Inference under a Wright-Fisher model using an accurate beta approximation

Paula Tataru*, Thomas Bataillon*, and Asger Hobolth*

*Bioinformatics Research Centre, Aarhus University, Aarhus, 8000, Denmark

Running title: An accurate beta approximation

Key words: Wright-Fisher, beta, pure drift, linear evolutionary pressures, divergence times

Corresponding author:

Paula Tataru

Bioinformatics Research Centre

Aarhus University

C.F. Mllers All 8

Aarhus C 8000, Denmark

paula@cs.au.dk

Abstract

The large amount and high quality of genomic data available today enables, in principle, accurate inference of evolutionary history of observed populations. The Wright-Fisher model is one of the most widely used models for this purpose. It describes the stochastic behavior in time of allele frequencies and the influence of evolutionary pressures, such as mutation and selection. Despite its simple mathematical formulation, exact results for the distribution of allele frequency (DAF) as a function of time are not available in closed analytic form. Existing approximations build on the computationally intensive diffusion limit, or rely on matching moments of the DAF. One of the moment-based approximations relies on the beta distribution, which can accurately describe the DAF when the allele frequency is not close to the boundaries (zero and one). Nonetheless, under a Wright-Fisher model, the probability of being on the boundary can be positive, corresponding to the allele being either lost or fixed. Here, we introduce the beta with spikes, an extension of the beta approximation, which explicitly models the loss and fixation probabilities as two spikes at the boundaries. We show that the addition of spikes greatly improves the quality of the approximation. We additionally illustrate, using both simulated and real data, how the beta with spikes can be used for inference of divergence times between populations, with comparable performance to existing state-of-the-art method.

INTRODUCTION

Advances in sequencing technologies have revolutionized the collection of genomic data, increasing both the volume and quality of available sequenced individuals from a large variety of populations and species (Romiguier *et al.* 2014; Gudbjartsson *et al.* 2015). These data, which may involve up to millions of single nucleotide polymorphisms (SNPs), contain information about the evolutionary history of the observed populations. There has been great focus in the recent years on inferring such histories and, to this end, one of the most widely used models is the Wright-Fisher (Gautier *et al.* 2010; Sirén *et al.* 2011; Malaspinas *et al.* 2012; Pickrell and Pritchard 2012; Gautier and Vitalis 2013; Steinrücken *et al.* 2014; Terhorst *et al.* 2015).

The Wright-Fisher model characterizes the evolution of a randomly mating population of finite size in discrete non-overlapping generations. The model describes the stochastic behavior in time of the number of copies (frequency) of alleles at a locus. The frequency is influenced by a series of factors, such as random genetic drift, mutations, migrations, selection, and changes in population size. When inferring the evolutionary history of a population, the effects of the different factors have to be untangled. The frequency varies from one generation to the next due to random sampling of a finite sized population (random genetic drift). Mutations, migrations and selection affect the sampling probability in a deterministic manner. We collectively refer to these as evolutionary pressures. Mutations and migrations result in linear changes of the sampling probability, while selection is a non-linear pressure (Kimura 1964; Crow and Kimura 1970) and is therefore more difficult to study analytically.

A crucial step for carrying out statistical inference in the Wright-Fisher model is the determination of the distribution of the allele frequency (DAF) as a function of time, conditional on an initial frequency. Even though the Wright-Fisher model has a very simple mathematical formulation, no tractable analytical form exists for the DAF (Ewens 2004). Therefore, various approximations have been developed, ranging from purely analytical to

purely numerical. They generally either build on the diffusion limit of the Wright-Fisher, or rely on matching moments of the true DAF. Both types of approximations have been used successfully for inference of populations divergence times (Sirén *et al.* 2011; Gautier and Vitalis 2013), populations admixture (Pickrell and Pritchard 2012), SNPs under selection (Gautier *et al.* 2010) and selection coefficients from time serial data (Malaspinas *et al.* 2012; Steinrücken *et al.* 2014; Terhorst *et al.* 2015).

Wright (1945) was the first to use the diffusion approximation to determine the stationary DAF. Kimura (1955) solved the diffusion limit and found the time-dependent distribution for pure drift, and Crow and Kimura (1956) extended the solution to include linear evolutionary pressures. However, these contain infinite sums, making their use cumbersome in practice. After decades dominated by inference based on the dual coalescent process (Rosenberg and Nordborg 2002; Hoban *et al.* 2012), diffusion has recently received increasing attention, and researchers have started to investigate other ways to solving analytically or approximating the diffusion equation (McKane and Waxman 2007; Waxman 2011; Malaspinas *et al.* 2012; Song and Steinrücken 2012; Zhao *et al.* 2013; Steinrücken *et al.* 2013; Steinrücken *et al.* 2014).

Moment-based approximations are less ambitious in that they aim at fitting mathematical convenient distributions by equating the first moments of the true DAF. Such approximations typically use either the normal distribution (Nicholson *et al.* 2002; Coop *et al.* 2010; Gautier *et al.* 2010; Pickrell and Pritchard 2012; Terhorst *et al.* 2015) or the beta distribution (Balding and Nichols 1995; Balding and Nichols 1997; Sirén *et al.* 2011; Sirén 2012). The rationale behind the use of these distributions is two-fold. Firstly, they are motivated by the diffusion limit: the normal distribution is the resulting DAF when drift is small (Nicholson *et al.* 2002), while the beta distribution is the stationary DAF under linear evolutionary pressures (Wright 1945; Crow and Kimura 1956). Secondly, they are entirely determined by their mean and variance. One major difference between the two is their support. Because the normal distribution is defined over the whole real line, it needs to be truncated to $[0, 1]$ (Nicholson

et al. 2002; Coop *et al.* 2010; Gautier *et al.* 2010). The truncated normal distribution has two atoms at zero and one (corresponding to the allele being lost or fixed) containing the densities in the intervals $(-\infty, 0]$ and $[1, \infty)$, respectively. However, the truncation procedure leads to a variance that no longer matches the variance of the true DAF (Gautier and Vitalis 2013). Alternatively, the full distribution can be applied for intermediary frequencies only, when the probabilities of lying outside the zero and one boundaries are small and can therefore be ignored (Pickrell and Pritchard 2012; Terhorst *et al.* 2015). Unlike the normal distribution, the beta distribution has the interval $[0, 1]$ as support, but, due to its continuous nature, the probabilities at the boundaries will always be zero. Under a Wright-Fisher model, the loss and fixation events have a positive probability. The beta distribution provides a good fit for intermediary frequencies, but fails at capturing the non-zero boundary probabilities, as illustrated for pure drift in Figure 1A – C. When time is small, most of the probability mass is found close to the initial value x_0 (Figure 1A). As time becomes larger, the allele frequency drifts away from x_0 and more and more probability accumulates at the boundaries (Figure 1B and C).

Here, we propose an accurate extension of the beta distribution under linear evolutionary pressures, entitled the beta with spikes, which explicitly models the probabilities at the boundaries. We show that the addition of spikes greatly improves the fit to the true DAF. We use simulation experiments and published chimpanzee exome data to demonstrate that the beta with spikes can be used for inference of population divergence times under pure drift, with performance comparable with a state-of-the-art diffusion-based method, and less computational burden. We additionally discuss how the beta with spikes can be used in future development to account for variable population size and selection.

Here, we propose an accurate extension of the beta distribution under linear evolutionary pressures, entitled the beta with spikes, which explicitly models the probabilities at the boundaries. We show that the addition of spikes greatly improves the fit to the true DAF. We use simulation experiments and published chimpanzee exome data to demonstrate that

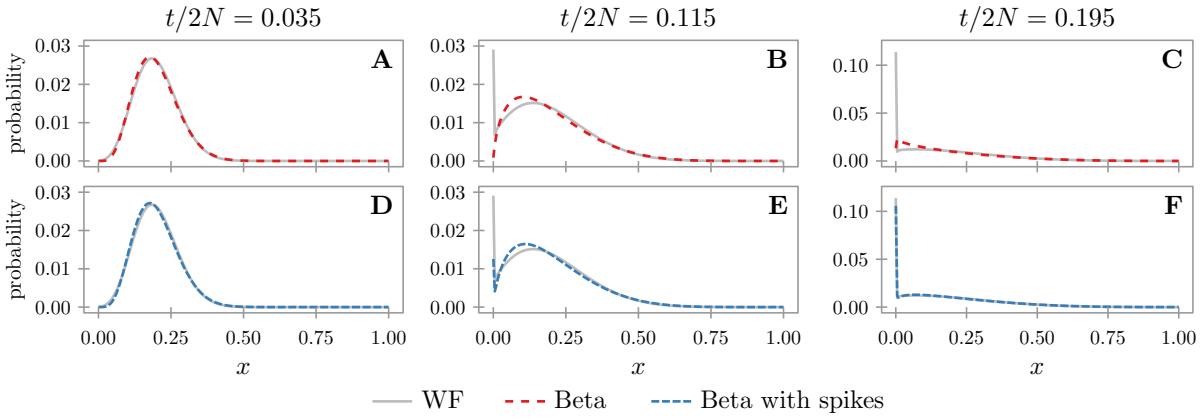


Figure 1: Fit of the beta and beta with spikes approximations. The figure shows the true discrete DAF as given by the Wright-Fisher model with a population size $2N = 200$ under pure drift, and the corresponding discretized beta (A – C) and beta with spikes (D – F) approximations. The distributions are conditional on an initial frequency $x_0 = 0.2$ and for different time points: $t/2N = 0.035$ (A, D), $t/2N = 0.115$ (B, E) and $t/2N = 0.195$ (C, F), where t is the number of discrete generations that the population has evolved. The discretization procedure is detailed in the Supplementary Material.

the beta with spikes can be used for inference of population divergence times under pure drift, with performance comparable with a state-of-the-art diffusion-based method, and less computational burden.

THE BETA WITH SPIKES APPROXIMATION

Consider a diploid randomly mating population of size $2N$ and a biallelic locus with alleles A_1 and A_2. Under a Wright-Fisher model, the count of one of the alleles, A_1, at the discrete generation t is a random variable $Z_t \in \{0, 1, \ldots, 2N\}$. Let $X_t = Z_t/(2N)$ be the corresponding allele frequency. The evolution of Z_t is shaped by random genetic drift and deterministic evolutionary pressures. We capture the joint effect of the deterministic pressures in $g(x)$, a polynomial in the allele frequency $0 \leq x \leq 1$. Conditional on Z_t, Z_{t+1} follows a binomial

distribution (Ewens 2004)

$$Z_{t+1} \mid Z_t = z_t \sim \mathrm{Bin}(2N,\, g(x_t)). \tag{1}$$

Here, we only consider linear evolutionary pressures, such as mutation and migration. Then $g(x)$ takes the form

$$g(x) = (1-a)\,x + b. \tag{2}$$

The parameters a and b verify that $0 \leq b \leq a \leq 1$ such that $0 \leq g(x) \leq 1$ for all $0 \leq x \leq 1$. The case where $a = 1$, for which $g(x) = b$ for all $0 \leq x \leq 1$, has no biological meaning and we therefore assume that $a \neq 1$.

Under pure drift, $a = b = 0$. If mutations happen with probabilities u (from A_1 to A_2) and v (from A_2 to A_1), then $a = u + v$ and $b = v$. Migration can be modeled, for example, by assuming that individuals can migrate away from the population under study and that there is an influx of individuals from a large population with constant frequency x_c. Then, with probabilities m_1 and m_2, individuals migrate from and to the population under study, respectively. We have $a = m_1$ and $b = m_2 x_c$. Mutation and migration can be modeled jointly, resulting in $a = m_1 + (1 - m_1)(u + v)$ and $b = (1 - m_1)v + m_2 x_c$. In the following, we treat the general linear case.

We are interested in the distribution of allele frequency (DAF) X_t conditional on $X_0 = x_0$, as a function of the generation t,

$$f(x;t) = \mathbb{P}\left(\, X_t = x \mid X_0 = x_0 \,\right). \tag{3}$$

For simpler notation, we leave out the explicit condition on $X_0 = x_0$, and implicit condition on population size and evolutionary pressures.

Under the beta approximation, the DAF is

$$f_B(x;t) = \frac{x^{\alpha_t - 1}\,(1-x)^{\beta_t - 1}}{\mathrm{B}\left(\,\alpha_t, \beta_t\,\right)}, \tag{4}$$

where $\mathrm{B}(\alpha, \beta)$ is the beta function. The two shape parameters of the beta distribution are entirely determined by its mean and variance,

$$\begin{aligned} \alpha_t &= \left(\frac{\mathbb{E}[X_t](1-\mathbb{E}[X_t])}{\mathrm{Var}(X_t)} - 1 \right) \mathbb{E}[X_t], \\ \beta_t &= \left(\frac{\mathbb{E}[X_t](1-\mathbb{E}[X_t])}{\mathrm{Var}(X_t)} - 1 \right) (1-\mathbb{E}[X_t]). \end{aligned} \tag{5}$$

Therefore, in order to fit f_B to f, we need to calculate $\mathbb{E}[X_t]$ and $\mathrm{Var}(X_t)$. These can be obtained in closed analytical form (see Supplementary Material for full derivation). The mean is entirely determined by the initial frequency x_0 and the parameters a and b of the linear evolutionary pressures, while the variance also depends on the population size. Under pure drift ($a = b = 0$) we have

$$\begin{aligned} \mathbb{E}[X_t] &= x_0, \\ \mathrm{Var}(X_t) &= x_0(1-x_0)\left(1 - \left(1 - \frac{1}{2N}\right)^t\right). \end{aligned} \tag{6}$$

When $a \neq 0$ we get

$$\begin{aligned} \mathbb{E}[X_t] = {} & \frac{b}{a} + (1-a)^t \left(x_0 - \frac{b}{a}\right), \\ \mathrm{Var}(X_t) = {} & \frac{b}{a}\left(1-\frac{b}{a}\right) \frac{1-(1-a)^{2t}\left(1-\frac{1}{2N}\right)^t}{2N-(1-a)^2(2N-1)} \\ & + \left(1-\frac{2b}{a}\right)\left(x_0-\frac{b}{a}\right)(1-a)^t \frac{1-(1-a)^t\left(1-\frac{1}{2N}\right)^t}{2N-(1-a)(2N-1)} \\ & - \left(x_0-\frac{b}{a}\right)^2 (1-a)^{2t}\left(1-\left(1-\frac{1}{2N}\right)^t\right). \end{aligned} \tag{7}$$

In the limit of infinite population size, the above formulas are equivalent to the mean and variance obtained by Sirén (2012) (up to some minor typographical errors, as confirmed by correspondence with the author; see also Supplementary Material).

To account for loss and fixation probabilities, we surround the beta distribution with two

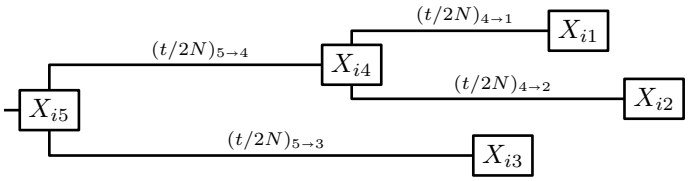
spikes

$$\begin{aligned} f_B^\star(x;t) = \quad & \mathbb{P}\left(X_t = 0\right) \cdot \delta(x) \\ & + \mathbb{P}\left(X_t = 1\right) \cdot \delta(1-x) \\ & + \mathbb{P}\left(X_t \notin \{0,1\}\right) \cdot \frac{x^{\alpha_t^\star - 1}\,(1-x)^{\beta_t^\star - 1}}{\mathrm{B}\left(\alpha_t^\star, \beta_t^\star\right)}, \end{aligned} \tag{8}$$

where $\delta(x)$ is the Dirac delta function and $\mathbb{P}\left(X_t \notin \{0,1\}\right) = 1 - \mathbb{P}\left(X_t = 0\right) - \mathbb{P}\left(X_t = 1\right)$. To fit $f_B^\star$ to f, we need to determine the mean and variance of X_t conditional on polymorphism ($X_t \notin \{0,1\}$), and the probabilities $\mathbb{P}\left(X_t = 0\right)$ and $\mathbb{P}\left(X_t = 1\right)$ of loss and fixation, respectively. Given $\mathbb{E}\left[X_t\right]$, $\mathrm{Var}\left(X_t\right)$, $\mathbb{P}\left(X_t = 0\right)$ and $\mathbb{P}\left(X_t = 1\right)$, the conditional mean and variance can easily be calculated (see Supplementary Material). Therefore, we only require means of calculating the loss and fixation probabilities in order to fully specify the beta with spikes approximation. We use a recursive approach where we calculate the probabilities for X_{t+1} by relying on the approximated $f_B^\star(x;t)$. We additionally assume that a and b are small to obtain the following approximation for loss and fixation probabilities (see Supplementary Material for full derivation)

$$\begin{aligned} \mathbb{P}\left(X_{t+1} = 0\right) \approx \quad & \mathbb{P}\left(X_t = 0\right) \cdot (1-b)^{2N} \\ & + \mathbb{P}\left(X_t = 1\right) \cdot (a-b)^{2N} \\ & + \mathbb{P}\left(X_t \notin \{0,1\}\right) \cdot (1-a)^{2N} \frac{\mathrm{B}\left(\alpha_t^\star, \beta_t^\star + 2N\right)}{\mathrm{B}\left(\alpha_t^\star, \beta_t^\star\right)}, \\ \mathbb{P}\left(X_{t+1} = 1\right) \approx \quad & \mathbb{P}\left(X_t = 0\right) \cdot b^{2N} \\ & + \mathbb{P}\left(X_t = 1\right) \cdot (1-a+b)^{2N} \\ & + \mathbb{P}\left(X_t \notin \{0,1\}\right) \cdot (1-a)^{2N} \frac{\mathrm{B}\left(\alpha_t^\star + 2N, \beta_t^\star\right)}{\mathrm{B}\left(\alpha_t^\star, \beta_t^\star\right)}. \end{aligned} \tag{9}$$

Figure 1D – F depicts the beta with spikes approximation for the same cases as in Figure 1A – C. When time is small (Figure 1A and D), the beta and beta with spikes distributions are equivalent, but as the time becomes larger, the advantage of adding the spikes becomes evident. As illustrated in supplementary Figure S1, the addition of spikes



Sample size	Allele count
n_{i1}	z_{i1}
n_{i2}	z_{i2}
n_{i3}	z_{i3}

Figure 2: History of three populations in the present. The ancestral population 5 splits in populations 3 and 4, which further splits in populations 2 and 1. For each SNP i and present population $j \in \{1, 2, 3\}$, the data consists of the sample size n_{ij} and allele count z_{ij}. The branch length between populations k and j is given as $(t/2N)_{k\to j}$ and represents the scaled number of generations that population j evolved since the split from the ancestral population k. The unknown allele frequencies of each population are denoted as X_{ij}, with $1 \leq j \leq 5$.

drastically improves the fit of the beta approximation to the true DAF under a Wright-Fisher model.

INFERENCE OF DIVERGENCE TIMES

To further illustrate the advantage of incorporating the spikes, we inferred divergence times between populations, using both simulated data and exome sequencing data from three chimpanzee subspecies (Bataillon *et al.* 2015).

Populations are represented as successive descendants of a single ancestral population. We assume that after each split, the new populations evolved in isolation (no migration) under pure drift. A rooted tree (Figure 2) can be used to describe the joint history of several present populations, located at the leaves, while the common ancestral population is represented as the root. The data $\mathcal{D} = \{(z_{ij}, n_{ij}) \mid 1 \leq i \leq I, 1 \leq j \leq J\}$ consist of I independent SNPs for J populations in the present: the (arbitrarily defined) reference (A_1) allele count z_{ij} in a sample of size n_{ij} ($0 \leq z_{ij} \leq n_{ij}$) for each locus $1 \leq i \leq I$ and population $1 \leq j \leq J$.

Conditional on the topology (i.e. tree without branch lengths), we inferred the scaled branch lengths by numerically maximizing the likelihood of the data.

Likelihood of the data: Assuming Hardy-Weinberg equilibrium, the probability of observing z_{ij} alleles in a sample of size n_{ij} given the population allele frequency x_{ij} is given by the binomial distribution

$$\mathbb{P}\left(z_{ij} \mid n_{ij}, x_{ij} \right) = \binom{n}{z_{ij}} x_{ij}^{z_{ij}} \ (1 - x_{ij})^{n_{ij} - z_{ij}} \ . \tag{10}$$

However, the allele frequencies x_{ij} are unobserved and the likelihood of the data $\mathcal{D}_i$ for SNP i is obtained by integrating over the unknown allele frequencies

$$\begin{aligned} L(\mathcal{D}_i; \Theta, \pi) = \int_0^1 \dots \int_0^1 & f(X_{i1}, X_{i2}, \dots, X_{iJ} \mid \Theta, \pi) \\ & \cdot \prod_{j=1}^{J} \mathbb{P}\left(z_{ij} \mid n_{ij}, X_{ij} \right) \ \mathrm{d}X_{i1} \dots \mathrm{d}X_{iJ}, \end{aligned} \tag{11}$$

where $f(X_{i1}, X_{i2}, \dots, X_{iJ} \mid \Theta, \pi)$ is the joint distribution of the X_{ij}'s at the leaves. The likelihood is a function of the scaled branch lengths, denoted here as Θ, and π, the unknown DAF at the root. The joint distribution $f(X_{i1}, X_{i2}, \dots, X_{iJ} \mid \Theta, \pi)$ is, in turn, an integral over the allele frequencies in the ancestral populations, represented as internal nodes in the tree. We approximate the integrals with sums by discretizing the allele frequencies. The discretized joint distribution is then obtained using a peeling algorithm (Felsenstein 1981), where the transition probabilities on each branch are given by the DAF (see Supplementary Material for details). As the SNPs are assumed to be independent, the full likelihood is a product over the SNPs,

$$L(\mathcal{D}; \Theta, \pi) = \prod_{i=1}^{I} L(\mathcal{D}_i; \Theta, \pi). \tag{12}$$

As SNP data contains only polymorphic sites, we further condition the above likelihood on polymorphic data as follows

$$L(\mathcal{D}_i; \Theta, \pi \mid \text{polymorphism}_i) = \frac{L(\mathcal{D}_i; \Theta, \pi)}{\mathbb{P}\left(\text{polymorphism}_i \mid \Theta, \pi \right)}, \tag{13}$$

where

$$\mathbb{P}\left(\text{ polymorphism}_i \mid \Theta, \pi\right) = 1 - L(\mathcal{D}_i^0; \Theta, \pi) - L(\mathcal{D}_i^1; \Theta, \pi). \tag{14}$$

Here, $\mathcal{D}_i^0$ and $\mathcal{D}_i^1$ are data corresponding to site i where the allele was lost or fixed, respectively, in the samples from all populations,

$$\mathcal{D}_i^0 = \{(0, n_{ij}) \mid 1 \leq j \leq J\}, \qquad \mathcal{D}_i^1 = \{(n_{ij}, n_{ij}) \mid 1 \leq j \leq J\}. \tag{15}$$

We treat π, the root DAF, as a nuisance parameter assumed to be a beta distribution. For a given topology (i.e. tree without branch lengths), the most likely branch lengths and shape parameters of π can be recovered by numerically maximizing the likelihood conditional on polymorphism.

Simulated data: Using the topology depicted in Figure 2, we simulated multiple data sets containing independent SNPs under a Wright-Fisher model, given an acenstral frequency X_{i5} sampled from π, the root DAF, which we set to be a beta distribution. We used two different scenarios, labeled I and II, summarized in Table 1. Scenario I has a uniform π and large sample sizes, while scenario II is built to produce data that resembles the chimpanzee exome data analyzed below. For this, we used the chimpanzee sample sizes, and scaled branch lengths and root DAF as inferred by the beta with spikes on the chimpanzee data (see also Table 2).

For each simulated data set, we estimated the branch lengths using both the beta and beta with spikes as described previously. We additionally ran Kim Tree (Gautier and Vitalis 2013) using the default settings. Kim Tree is a method designed for inference of divergence times between populations evolving under pure drift. It uses Kimura's solution to the diffusion limit for the DAF (Kimura 1955) and relies on a Bayesian MCMC approach. Here, we use the posterior means of the branch lengths as point estimates.

All methods estimate well the branches leading to populations 1 and 2 (Figure 3). Beta with spikes estimates the branch lengths more accurately and with lower variance than the

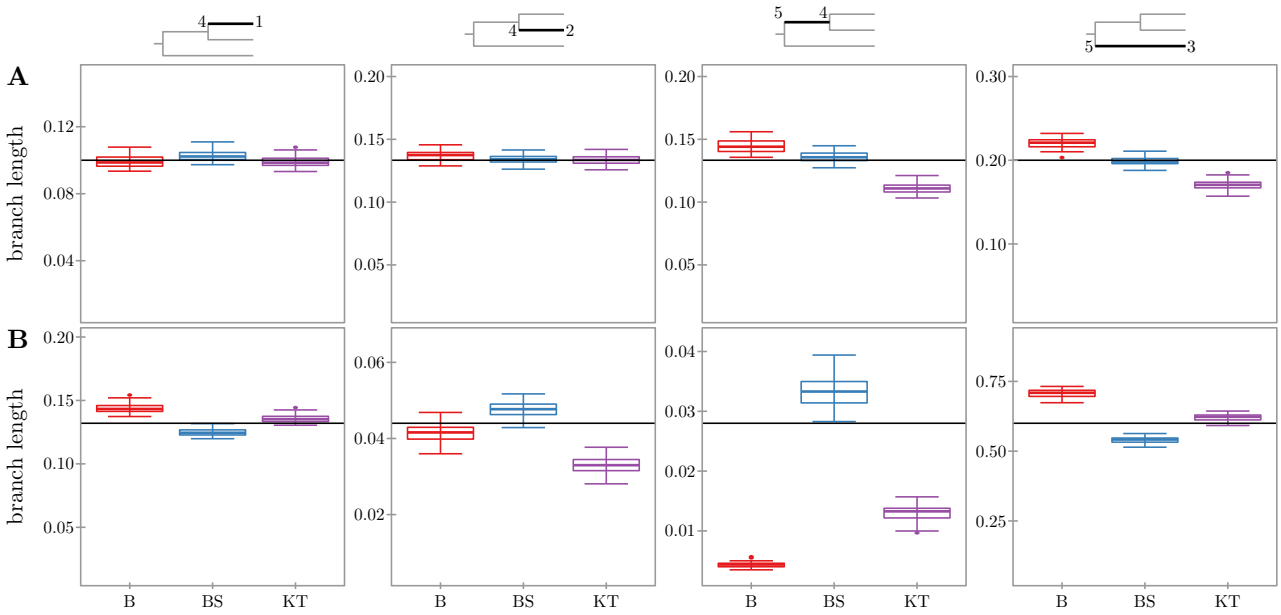


Figure 3: Inference of divergence times for simulation scenarios I (A) and II (B). The figure shows boxplots summarizing the inferred lengths for the four branches of the tree, indicated at the top of each column in black. The inferred lengths are plotted for beta (B), beta with spikes (BS) and Kim Tree (Gautier and Vitalis 2013) (KT). The (true) simulated length of each branch is plotted as a horizontal line. Each plot is scaled relative to the corresponding simulated branch length τ, with the limits of the y-axis set to $[\tau \cdot 0.1, \tau \cdot 1.5]$.

beta approximation (see also supplementary Figure S2). Despite the fact that the spikes probabilities do not perfectly match the true loss and fixation probabilities (Figure 1E and F), this seems to have little effect on the accuracy of branch length estimation for beta with spikes. For both scenarios, the branch leading to population 2 and the inner branch from the root to population 4 have similar lengths, but the beta approximation and Kim Tree provide a worse estimate for the inner branch. This could be due to the fact that there is no data available resulting directly from the evolution on that branch, making the estimation problem harder. A similar result was obtained by Gautier and Vitalis (2013), where trees with the same topology were used. Interestingly, beta with spikes recovers the inner branch much more accurately than either beta and Kim Tree. When measuring the accuracy of the

Table 1: Simulation study scenarios.

	scenario I	scenario II
$(t/2N)_{4\to1}$	$40/(2\cdot200)=0.1$	$132/(2\cdot500)=0.132$
$(t/2N)_{4\to2}$	$40/(2\cdot150)=0.133$	$44/(2\cdot500)=0.044$
$(t/2N)_{5\to4}$	$40/(2\cdot150)=0.133$	$14/(2\cdot250)=0.028$
$(t/2N)_{5\to3}$	$80/(2\cdot200)=0.2$	$300/(2\cdot250)=0.6$
shape parameters of π	1, 1	0.0188, 0.0195
number of SNPs	$5,000$	$10,000$
sample sizes n_{i1}, n_{i2}, n_{i3}	100, 100, 100	22, 24, 12
replicates	50	50

The table indicates the values used for the branch lengths (t), population sizes (N) and scaled branch lengths ($t/2N$), shape parameters of the beta distribution π, the root DAF, number of SNPs and sample sizes used in the two simulation scenarios.

inferred lengths as an average over all four branches (supplementary Table S1), it is clear that beta with spikes outperforms Kim Tree for both scenarios.

Chimpanzee data: The chimpanzee data analyzed here consisted of allele counts of autosomal synonymous SNPs obtained from exome sequencing of the Eastern, Central and Western chimpanzee subspecies (Bataillon *et al.* 2015) for 11, 12 and 6 individuals, respectively. From the original data set containing 59,905 synonymous SNPs, we filtered the SNPs where there was missing data, obtaining a total of 42,063 SNPs. We inferred the scaled branch lengths (Figure 4 and Table 2) using beta, beta with spikes and Kim Tree on the full data set and on 50 smaller data sets containing only 10,000 randomly sampled SNPs. Beta with spikes and Kim Tree infer comparable branch lengths, with the exception of the branch leading to the Western chimpanzee subspecies (population 3). We additionally report in Table 2 the likelihood of the full data calculated using beta with spikes for the branch lengths

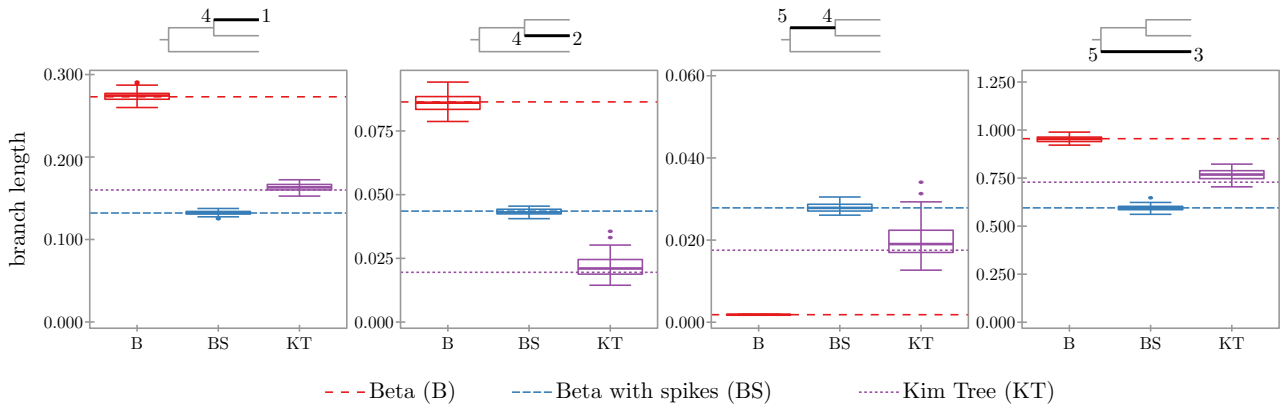


Figure 4: Inference of divergence times for the chimpanzee exome data. The figure shows boxplots summarizing the inferred lengths using 50 data sets with 10,000 SNPs that where randomly sampled from the full data set. The corresponding tree branches are indicated at the top of each plot in black. The inferred lengths are plotted for beta (B), beta with spikes (BS) and Kim Tree (Gautier and Vitalis 2013) (KT). The non-solid lines indicate the inferred lengths when running the methods on the full data set of 42,064 SNPs. The populations at the leaves are: Eastern (1), Central (2) and Western (3). Each plot is scaled relative to the corresponding branch length τ inferred by the beta with spikes on the full data set. The limits of the y-axis are set to $[\tau \cdot 0.05, \tau \cdot 1.5]$.

inferred using the three methods and the ones reported in the original study (Bataillon *et al.* 2015). Bataillon *et al.* (2015) used an ABC approach to fit a demographic model to the synonymous SNPs. Their results are consistent with the ones obtained here for the branches leading to the Eastern (population 1) and Central (population 2) chimpanzees. However, we obtain very different estimates for the remaining two branches. The likelihood in Table 2 indicates that the differences between beta with spikes and beta / Kim Tree / ABC are not merely a result of the numerical optimization being trapped in a local optimum, as the branch lengths obtained by the beta with spikes have the highest likelihood. The discrepancy between the beta with spikes and the ABC results is, perhaps, not surprising, as the difference in inferred branch lengths seems to correlate with the goodness of fit of the ABC

Table 2: Inferred scaled branch lengths for the chimpanzee exome data.

Method	$(t/2N)_{4\to1}$	$(t/2N)_{4\to2}$	$(t/2N)_{5\to4}$	$(t/2N)_{5\to3}$	log L
Beta	0.273	0.086	0.002	0.955	-209646
Beta with spikes	0.132	0.044	0.028	0.595	-204045
Kim Tree	0.160	0.019	0.018	0.729	-205838
ABC (Bataillon *et al.* 2015)	0.183	0.027	0.333	1.914	-233802

The notation follows the one in Figure 2 and the populations (1 to 5) correspond to the ones in Figure 4, with the leaves population: Eastern (1), Central (2) and Western (3). The last column shows the corresponding log likelihood calculated using beta with spikes.

demographic model to the observed data. Bataillon *et al.* (2015) report that their inferred demographic model shows a very good fit for the Central chimpanzees (difference in inferred branch length: 0.017), a slightly less good fit for the Eastern chimpanzees (difference in inferred branch length: 0.051) and a poorer fit for the Western chimpanzees (difference in inferred branch length: 1.319).

DISCUSSION

We have developed a new approximation to the distribution of allele frequency (DAF) as a function of time, conditional on an initial frequency, under a Wright-Fisher model with linear evolutionary pressures. Our work provides an accurate extension of the beta approximation (Balding and Nichols 1995; Balding and Nichols 1997; Sirén *et al.* 2011; Sirén 2012). As noted by Gautier and Vitalis (2013), the beta distribution ignores the possibility of loss or fixation of alleles. We addressed this issue by explicitly modeling the loss and fixation probabilities as two spikes at the boundaries. We showed that the addition of the spikes improves the quality of the approximation and results in more exact inference of divergence times between populations. We expect the beta with spikes to provide a less accurate approximation to the true DAF than the diffusion limit. Nevertheless, we showed that it can infer divergence times

just as accurately as Kim Tree (Gautier and Vitalis 2013), a software built for inference of divergence times using Kimura's solution to the diffusion limit (Kimura 1955).

Computational complexity: The advantage of the beta with spikes becomes more clear when one considers its computational complexity. Diffusion methods rely on heavy computations, such as calculations of Gegenbauer polynomials (Gautier and Vitalis 2013), spectral decomposition of large matrices (Steinrücken *et al.* 2013; Steinrücken *et al.* 2014) or matrix inverse (Zhao *et al.* 2013). In contrast, the beta with spikes requires operations which are performed in constant time per iteration. Perhaps the most expensive evaluation is the beta function used in the loss and fixation probabilities, but very efficient approximations exist for this (Abramowitz and Stegun 1964). The difference in computational complexity is noticeable when comparing the running times of beta with spikes, implemented in `python 2.7`, and Kim Tree, implemented in `Fortran`. For the chimpanzee data set of 42,063 SNPs, beta with spikes ran in just under 5 minutes, while Kim Tree took almost an hour, even though `python 2.7` is a programming language less efficient than `Fortran`. We also note that the two inference methods are inherently different, as here we used a numerical optimization procedure, while Kim Tree uses a Bayesian MCMC approach.

Extensions: We end this section by discussing possible extensions of the beta with spikes approximation and how these can be used in inference problems. Throughout this paper, we assumed that the population size is constant. Due to its recursive formulation, the beta with spikes lends itself naturally to incorporating variable population size, without any increase in computational complexity. This can then be used for inference of population size backwards in time, similar to methods relying on the coalescent with recombination (Li and Durbin 2011; Sheehan *et al.* 2013; Schiffels and Durbin 2014). A recently published method (Liu and Fu 2015) illustrates that allele frequency data, summarized as site frequency spectra, can be efficiently used for inference of variable population size backwards in time. Even

though Liu and Fu (2015) assume sites are independent and do not use linkage information, their method can handle larger data sets than Li and Durbin (2011), which leads to more accurate inference of population sizes for the recent past. The results obtained by Liu and Fu (2015) indicate that the beta with spikes could be successfully used for such demographic inference.

Another extension of the presented approximation would be to incorporate selection, which is a non-linear evolutionary pressure. In the recent years, there has been a great focus on inference of selection coefficients from time-series data under a Wright-Fisher model (Malaspinas *et al.* 2012; Bank *et al.* 2014; Steinrücken *et al.* 2014; Foll *et al.* 2015; Terhorst *et al.* 2015). A newly developed statistical method aims at modeling the evolution of multi-locus alleles under a Wright-Fisher model with selection (Terhorst *et al.* 2015), by fitting a multivariate normal distribution from the first moments of the DAF. Using the approach of Terhorst *et al.* (2015) for moment calculation, the beta with spikes can be extended to non-linear evolutionary pressures. Terhorst *et al.* (2015) do not treat the loss and fixation probabilities. However, as selection is expected to drive allele frequencies towards the boundaries faster than pure drift, including the explicit spikes becomes crucial.

AVAILABILITY

The beta, beta with spikes approximations, inference of divergence times and simulation under a Wright-Fisher model were implemented in `python 2.7`. The code is freely available at `https://github.com/paula-tataru/SpikeyTree`.

ACKNOWLEDGMENTS

It is a pleasure to thank Thomas Mailund for helpful discussions. This work has been supported, in part, by the European Research Council under the European Unions Seventh Framework Program (FP7/20072013, ERC grant number 311341) and the Danish Research Council (grant number DFF–4002-00382).

LITERATURE CITED

Abramowitz, M. and I. A. Stegun, 1964 *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation.

Balding, D. J. and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12.

Balding, D. J. and R. A. Nichols, 1997 Significant genetic correlations among Caucasians at forensic DNA loci. Heredity *78*(6): 583–589.

Bank, C., G. B. Ewing, A. Ferrer-Admettla, M. Foll, and J. D. Jensen, 2014 Thinking too positive? Revisiting current methods of population genetic selection inference. Trends in Genetics *30*(12): 540–546.

Bataillon, T., J. Duan, C. Hvilsom, X. Jin, Y. Li, L. Skov, S. Glemin, K. Munch, T. Jiang, Y. Qian, et al., 2015 Inference of purifying and positive selection in three subspecies of chimpanzees (Pan troglodytes) from exome sequencing. Genome biology and evolution *7*(4): 1122–1132.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics *185*(4): 1411–1423.

Crow, J. and M. Kimura, 1956 Some genetic problems in natural populations. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 4, pp. 1–22. Univ of California Press.

Crow, J. F. and M. Kimura, 1970 *An introduction to population genetics theory.* New York, Evanston and London: Harper & Row, Publishers.

Ewens, W. J., 2004 *Mathematical Population Genetics 1: I. Theoretical Introduction*, Volume 27. Springer Science & Business Media.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood ap-

proach. Journal of molecular evolution *17*(6): 368–376.

Foll, M., H. Shim, and J. D. Jensen, 2015 WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular ecology resources *15*(1): 87–98.

Gautier, M., T. D. Hocking, and J.-L. Foulley, 2010 A Bayesian outlier criterion to detect SNPs under selection in large data sets. PloS one *5*(8): e11913.

Gautier, M. and R. Vitalis, 2013 Inferring population histories using genome-wide allele frequency data. Molecular biology and evolution *30*(3): 654–668.

Gudbjartsson, D. F., H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, et al., 2015 Large-scale whole-genome sequencing of the Icelandic population. Nature genetics *47*(5): 435–444.

Hoban, S., G. Bertorelle, and O. E. Gaggiotti, 2012 Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics *13*(2): 110–122.

Kimura, M., 1955 Solution of a process of random genetic drift with a continuous model. Proceedings of the National Academy of Sciences of the United States of America *41*(3): 144.

Kimura, M., 1964 Diffusion models in population genetics. Journal of Applied Probability *1*(2): 177–232.

Li, H. and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature *475*(7357): 493–496.

Liu, X. and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. Nature genetics 47: 555–559.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012 Estimating allele age and selection coefficient from time-serial data. Genetics *192*(2): 599–607.

McKane, A. and D. Waxman, 2007 Singular solutions of the diffusion equation of population genetics. Journal of Theoretical Biology *247*(4): 849–858.

Nicholson, G., A. V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson, and P. Donnelly, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) *64*(4): 695–715.

Pickrell, J. K. and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics *8*(11): e1002967.

Romiguier, J., P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, et al., 2014 Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature *515*(7526): 261–263.

Rosenberg, N. A. and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics *3*(5): 380–390.

Schiffels, S. and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nature genetics.

Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics *194*(3): 647–662.

Sirén, J., 2012 Statistical models for inferring the structure and history of populations from genetic data. Ph. D. thesis, University of Helsinki, Faculty of Science, Department of Mathematics and Statistics.

Sirén, J., P. Marttinen, and J. Corander, 2011 Reconstructing population histories from single nucleotide polymorphism data. Molecular biology and evolution *28*(1): 673–683.

Song, Y. S. and M. Steinrücken, 2012 A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. Genetics *190*(3): 1117–1129.

Steinrücken, M., A. Bhaskar, and Y. S. Song, 2014 A novel spectral method for inferring general diploid selection from time series genetic data. The annals of applied statistics *8*(4): 2203.

Steinrücken, M., Y. R. Wang, and Y. S. Song, 2013 An explicit transition density expansion for a multi-allelic Wright-Fisher diffusion with general diploid selection. Theoretical population biology 83: 1–14.

Terhorst, J., C. Schlötterer, and Y. S. Song, 2015 Multi-locus analysis of genomic time series data from experimental evolution. PLoS Genetics *11*(4): e1005069.

Waxman, D., 2011 A compact result for the time-dependent probability of fixation at a neutral locus. Journal of Theoretical Biology *274*(1): 131–135.

Wright, S., 1945 The differential equation of the distribution of gene frequencies. Proceedings of the National Academy of Sciences of the United States of America *31*(12): 382.

Zhao, L., X. Yue, and D. Waxman, 2013 Complete Numerical Solution of the Diffusion Equation of Random Genetic Drift. Genetics *194*(4): 973–985.