. Two-pulse admixture models for scenarios that are truly single-pulse (left), two-pulse with a first admixture pulse 20 generations prior to sample (middle), and two-pulse models with a first admixture pulse 200 generations prior to sampling (right). From top to bottom, the second admixture pulse occurred 250, 500, 1000, 2500, and 5000 generations prior to sampling. The point colors indicate the proportion of ancestry in the sampled population that entered during the first admixture pulse in backward time with a gradient running from 0 (red) to the maximum, 0.5 (yellow). Dashed lines indicate the correct timing of simulated ancestry pulses. In two-pulse models, the first pulse contained 10% of the final ancestry and the second contributed 14% of ancestry to the sampled population. In single-pulse models, all of the non-native ancestry is contributed by one pulse. These values were selected to be consistent with those from admixed sub-Saharan populations of this species (Pool *et al.* 2012). Note that axes may differ between subplots.

possible that ancient cosmopolitan admixture contributed a small amount to the ancestry of the SD and RG populations, the simplest model, a single-pulse model with cosmopolitan ancestry pulsing into African ancestry, provides a reasonable description of the demographic history of RG and SD. The estimated admixture times, 140 and 437 generations in a single pulse model, suggest that the observed cosmopolitan ancestry has entered these populations exceptionally recently despite earlier contact and extensive commerce among human groups.

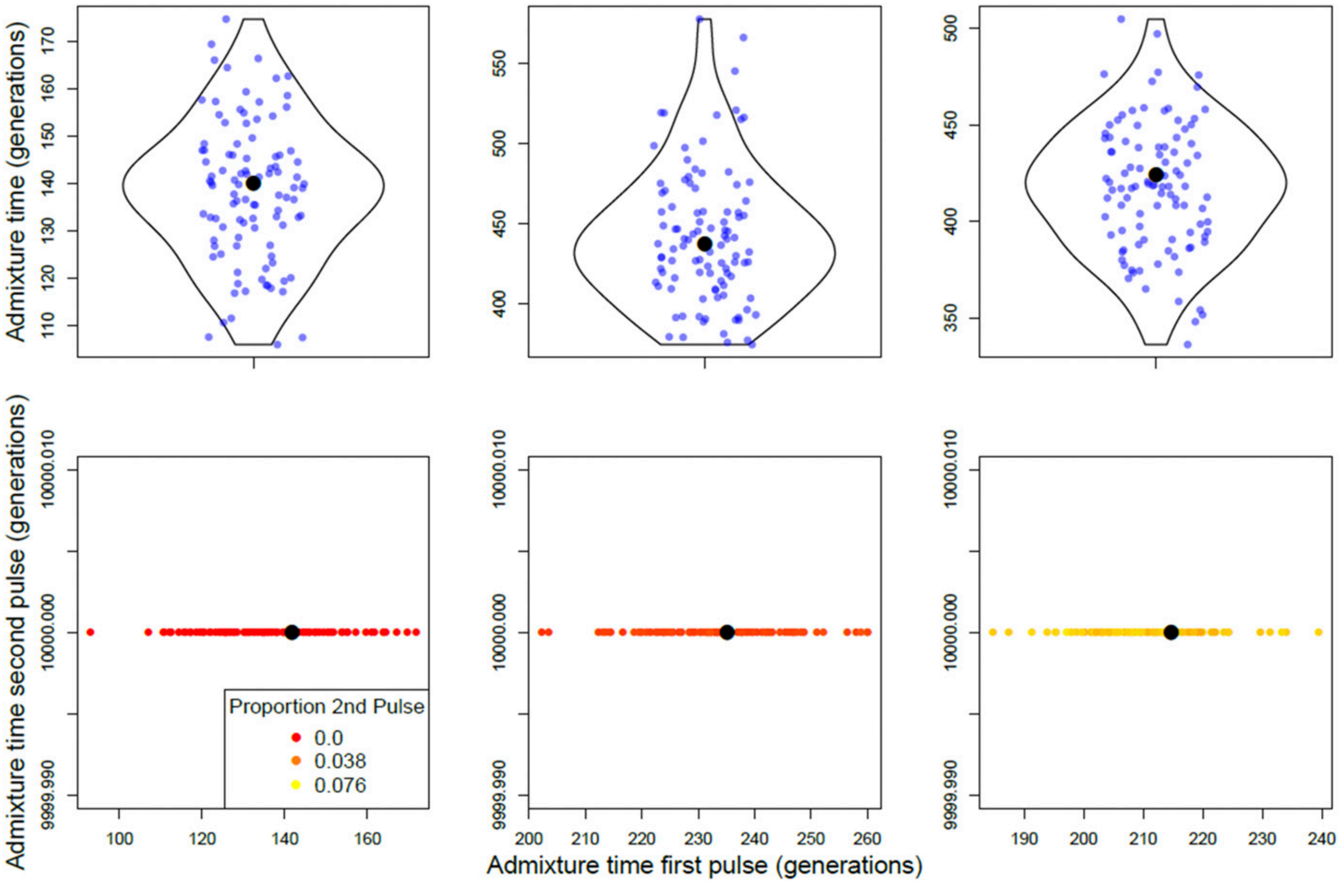

Figure 7 Single- and double-pulse models applied to real genotype data generated from sub-Saharan African populations of *D. melanogaster*. In the plots shown, cosmopolitan ancestry pulsed into native African ancestry once (top row) or twice (bottom row). The panels from left to right are populations RG, SD, and EA. Single- and double-pulse models were bootstrapped 100 times. Each point in the top row corresponds to a bootstrap estimate of the timing of admixture. The bottom row shows bootstrap estimates for the timing of the second admixture pulse in backward time. The point colors indicate the proportion of ancestry in the sampled population that entered during the second (in backward time) admixture pulse, with a gradient running from 0 (red) to the maximum, 0.076 (yellow). Black dots indicate the bootstrapped time optima. Note that axes may differ between subplots.

To test how our method performs when ancestral populations are omitted, we applied a two-pulse model to genotype data (EA) from Gambella, Ethiopia, which is believed to have West African, Ethiopian, and cosmopolitan ancestry (Lack *et al.* 2015). Instead of modeling all three ancestral populations, we omitted West African as an ancestral population to mimic a scenario where the ancestral populations of an admixed population have not all been identified. The results of the two ancestral state model estimated an ancient admixture pulse composing 6.5% of the total cosmopolitan ancestry present in the population at the time of sampling (Figure 7). Importantly, we note that 6.5% is substantially more than we found in any of our simulated single-pulse datasets. A second pulse ∼200 generations ago contributed 28% of cosmopolitan ancestry. Since this two-pulse model did not include ancestry from a West African population, it is likely that our method fit a portion of the allele frequency differentiated between the ancestral populations into the ancient pulse of cosmopolitan ancestry. Thus, a large proportion of ancestry contributed by an ancient admixture pulse might indicate the need to identify additional ancestral populations. A two-pulse model including three ancestral populations is therefore more appropriate for studying the admixture history of EA.

Application to three-population mixture in Gambella, Ethiopia (EA)

Our results above emphasize the importance of identifying and incorporating variation data from all ancestral populations. Here, we use a two-pulse, three-ancestral-population, admixture model to estimate the order and timing of admixture events into EA (Figure 1c and Figure 8).

The native ancestry of *D. melanogaster* from EA is unknown. Previous studies have suggested both cosmopolitan and sub-Saharan ancestry contribute to the genetic variation in EA (Figure 1c) (Lack *et al.* 2015). To determine the most likely native ancestry of EA, we applied three permutations of a two-pulse model to population data from Gambella, Ethiopia. We used unadmixed populations from Ethiopia (EF), cosmopolitan (FR), and West African (AF) to represent ancestral populations of EA, and used our program to estimate the timing of admixture in each admixture model. Our three models were: EF and AF pulsing into resident FR ancestry, AF

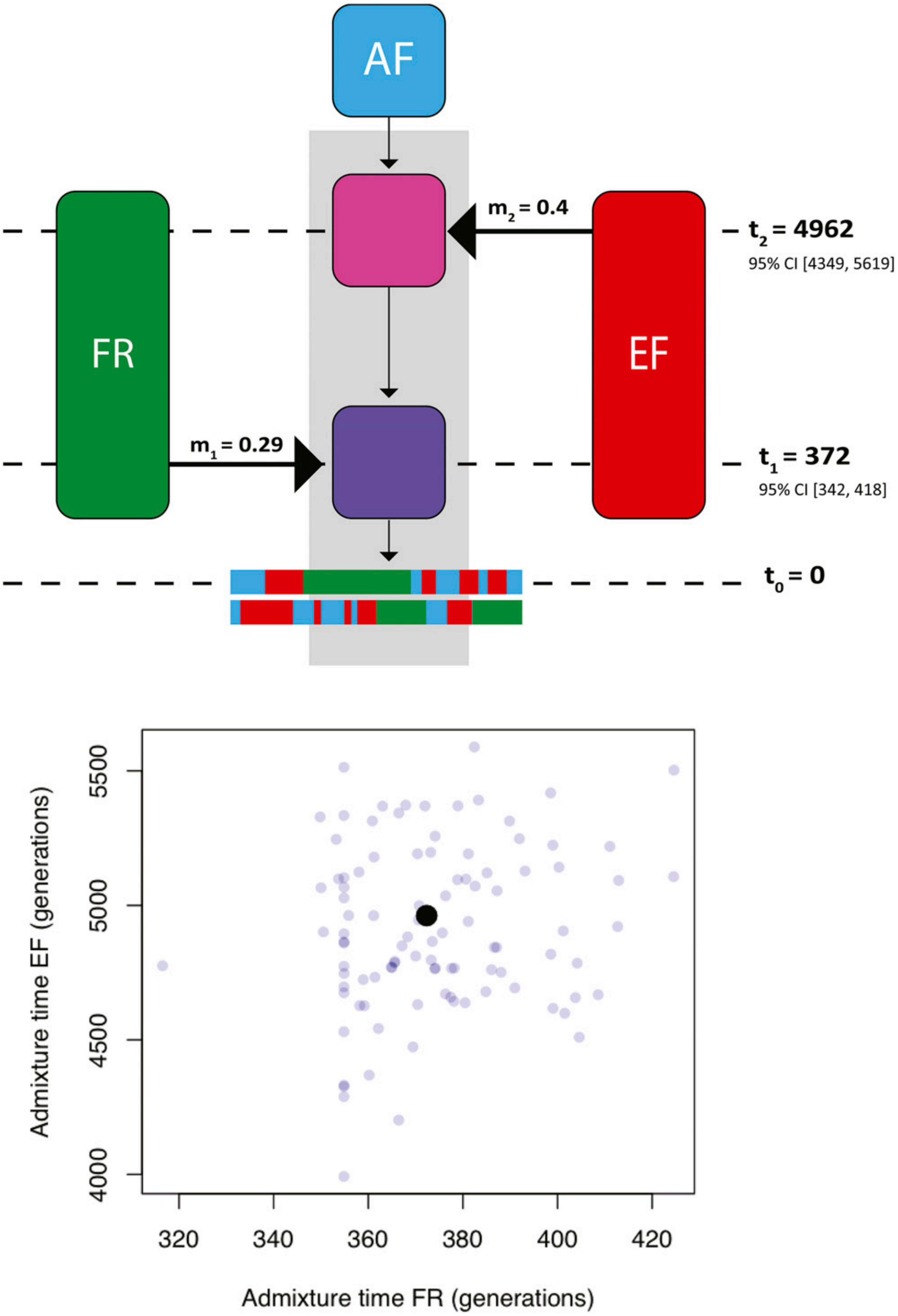

Figure 8 Most likely admixture model fitted to real data generated from Gambella, Ethiopia (EA). Schematic of double-pulse admixture model with three ancestral populations and ancestry types is shown (top). Note the schematic is not drawn to scale. A total of 1000 bootstraps were run of FR and EA pulsing into AF native ancestry. Real bootstrap data are shown (bottom), where each point is a bootstrap estimate of FR and EA admixture pulses and the black point represents the optimal admixture model from the full data set. The optima and 95% confidence intervals around the estimated timing of admixture are reported as part of the schematic.

and FR pulsing into EF ancestry, and EF and FR pulsing into AF ancestry.

When FR is supplied as the resident population in the two pulse model, EF and AF both enter the population 1140 generations ago. This is implausible for a variety of reasons. First, even without considering this result, it is unlikely that cosmopolitan ancestry is the native ancestry type in Gambella, Ethiopia. Second, it is exceptionally unlikely that two distinct ancestral populations enter the admixed population at precisely the same time. Finally, the likelihood of the FR native ancestry model is less favored relative to an EF or AF native model by 11,000 log likelihood units. We therefore consider a model with native African ancestry, EF or AF, to most likely describe the admixture history of Gambella, Ethiopia *D. melanogaster.*

When EF and AF were modeled as native ancestry, resulting admixture histories were nearly identical, suggesting either EF or AF could be native to Gambella since these are equivalent models. Using a two-pulse, three-ancestral-population model, we estimated the timing of EF and FR admixture pulses in AF native ancestry (Figure 1c). The second pulse in backward time is estimated to have occurred ~4962, 95% CI [4349, 5619], generations ago, and suggests that EA admixed with genetically distinct sub-Saharan African populations in the relatively distant past (Figure 8). The most recent pulse introduced FR ancestry 372, 95% CI [342, 418] generations ago, suggesting EA recently admixed with cosmopolitan populations, which is strikingly consistent with our estimates from the other admixed populations in sub Saharan Africa (Figure 3 and Figure 8). Additionally, the relatively recent admixture between West African and Ethiopian *D. melanogaster*, suggests that sub-Saharan populations of this species were isolated until very recently (*i.e.*, ~338 years ago).

We note that our estimate of admixture time is congruent with an estimate of divergence between Ethiopian and Central African populations. Kern and Hey (2017) estimated Ethiopian and West African *D. melanogaster* diverged ~3628 years ago (or ~50,000 generations). Our estimates of admixture time occur sufficiently far after this estimated Ethiopian and African divergence time and provide confidence that our three-population admixture model of Gambella (EA) *D. melanogaster* is consistent with previous investigations of demographic patterns of *D. melanogaster* from sub-Saharan Africa. These data therefore indicate that our approach can be useful in estimating the timing of multiple admixture events in the history of a population using real genotype data.

Conclusion

Admixture histories can be complex with numerous distinct ancestral populations contributing genetic material to a recipient population at multiple times in the past. In this work, we developed coalescent theory as well as a method to estimate the timing of multiple admixture events in an admixed population. We applied our model to forward-time simulated and real sequence data. The results of our simulations suggest that our model can discern between ancestry pulses occurring far apart in time from each other. Moreover, our method excels when the populations contributing ancestry are genetically diverse from each other.

Our approach, when applied to real admixed *D. melanogaster* populations, is consistent with previous results on the African origin and admixture of the *D. melanogaster* species. We find that cosmopolitan ancestry has entered very recently, and is best accommodated by our framework using a single-pulse model. Additionally, we demonstrate that more complex admixture patterns have shaped the ancestry of Gambella, Ethiopia. We indicate that two ancestry pulses from distinct ancestry types are necessary to explain patterns of genetic variation within this population.

This method improves the estimates of multiple pulse admixture models while accommodating the possibility of more than one source population. Although not a focus of this study, our method is designed to accommodate read pile-up data, samples of arbitrary ploidy, and samples with unknown admixture history. Taken all together, we anticipate that our approach will facilitate continued exploration of admixture's contributions to fundamental biological processes such as adaptation, ecological divergence, and speciation.

Acknowledgments

We thank Audrey T. Ford for helpful edits to the manuscript, and the Corbett-Detig and Nielsen laboratories for suggestions and feedback on the project. This work was supported by the National Institute of General Medical Sciences/ National Institutes of Health (NIH) grant R35GM128932-01 and the Sloan Research Fellowship (to R.C.-D.). P.M. was funded by NIH training grant T32 HG008345.

Literature Cited

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of Drosophila melanogaster are very different at the DNA level. Nature 365: 548–550. https://doi.org/10.1038/365548a0

Bergland, A. O., R. Tobler, J. Gonzalez, P. Schmidt, and D. Petrov, 2016 Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in Drosophila melanogaster. Mol. Ecol. 25: 1157–1174. https://doi.org/10.1111/mec.13455

Bukowicki, M., S. U. Franssen, and C. Schlötterer, 2016 High rates of phasing errors in highly polymorphic species with low levels of linkage disequilibrium. Mol. Ecol. Resour. 16: 874–882. https://doi.org/10.1111/1755-0998.12516

Caracristi, G., and C. Schlötterer, 2003 Genetic differentiation between American and European Drosophila melanogaster populations could be attributed to admixture of African alleles. Mol. Biol. Evol. 20: 792–799. https://doi.org/10.1093/molbev/msg091

Chakraborty, R., and K. M. Weiss, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA 85: 9119–9123. https://doi.org/10.1073/pnas.85.23.9119

Chen, G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142. https://doi.org/10.1101/gr.083634.108

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

Dias-Alves, T., J. Mairal, and M. G. Blum, 2018 Loter: a software package to infer local ancestry for a wide range of species. *Mol. Biol. Evol*. 35: 2318–2326. https://doi.org/10.1093/molbev/msy126

Duchen, P., D. Živković, S. Hutter, W. Stephan, and S. Laurent, 2013 Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. Genetics 193: 291–301. https://doi.org/10.1534/genetics.112.145912

Evanno, G., S. Regnaut, and J. Goudet, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. 14: 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

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

Gravel, S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio *et al.*, 2013 Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. 9: e1004023. https://doi.org/10.1371/journal.pgen.1004023

Hellenthal, G., G. B. Busby, G. Band, J. F. Wilson, C. Capelli *et al.*, 2014 A genetic atlas of human admixture history. Science 343: 747–751. https://doi.org/10.1126/science.1243518

Hufford, M. B., P. Lubinksy, T. Pyhäjärvi, M. T. Devengenzo, N. C. Ellstrand *et al.*, 2013 The genomic signature of crop-wild introgression in maize. PLoS Genet. 9: e1003477. https://doi.org/10.1371/journal.pgen.1003477

Kao, J. Y., A. Zubair, M. P. Salomon, S. V. Nuzhdin, and D. Campo, 2015 Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south-eastern United States and Caribbean islands. Mol. Ecol. 24: 1499–1509. https://doi.org/10.1111/mec.13137

Kern, A. D., and J. Hey, 2017 Exact calculation of the joint allele frequency spectrum for isolation with migration models. Genetics 207: 241–253. https://doi.org/10.1534/genetics.116.194019

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the Drosophila melanogaster species subgroup, pp. 159–225 in *Evolutionary Biology*. Springer, New York. https://doi.org/10.1007/978-1-4613-0931-4_4

Lachance, J., and J. R. True, 2010 X-autosome incompatibilities in Drosophila melanogaster: tests of Haldane's rule and geographic patterns within species. Evolution 64: 3035–3046.

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig *et al.*, 2015 The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics 199: 1229–1241. https://doi.org/10.1534/genetics.115.174664

Lack, J. B., J. D. Lange, A. D. Tang, R. B. Corbett-Detig, and J. E. Pool, 2016 A thousand fly genomes: an expanded Drosophila genome nexus. Mol. Biol. Evol. 33: 3308–3313. https://doi.org/10.1093/molbev/msw195

Liang, M., and R. Nielsen, 2014 The lengths of admixture tracts. Genetics 197: 953–967. https://doi.org/10.1534/genetics.114.162362

Liang, M., and R. Nielsen, 2016 Estimating the timing of multiple admixture events using 3-locus linkage disequilibrium. bioRxiv 2016: 078378.

Loh, P.-R., M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell *et al.*, 2013 Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193: 1233–1254. https://doi.org/10.1534/genetics.112.147330

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante, 2013 Rfmix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. 93: 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Moorjani, P., N. Patterson, J. N. Hirschhorn, A. Keinan, L. Hao *et al.*, 2011 The history of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genet. 7: e1001373. https://doi.org/10.1371/journal.pgen.1001373

Nelder, J. A., and R. Mead, 1965 A simplex method for function minimization. Comput. J. 7: 308–313. https://doi.org/10.1093/comjnl/7.4.308

Ni, X., K. Yuan, X. Yang, Q. Feng, W. Guo *et al.*, 2018 Inference of multiple-wave admixtures by length distribution of ancestral tracks. Heredity 121: 52–63. https://doi.org/10.1038/s41437-017-0041-2

Ometto, L., S. Glinka, D. De Lorenzo, and W. Stephan, 2005 Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. 22: 2119–2130. https://doi.org/10.1093/molbev/msi207

Paşaniuc, B., S. Sankararaman, G. Kimmel, and E. Halperin, 2009 Inference of locus-specific ancestry in closely related populations. Bioinformatics 25: i213–i221. https://doi.org/10.1093/bioinformatics/btp197

Pickrell, J. K., N. Patterson, P.-R. Loh, M. Lipson, B. Berger *et al.*, 2014 Ancient West Eurasian ancestry in Southern and Eastern Africa. Proc. Natl. Acad. Sci. USA 111: 2632–2637. https://doi.org/10.1073/pnas.1313787111

Pool, J. E., 2015 The mosaic ancestry of the Drosophila genetic reference panel and the D. melanogaster reference genome reveals a network of epistatic fitness interactions. Mol. Biol. Evol. 32: 3236–3251.

Pool, J. E., and R. Nielsen, 2009 Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719. https://doi.org/10.1534/genetics.108.098095

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno *et al.*, 2012 Population genomics of sub-Saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet. 8: e1003080. https://doi.org/10.1371/journal.pgen.1003080

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5: e1000519. https://doi.org/10.1371/journal.pgen.1000519

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Pugach, I., R. Matveev, V. Spitsyn, S. Makarov, I. Novgorodov *et al.*, 2016 The complex admixture history and recent southern origins of Siberian populations. Mol. Biol. Evol. 33: 1777–1795. https://doi.org/10.1093/molbev/msw055

Rieseberg, L. H., O. Raymond, D. M. Rosenthal, Z. Lai, K. Livingstone *et al.*, 2003 Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301: 1211–1216. https://doi.org/10.1126/science.1086949

Sanderson, J., H. Sudoyo, T. M. Karafet, M. F. Hammer, and M. P. Cox, 2015 Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. Genetics 200: 469–481. https://doi.org/10.1534/genetics.115.176842

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin, 2008 Estimating local ancestry in admixed populations. Am. J. Hum. Genet. 82: 290–303. https://doi.org/10.1016/j.ajhg.2007.09.022

Sankararaman, S., N. Patterson, H. Li, S. Pääbo, and D. Reich, 2012 The date of interbreeding between Neandertals and modern humans. PLoS Genet. 8: e1002947. https://doi.org/10.1371/journal.pgen.1002947

Sankararaman, S., S. Mallick, M. Dannemann, K. Prüfer, J. Kelso *et al.*, 2014 The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507: 354–357. https://doi.org/10.1038/nature12961

Skotte, L., T. S. Korneliussen, and A. Albrechtsen, 2013 Estimating individual admixture proportions from next generation sequencing data. Genetics 195: 693–702. https://doi.org/10.1534/genetics.113.154138

Thornton, K., and P. Andolfatto, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. Genetics 172: 1607–1619. https://doi.org/10.1534/genetics.105.048223

Ting, C.-T., A. Takahashi, and C.-I. Wu, 2001 Incipient speciation by sexual isolation in Drosophila: concurrent evolution at multiple loci. Proc. Natl. Acad. Sci. USA 98: 6709–6713. https://doi.org/10.1073/pnas.121418898

Communicating editor: J. Novembre