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

Supplementary Material

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

Contents

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

Conditional mean and variance

If X is a discrete random variable with values between 0 and 1, its mean conditional on $X \notin \{0,1\}$ can be calculated as follows

$$\begin{aligned}
\mathbb{E}\,[\,X \mid X \notin \{0,1\}\,] &= \sum_{x\colon x \notin \{0,1\}} x \cdot \mathbb{P}\,(\,X = x \mid X \notin \{0,1\}\,) \\
&= \sum_{x\colon x \notin \{0,1\}} x \cdot \frac{\mathbb{P}\,(\,X = x\,)}{\mathbb{P}\,(\,X \notin \{0,1\}\,)} \\
&= \frac{1}{\mathbb{P}\,(\,X \notin \{0,1\}\,)} \sum_{x\colon x \notin \{0,1\}} x \cdot \mathbb{P}\,(\,X = x\,) \\
&= \frac{1}{\mathbb{P}\,(\,X \notin \{0,1\}\,)} \left(\mathbb{E}\,[\,X\,] - 0 \cdot \mathbb{P}\,(\,X = 0\,) - 1 \cdot \mathbb{P}\,(\,X = 1\,)\right) \\
&= \frac{\mathbb{E}\,[\,X\,] - \mathbb{P}\,(\,X = 1\,)}{\mathbb{P}\,(\,X \notin \{0,1\}\,)}.
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
\mathbb{E}\left[\,X^2 \mid X \notin \{0,1\}\,\right] &= \frac{\mathbb{E}\,[\,X^2\,] - \mathbb{P}\,(\,X = 1\,)}{\mathbb{P}\,(\,X \notin \{0,1\}\,)} \\
&= \frac{\operatorname{Var}\,(\,X\,) + \mathbb{E}\,[\,X\,]^2 - \mathbb{P}\,(\,X = 1\,)}{\mathbb{P}\,(\,X \notin \{0,1\}\,)},
\end{aligned}$$

from which

$$\begin{aligned}
\operatorname{Var}\,(\,X \mid X \notin \{0,1\}\,) &= \mathbb{E}\left[\,X^2 \mid X \notin \{0,1\}\,\right] - \mathbb{E}\,[\,X \mid X \notin \{0,1\}\,]^2 \\
&= \frac{\operatorname{Var}\,(\,X\,) + \mathbb{E}\,[\,X\,]^2 - \mathbb{P}\,(\,X = 1\,)}{\mathbb{P}\,(\,X \notin \{0,1\}\,)} - \mathbb{E}\,[\,X \mid X \notin \{0,1\}\,]^2\,.
\end{aligned}$$

Derivation of mean and variance of X_t

To calculate the mean and variance of X_t under the Wright-Fisher model, we rely on the laws of total mean and variance, respectively. Recall that $X_t = Z_t/2N$ and

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

where $x_t = z_t/2N$. The evolutionary pressures $g(x)$ verify that $0 \leq g(x) \leq 1$ for all $0 \leq x \leq 1$. In the following, g is a linear function in the allele frequency, $g(x) = (1-a)\,x + b$. The parameters a and b verify that $0 \leq b \leq a < 1$ and typically, $a << 1$. We note that if $a = 0$, then $b = 0$. In the derivations below, we condition implicitly on $X_0 = x_0$, population size $2N$ and evolutionary pressures.

Let us start with the mean and variance of X_{t+1} conditional on $X_t = x_t$, given by

$$\begin{aligned}
\mathbb{E}\,[\,X_{t+1} \mid X_t = x_t\,] &= \frac{1}{2N}\,\mathbb{E}\,[\,Z_{t+1} \mid Z_t = z_t\,] \\
&= \frac{1}{2N}\,2N\,g(x_t) \\
&= g(x_t), \\
\mathrm{Var}\,(\,X_{t+1} \mid X_t = x_t\,) &= \frac{1}{4N^2}\,\mathrm{Var}\,(\,Z_{t+1} \mid Z_t = z_t\,) \\
&= \frac{1}{4N^2}\,2N\,g(x_t)\,(1 - g(x_t)) \\
&= \frac{1}{2N}\,g(x_t)\,(1 - g(x_t)).
\end{aligned}$$

First, using the law of total expectation, we have that

$$\begin{aligned}
\mathbb{E}\,[\,X_t\,] &= \mathbb{E}\,[\,\mathbb{E}\,[\,X_t \mid X_{t-1}\,]\,] \\
&= \mathbb{E}\,[\,g(X_{t-1})\,] \\
&= \mathbb{E}\,[\,(1-a)\,X_{t-1} + b\,] \\
&= (1-a)\,\mathbb{E}\,[\,X_{t-1}\,] + b \\
&= (1-a)\,\mathbb{E}\,[\,\mathbb{E}\,[\,X_{t-1} \mid X_{t-2}\,]\,] + b \\
&= \dots \\
&= (1-a)^t\,x_0 + b\sum_{i=0}^{t-1}(1-a)^i.
\end{aligned}$$

When $a = b = 0$, the mean becomes

$$\mathbb{E}\,[\,X_t\,] = x_0.$$

If $a \neq 0$,

$$\sum_{i=0}^{t-1}(1-a)^i = \frac{1-(1-a)^t}{a},$$

and this gives

$$\mathbb{E}\left[\, X_t \,\right] = \frac{b}{a} + (1-a)^t \left(x_0 - \frac{b}{a} \right).$$

We use a similar approach to determine the variance of X_t, this time relying on the law of total variance

$$\begin{aligned}
\operatorname{Var}(\, X_t \,) &= \mathbb{E}\left[\, \operatorname{Var}(\, X_t \mid X_{t-1} \,) \,\right] + \operatorname{Var}(\, \mathbb{E}\left[\, X_t \mid X_{t-1} \,\right] \,) \\
&= \mathbb{E}\left[\frac{1}{2N}\, g(X_{t-1})\,(1 - g(X_{t-1})) \right] + \operatorname{Var}(\, g(X_{t-1}) \,) \\
&= \frac{1}{2N}\,\mathbb{E}\left[\, g(X_{t-1}) \,\right] - \frac{1}{2N}\,\mathbb{E}\left[\, g(X_{t-1})^2 \,\right] + \operatorname{Var}(\, g(X_{t-1}) \,) \\
&= \frac{1}{2N}\,\mathbb{E}\left[\, g(X_{t-1}) \,\right] - \frac{1}{2N}\operatorname{Var}(\, g(X_{t-1}) \,) - \frac{1}{2N}\,\mathbb{E}\left[\, g(X_{t-1}) \,\right]^2 + \operatorname{Var}(\, g(X_{t-1}) \,) \\
&= \frac{1}{2N}\,\mathbb{E}\left[\, g(X_{t-1}) \,\right](1 - \mathbb{E}\left[\, g(X_{t-1}) \,\right]) + \left(1 - \frac{1}{2N}\right)\operatorname{Var}(\, g(X_{t-1}) \,) \\
&= \frac{1}{2N}\,\mathbb{E}\left[\, X_t \,\right](1 - \mathbb{E}\left[\, X_t \,\right]) + \left(1 - \frac{1}{2N}\right)(1-a)^2 \operatorname{Var}(\, X_{t-1} \,).
\end{aligned}$$

Iterating the above,

$$\operatorname{Var}(\, X_t \,) = \frac{1}{2N}\sum_{i=1}^{t}(1-a)^{2(t-i)}\left(1 - \frac{1}{2N}\right)^{t-i}\mathbb{E}\left[\, X_i \,\right](1 - \mathbb{E}\left[\, X_i \,\right]).$$

Let us observe that, for any c,

$$\begin{aligned}
\frac{1}{2N}\sum_{i=1}^{t}(1-a)^{c\,(t-i)}\left(1 - \frac{1}{2N}\right)^{t-i} &= \frac{1}{2N}\cdot\frac{1-(1-a)^{c\,t}\left(1-\frac{1}{2N}\right)^t}{1-(1-a)^c\left(1-\frac{1}{2N}\right)} \\
&= \frac{1-(1-a)^{c\,t}\left(1-\frac{1}{2N}\right)^t}{2N-(1-a)^c\,(2N-1)}.
\end{aligned}$$

When $a = b = 0$ and using $c = 0$ in the above, the variance becomes

$$\operatorname{Var}(\, X_t \,) = x_0(1-x_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right].$$

If $a \neq 0$,

$$\begin{aligned}
\mathbb{E}\left[\, X_i \,\right](1 - \mathbb{E}\left[\, X_i \,\right]) = \frac{b}{a}&\left(1 - \frac{b}{a}\right) \\
&+ \left(1 - \frac{2b}{a}\right)(1-a)^i\left(x_0 - \frac{b}{a}\right) \\
&- (1-a)^{2i}\left(x_0 - \frac{b}{a}\right)^2,
\end{aligned}$$

and using $c = 1$ and $c = 2$, respectively, we obtain the variance

$$\begin{aligned}
\operatorname{Var}\left(X_t \right) = {} & \frac{b}{a} \left(1 - \frac{b}{a}\right) \frac{1}{2N} \sum_{i=1}^{t} (1-a)^{2(t-i)} \left(1 - \frac{1}{2N}\right)^{t-i} \\
& + \left(1 - \frac{2b}{a}\right) \left(x_0 - \frac{b}{a}\right) (1-a)^t \frac{1}{2N} \sum_{i=1}^{t} (1-a)^{t-i} \left(1 - \frac{1}{2N}\right)^{t-i} \\
& - \left(x_0 - \frac{b}{a}\right)^2 (1-a)^{2t} \frac{1}{2N} \sum_{i=1}^{t} \left(1 - \frac{1}{2N}\right)^{t-i} \\
= {} & \frac{b}{a} \left(1 - \frac{b}{a}\right) \frac{1 - (1-a)^{2t} \left(1 - \frac{1}{2N}\right)^t}{2N - (1-a)^2 \left(2N - 1\right)} \\
& + \left(1 - \frac{2b}{a}\right) \left(x_0 - \frac{b}{a}\right) (1-a)^t \frac{1 - (1-a)^t \left(1 - \frac{1}{2N}\right)^t}{2N - (1-a) \left(2N - 1\right)} \\
& - \left(x_0 - \frac{b}{a}\right)^2 (1-a)^{2t} \left(1 - \left(1 - \frac{1}{2N}\right)^t\right).
\end{aligned}$$

See the parameter scaling section for a comparison with the derivations obtained by Sirén (2012). We note that Sirén (2012) relies on approximations resulting from the infinite population limit, while the above equations hold for any population size.

The derivations for the mean and variance use the linearity of the evolutionary pressures through the simplification that

$$\begin{aligned}
\mathbb{E}\left[(1-a)\, X_t + b \right] &= (1-a)\, \mathbb{E}\left[X_t \right] + b, \\
\operatorname{Var}\left((1-a)\, X_t + b \right) &= (1-a)^2 \operatorname{Var}\left(X_t \right).
\end{aligned}$$

When $g(x)$ is a polynomial of higher order, such as in the case of selection, the derivation requires higher moments of X_t, leading to an explosion in the moments needed and rendering the above approach untractable in such situations.

Derivation of loss and fixation probabilities of X_t

To determine $\mathbb{P}\left(X_{t+1} = 0 \right)$ and $\mathbb{P}\left(X_{t+1} = 1 \right)$, we use the law of total probability in an approach similar to the above. Additionally, we rely on the approximation that X_t follows a known $f_B^\star$ beta with spikes distribution to obtain

$$\begin{aligned}
\mathbb{P}\,(\,X_{t+1}=0\,) &= \int_0^1 \mathbb{P}\,(\,X_{t+1}=0 \mid X_t = x\,)\cdot f_B^\star(x;t)\,\mathrm{d}x \\
&= \mathbb{P}\,(\,X_{t+1}=0 \mid X_t=0\,)\cdot\mathbb{P}\,(\,X_t=0\,)+\mathbb{P}\,(\,X_{t+1}=0 \mid X_t=1\,)\cdot\mathbb{P}\,(\,X_t=1\,) \\
&\qquad + \mathbb{P}\,(\,X_t\notin\{0,1\}\,)\cdot\int_0^1 \mathbb{P}\,(\,X_{t+1}=0 \mid X_t=x\,)\cdot\frac{x^{\alpha_t^\star-1}(1-x)^{\beta_t^\star-1}}{\mathrm{B}\,(\,\alpha_t^\star,\beta_t^\star\,)}\,\mathrm{d}x \\
&= \mathbb{P}\,(\,X_t=0\,)\cdot(1-g(0))^{2N}+\mathbb{P}\,(\,X_t=1\,)\cdot(1-g(1))^{2N} \\
&\qquad + \mathbb{P}\,(\,X_t\notin\{0,1\}\,)\cdot\int_0^1 (1-g(x))^{2N}\cdot\frac{x^{\alpha_t^\star-1}(1-x)^{\beta_t^\star-1}}{\mathrm{B}\,(\,\alpha_t^\star,\beta_t^\star\,)}\,\mathrm{d}x,
\end{aligned}$$

where $\mathrm{B}\,(\,\alpha,\beta\,)$ is the beta function.

To calculate the above integral for linear evolutionary pressures, we rely on the hypergeometric function. Let ${}_2F_1(-m,b;c;z)$ ((Erdélyi *et al.* 1953), 2.1.3) be the hypergeometric function for $m\in\mathbb{N}$, $c,d\in\mathbb{R}_+$ and $z\in\mathbb{R}$, given by

$${}_2F_1(-m,c;c+d;z)=\frac{1}{\mathrm{B}\,(\,c,d\,)}\int_0^1 x^{c-1}(1-x)^{d-1}(1-z\,x)^m\,\mathrm{d}x.$$

We have that (recall that $0\le b<1$)

$$\begin{aligned}
&\int_0^1 (1-g(x))^{2N}\cdot\frac{x^{\alpha_t^\star-1}(1-x)^{\beta_t^\star-1}}{\mathrm{B}\,(\,\alpha_t^\star,\beta_t^\star\,)}\,\mathrm{d}x \\
&\qquad = \frac{1}{\mathrm{B}\,(\,\alpha_t^\star,\beta_t^\star\,)}\int_0^1 ((1-b)-(1-a)\,x)^{2N}x^{\alpha_t^\star-1}(1-x)^{\beta_t^\star-1}\,\mathrm{d}x \\
&\qquad = \frac{(1-b)^{2N}}{\mathrm{B}\,(\,\alpha_t^\star,\beta_t^\star\,)}\int_0^1\left(1-\frac{1-a}{1-b}\,x\right)^{2N}x^{\alpha_t^\star-1}(1-x)^{\beta_t^\star-1}\,\mathrm{d}x \\
&\qquad = (1-b)^{2N}\,{}_2F_1\left(-2N,\alpha_t^\star;\alpha_t^\star+\beta_t^\star;\frac{1-a}{1-b}\right),
\end{aligned}$$

leading to the full expression for the loss probability

$$\begin{aligned}
\mathbb{P}\,(\,X_{t+1}=0\,) &= \mathbb{P}\,(\,X_t=0\,)\cdot(1-b)^{2N}+\mathbb{P}\,(\,X_t=1\,)\cdot(a-b)^{2N} \\
&\qquad + \mathbb{P}\,(\,X_t\notin\{0,1\}\,)\cdot(1-b)^{2N}\cdot{}_2F_1\left(-2N,\alpha_t^\star;\alpha_t^\star+\beta_t^\star;\frac{1-a}{1-b}\right).
\end{aligned}$$

Similarly, for $b \neq 0$ (if $b = 0$, see below), we obtain the fixation probability

$$\begin{aligned}\mathbb{P}\,(\,X_{t+1} = 1\,) = \mathbb{P}\,(\,X_t = 0\,) \cdot b^{2N} + \mathbb{P}\,(\,X_t = 1\,) \cdot (1-a+b)^{2N} \\ + \mathbb{P}\,(\,X_t \notin \{0,1\}\,) \cdot b^{2N} \cdot {}_2F_1\left(-2N, \alpha_t^\star; \alpha_t^\star + \beta_t^\star; -\frac{1-a}{b}\right).\end{aligned}$$

Approximation for small a and b The hypergeometric function can be cumbersome and slow to evaluate. Typically the parameters a and b are small and we can use that

$$\begin{aligned} 1 - g(x) &= (1-a)(1-x) + a - b && \approx (1-a)(1-x), \\ g(x) &= (1-a)\,x + b && \approx (1-a)\,x, \end{aligned}$$

to more easily reduce the above integrals to

$$\begin{aligned} \int_0^1 (1-g(x))^{2N} \cdot \frac{x^{\alpha_t^\star - 1}(1-x)^{\beta_t^\star - 1}}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)}\,\mathrm{d}x &\approx (1-a)^{2N} \cdot \int_0^1 \frac{x^{\alpha_t^\star - 1}(1-x)^{\beta_t^\star + 2N - 1}}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)}\,\mathrm{d}x \\ &= (1-a)^{2N} \cdot \frac{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star + 2N\,)}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)}, \\ \int_0^1 g(x)^{2N} \cdot \frac{x^{\alpha_t^\star - 1}(1-x)^{\beta_t^\star - 1}}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)}\,\mathrm{d}x &\approx (1-a)^{2N} \cdot \frac{\mathrm{B}\,(\,\alpha_t^\star + 2N, \beta_t^\star\,)}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)}, \end{aligned}$$

from which

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

For the results reported in the main text and below (in numerical accuracy and inference of divergence times sections), we used the above approximation for small a and b.

Approximation for large N A widely used assumption in the derivations based on the Wright-Fisher model, such as the diffusion limit, is that the population size N is large, and a and b are small such that

$$\lim_{N\to\infty} 2Na = A, \qquad \lim_{N\to\infty} 2Nb = B.$$

Additionally, the time is scaled by the population size, $\tau = t/2N$. We set $\Delta = 1/2N$. Because a and b are small, we build on the previous approximation.

Let $\Gamma(c)$ be the Gamma function and note that, for large N ((Erdélyi *et al.* 1953), 1.18),

$$\frac{\Gamma(N+c)}{\Gamma(N+c+d)} \approx \left(\frac{1}{N}\right)^{d} \left(1 - \frac{d\,(c+2d-1)}{2N}\right).$$

We then have

$$\begin{aligned}
\frac{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star + 2N\,)}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)} &= \frac{\Gamma(\alpha_t^\star)\,\Gamma(\beta_t^\star + 2N)}{\Gamma(\alpha_t^\star + \beta_t^\star + 2N)} \cdot \frac{\Gamma(\alpha_t^\star + \beta_t^\star)}{\Gamma(\alpha_t^\star)\,\Gamma(\beta_t^\star)} \\
&= \frac{\Gamma(\alpha_t^\star + \beta_t^\star)}{\Gamma(\beta_t^\star)} \cdot \frac{\Gamma(2N + \beta_t^\star)}{\Gamma(2N + \alpha_t^\star + \beta_t^\star)} \\
&\approx \frac{\Gamma(\alpha_t^\star + \beta_t^\star)}{\Gamma(\beta_t^\star)} \cdot \left(\frac{1}{2N}\right)^{\alpha_t^\star} \cdot \left(1 - \frac{\alpha_t^\star\,(2\alpha_t^\star + \beta_t^\star - 1)}{4N}\right) \\
&= \frac{\Gamma(\alpha_t^\star + \beta_t^\star)}{\Gamma(\beta_t^\star)} \cdot \Delta^{\alpha_t^\star} \cdot \left(1 - \frac{1}{2}\Delta\,\alpha_t^\star\,(2\alpha_t^\star + \beta_t^\star - 1)\right),
\end{aligned}$$

and, similarly,

$$\frac{\mathrm{B}\,(\,\alpha_t^\star + 2N, \beta_t^\star\,)}{\mathrm{B}\,(\,\alpha_t^\star, \beta_t^\star\,)} \approx \frac{\Gamma(\alpha_t^\star + \beta_t^\star)}{\Gamma(\alpha_t^\star)} \cdot \Delta^{\beta_t^\star} \cdot \left(1 - \frac{1}{2}\Delta\,\beta_t^\star\,(\alpha_t^\star + 2\beta_t^\star - 1)\right).$$

Using that

$$\begin{aligned}
\lim_{N\to\infty} (1-a)^{2N} &= e^{-A}, & \lim_{N\to\infty} (1-b)^{2N} &= e^{-B}, \\
\lim_{N\to\infty} (1-a+b)^{2N} &= e^{-(A-B)}, & \lim_{N\to\infty} (a-b)^{2N} &= 0,
\end{aligned}$$

we obtain the recursion in scaled time for loss and fixation probabilities to be

$$\begin{aligned}
\mathbb{P}\,(\,X_{\tau+\Delta} = 0\,) \approx{}& \mathbb{P}\,(\,X_\tau = 0\,) \cdot e^{-B} \\
&+ \mathbb{P}\,(\,X_\tau \notin \{0,1\}\,) \cdot e^{-A} \cdot \frac{\Gamma(\alpha_\tau^\star + \beta_\tau^\star)}{\Gamma(\beta_\tau^\star)} \cdot \Delta^{\alpha_\tau^\star} \cdot \left(1 - \frac{1}{2}\Delta\,\alpha_\tau^\star\,(2\alpha_\tau^\star + \beta_\tau^\star - 1)\right), \\
\mathbb{P}\,(\,X_{\tau+\Delta} = 1\,) \approx{}& \mathbb{P}\,(\,X_\tau = 1\,) \cdot e^{-(A-B)} \\
&+ \mathbb{P}\,(\,X_\tau \notin \{0,1\}\,) \cdot e^{-A} \cdot \frac{\Gamma(\alpha_\tau^\star + \beta_\tau^\star)}{\Gamma(\alpha_\tau^\star)} \cdot \Delta^{\beta_\tau^\star} \cdot \left(1 - \frac{1}{2}\Delta\,\beta_\tau^\star\,(\alpha_\tau^\star + 2\beta_\tau^\star - 1)\right).
\end{aligned}$$

Parameter scaling

As noted above, one common assumption is that the population size N is large and a and b are small. One central result of the diffusion limit is that the allele frequency distribution is entirely determined by the scaled time $\tau = t/2N$ and parameters $A = 2Na$ and $B = 2Nb$ (Ewens 2004). The same holds for the beta distribution. Using that

$$\begin{aligned} (1-a)^t &\approx e^{-A\tau}, & 2N - (1-a)(2N-1) &\approx 1 + A, \\ \left(1 - \frac{1}{2N}\right)^t &\approx e^{-\tau}, & 2N - (1-a)^2(2N-1) &\approx 1 + 2A, \end{aligned}$$

we obtain the mean and variance as a function of the scaled parameters to be

$$\mathbb{E}\left[\, X_\tau \,\right] = \begin{cases} x_0 & \text{if } a = b = 0, \\ \frac{B}{A} + e^{-A\tau}\left(x_0 - \frac{B}{A}\right) & \text{otherwise,} \end{cases}$$

$$\operatorname{Var}\left(\, X_\tau \,\right) = \begin{cases} x_0(1-x_0)\left(1 - e^{-\tau}\right) & \text{if } a = b = 0, \\ \frac{B}{A}\left(1 - \frac{B}{A}\right)\frac{1-e^{-(2A+1)\tau}}{1+2A} & \\ \qquad + \left(1 - \frac{2B}{A}\right)\left(x_0 - \frac{B}{A}\right)e^{-A\tau}\frac{1-e^{-(A+1)\tau}}{1+A} & \text{otherwise.} \\ \qquad - \left(x_0 - \frac{B}{A}\right)^2 e^{-2A\tau}\left(1 - e^{-\tau}\right) & \end{cases}$$

The above equations are equivalent to the ones by Sirén (2012) (up to some minor typographical errors, as confirmed by correspondence with the author).

The same property holds for the beta with spikes, as shown in the above derivation for large N, where the loss and fixation probability are written as functions of the scaled parameters.

Discretization of beta and beta with spikes

For the presented results, the beta and beta with spikes distributions need to be discretized in $K+1$ bins. We chose bins that, expect for the first and last bin, are centered around $\frac{k}{K}$ for $1 \leq k \leq K-1$, given by

$$\left[0, \frac{1}{2K}\right], \left[\frac{1}{2K}, \frac{3}{2K}\right], \ldots, \left[\frac{2k-1}{2K}, \frac{2k+1}{2K}\right], \ldots, \left[\frac{2K-3}{2K}, \frac{2K-1}{2K}\right], \left[\frac{2K-1}{2K}, 1\right],$$

Next to the $K+1$ probabilities corresponding to each bin, the beta with spikes distribution contains two extra spike probabilities for 0 and 1.

Numerical accuracy of the beta and beta with spikes models

To investigate how well the beta with spikes approximates the true distribution of allele frequency (DAF) and, in particular, if it provides a better approximation than the beta distribution, we compared the two with the DAF calculated directly from the Wright-Fisher. For this purpose, we discretized the approximated distributions using $K = 2N$. This leads to a unique mapping between the true discrete allele frequencies $k/2N$, $0 \leq k \leq 2N$ and the bins. As the first and last bins correspond to frequencies 0 and 1, respectively, and the beta with spikes contains explicit probabilities for these two frequencies, we merged the first and last two bins to $\left[0, \frac{3}{4N}\right]$ and $\left[\frac{4N-3}{4N}, 1\right]$ for calculating the discrete probability for frequencies $\frac{1}{2N}$ and $\frac{2N-1}{2N}$, respectively.

We used a population size $2N = 200$ and for a range of initial frequencies x_0, times t and parameters a and b, we calculated the Hellinger distance between the true and approximated distributions. For two discrete distributions $P = (p_1, \ldots, p_k)$ and $Q = (q_1, \ldots, q_k)$, the Hellinger distance is given by

$$h(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}.$$

The Hellinger distance lies between 0 and 1, with 0 indicating perfect match between the two distributions, while the value of 1 is achieved when P assigns probability zero to every set where Q assigns a positive probability, and vice versa.

The Hellinger distance for the beta and beta with spikes is given in Figure S1 and shows that the beta with spikes provides a better approximation than the beta, for the whole considered range of parameter values, initial frequencies and times. It is apparent from the figure that the beta distribution approximates well the true DAF when this is not close to the boundaries: either the initial frequency is close to 0.5 and the time is not too large, or

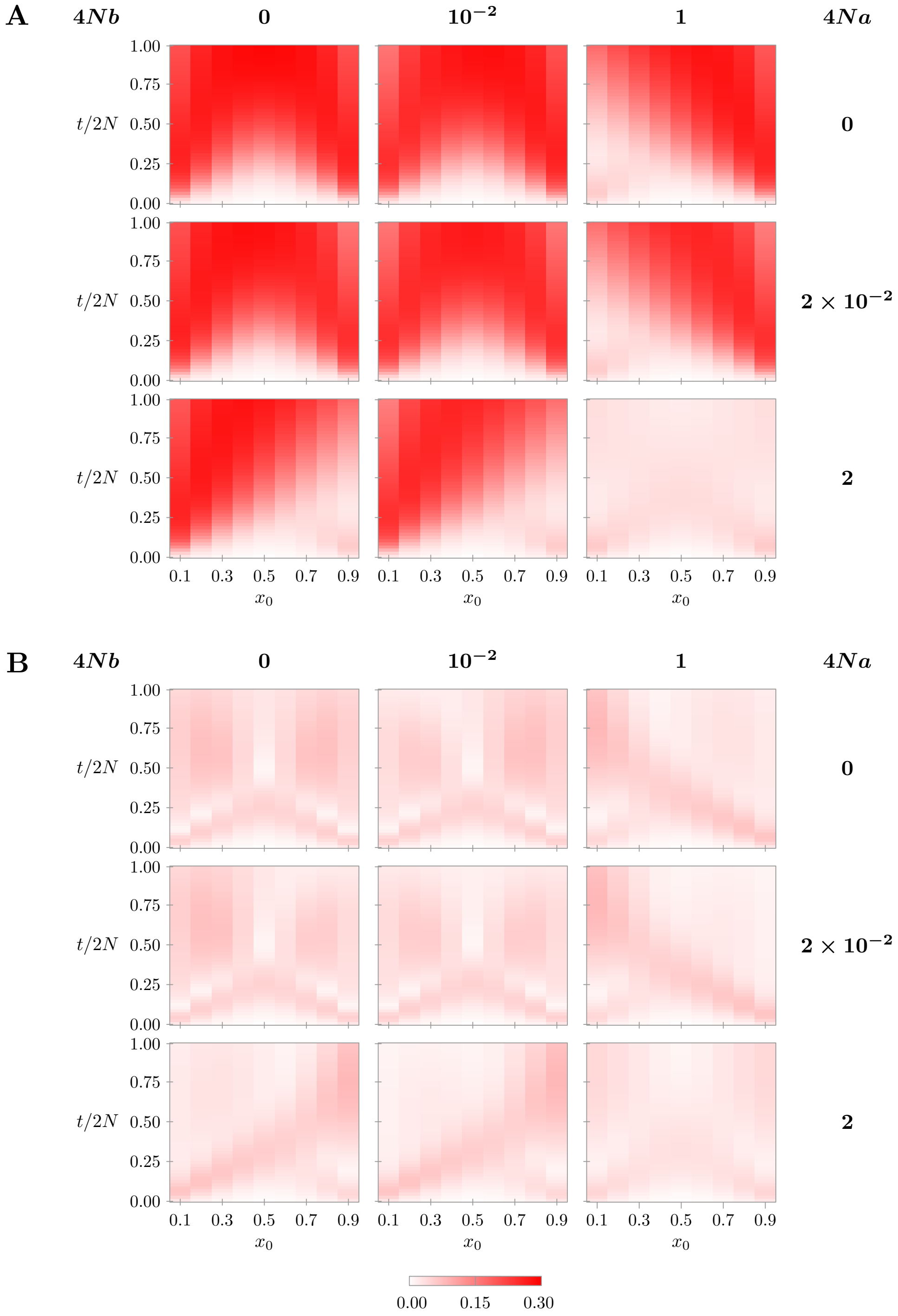

Figure S1: Numerical accuracy. The heatmaps show the Hellinger distance between the true DAF and the beta (A) and beta with spikes (B) as a function of x_0 and $t/2N$. Each row and column corresponds to specific values of the scaled parameters $4Na$ and $4Nb$.

the parameters a or b are large enough to keep the allele frequency away from 0 and 1. The beta with spikes has a more consistent behavior, as the corresponding Hellinger distance does not vary as much as for the beta distribution.

Likelihood calculation on a tree

The probability of observing the data for a given tree is a function of the scaled branch lengths, denoted here as Θ, and π, the DAF at the root. We are interested in calculating the likelihood $L(\mathcal{D}; \Theta, \pi)$ of the data $\mathcal{D} = \{(z_{ij}, n_{ij}) \mid 1 \leq i \leq I, 1 \leq j \leq J\}$, for I SNPs in J populations. The SNPs are assumed to be realizations of independent and identically distributed random variables (of dimension J). Then the full likelihood of the data can be written as a product over the independent sites

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

where $\mathcal{D}_i = \{(z_{ij}, n_{ij}) \mid 1 \leq j \leq J\}$ is the observed data for SNP i. We therefore present below how to calculate the likelihood for one SNP and, for notation simplicity, drop the index i.

Full data Assuming Hardy-Weinberg equilibrium, the probability of observing z alleles in a sample of size n, given the population allele frequency x, follows from the binomial distribution

$$\mathbb{P}\left(z \mid n, x \right) = \binom{n}{z} x^z \left(1 - x\right)^{n-z}.$$

However, the allele frequencies x_j, $1 \leq j \leq J$, are unobserved and the likelihood of the data is obtained by integrating over the unknown allele frequencies

$$L(\mathcal{D}; \Theta, \pi) = \int_0^1 \int_0^1 \dots \int_0^1 f(X_1, X_2, \dots, X_J \mid \Theta, \pi) \cdot \prod_{j=1}^{J} \mathbb{P}\left(z_j \mid n_j, X_j \right) \, \mathrm{d}X_1 \, \mathrm{d}X_2 \dots \mathrm{d}X_J,$$

where $f(X_1, X_2, \ldots, X_J \mid \Theta, \pi)$ is the joint distribution of the X_j's at the leaves. The joint distribution is, in turn, an integral over the allele frequencies in the ancestral populations, represented as internal nodes in the tree. To calculate the likelihood and the joint distribution, we discretize the allele frequencies in $K+1$ bins as detailed above. Let bin number $0 \leq k \leq K$ from before be $[l_k, u_k]$ (i.e. $l_k = \max\{0,\, 2k-1\}/2K$ and $u_k = \min\{2k+1,\, 1\}/2K$). Then, for each branch length $t/2N$ we can calculate the discrete transition probabilities as

$$\mathbb{P}\left(\, X_j \in [l_k, u_k] \mid X_l = k_0/K,\, t/2N \,\right) = \int_{l_k}^{u_k} f(x; k_0/K, t, N)\, \mathrm{d}x,$$

where $f(x; k_0/K, t, N)$ is the DAF over t generations in a population of size $2N$, conditional on a initial frequency k_0/K, $0 \leq k_0 \leq K$. The distribution f is replaced by either f_B for the beta, or $f_B^\star$ for the beta with spikes. When using the beta with spikes, we use two additional probabilities for $X_j = 0$ and $X_j = 1$. With these transition probabilities at hand, we can efficiently calculate the joint distribution using a peeling algorithm (Felsenstein 1981).

For the tree depicted in Figure 2, $\Theta = \Big((t/2N)_{5\to3}, (t/2N)_{5\to4}, (t/2N)_{4\to1}, (t/2N)_{4\to2}\Big)$ and conditional on the allele frequency in the ancestral population (at the root) to be k_5/K, we obtain

$$\begin{aligned} L(\mathcal{D}; \Theta \mid k_5/K) = \;& \left(\sum_{k_3=0}^{K} \mathbb{P}\left(\, X_3 \in [l_{k_3}, u_{k_3}] \mid X_5 = k_5/K,\, (t/2N)_{5\to3} \,\right) \mathbb{P}\left(\, z_3 \mid n_3, k_3 \,\right) \right) \\ & \cdot \left(\sum_{k_4=0}^{K} \mathbb{P}\left(\, X_4 \in [l_{k_4}, u_{k_4}] \mid X_5 = k_5/K,\, (t/2N)_{5\to4} \,\right) \right. \\ & \qquad \cdot \left(\sum_{k_2=0}^{K} \mathbb{P}\left(\, X_2 \in [l_{k_2}, u_{k_2}] \mid X_4 = k_4/K,\, (t/2N)_{4\to2} \,\right) \mathbb{P}\left(\, z_2 \mid n_2, k_2 \,\right) \right) \\ & \qquad \left. \cdot \left(\sum_{k_1=0}^{K} \mathbb{P}\left(\, X_1 \in [l_{k_1}, u_{k_1}] \mid X_4 = k_4/K,\, (t/2N)_{4\to1} \,\right) \mathbb{P}\left(\, z_1 \mid n_1, k_1 \,\right) \right) \right), \end{aligned}$$

and the full likelihood is obtained by summing over all possible ancestral frequencies

$$L(\mathcal{D}; \Theta, \pi) = \sum_{k_5=0}^{K} \pi(k_5/K)\, L(\mathcal{D}; \Theta \mid k_5/K).$$

The sums are slightly different when using beta with spikes, in order to correctly account for the loss and fixation probabilities.

Due to the binning, the above calculation provides an approximation which converges to the true likelihood as K increases.

Polymorphic data The above likelihood calculation assumes that the data contains both sites that are polymorphic, and sites that are fixed or lost in all populations. However, SNP data is restricted to polymorphic sites. We can calculate the likelihood of the data conditional on observing only polymorphic sites as follows

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

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

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

$$\mathcal{D}^0 = \{(0,n_j) \mid 1 \le j \le J\}, \qquad \mathcal{D}^1 = \{(n_j,n_j) \mid 1 \le j \le J\}.$$

The DAF at the root Let us assume that the DAF at the root is a beta with spikes distribution, with the sum of the spikes equal to p_{mono} (i.e. the probability that the allele has either frequency 0 or 1). Let π denote the beta distribution over $(0,1)$. In the following, we show that the likelihood conditional on polymorphic data is independent of p_{mono}.

We note that the probability of observing polymorphic data is zero if the allele frequency at the root is 0 or 1

$$L(\mathcal{D};\Theta \mid 0) = L(\mathcal{D};\Theta \mid 1) = 0,$$

from which

$$L(\mathcal{D};\Theta,\pi,p_{\text{mono}}) = (1-p_{\text{mono}}) \sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D};\Theta \mid k_5/K),$$

Similarly, for the unobserved monomorphic data we have that

$$L(\mathcal{D}^0;\Theta \mid 0) = L(\mathcal{D}^1;\Theta \mid 1) = 1, \qquad L(\mathcal{D}^0;\Theta \mid 1) = L(\mathcal{D}^1;\Theta \mid 0) = 0,$$

from which

$$
\begin{aligned}
&L(\mathcal{D}^0; \Theta, \pi, p_{\text{mono}}) + L(\mathcal{D}^1; \Theta, \pi, p_{\text{mono}}) \\
&\quad = p_{\text{mono}} + (1 - p_{\text{mono}}) \left(\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^0; \Theta \mid k_5/K) + \sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^1; \Theta \mid k_5/K) \right),
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\, \text{polymorphism} \mid \Theta, \pi, p_{\text{mono}} \,\right) \\
&\quad = 1 - L(\mathcal{D}^0; \Theta, \pi) - L(\mathcal{D}^1; \Theta, \pi) \\
&\quad = (1 - p_{\text{mono}}) \left(1 - \sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^0; \Theta \mid k_5/K) - \sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^1; \Theta \mid k_5/K) \right).
\end{aligned}
$$

Using the above, we obtain the likelihood conditional on polymorphism to be

$$
\begin{aligned}
&L(\mathcal{D}; \Theta, \pi, p_{\text{mono}} \mid \text{polymorphism}) \\
&\quad = \frac{L(\mathcal{D}; \Theta, \pi, p_{\text{mono}})}{\mathbb{P}\left(\, \text{polymorphism} \mid \Theta, \pi, p_{\text{mono}} \,\right)} \\
&\quad = \frac{(1 - p_{\text{mono}}) \displaystyle\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}; \Theta \mid k_5/K)}{(1 - p_{\text{mono}}) \left(1 - \displaystyle\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^0; \Theta \mid k_5/K) - \displaystyle\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^1; \Theta \mid k_5/K) \right)} \\
&\quad = \frac{\displaystyle\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}; \Theta \mid k_5/K)}{1 - \displaystyle\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^0; \Theta \mid k_5/K) - \displaystyle\sum_{k_5=1}^{K-1} \pi(k_5/K)\, L(\mathcal{D}^1; \Theta \mid k_5/K)} \\
&\quad = L(\mathcal{D}; \Theta, \pi \mid \text{polymorphism}).
\end{aligned}
$$

We note that for all the simulations and inference results reported here and in the main text, we used only polymorphic data and the above conditional likelihood.

Inference of divergence times: a simulation study

Given a topology, we estimated the scaled branch lengths (under pure drift) and the DAF at the root by numerically maximizing the likelihood using the L-BFGS-B algorithm (Byrd

Table S1: Summary of normalized differences.

		min	5th per	median	mean	95th per	max
Scenario I	Beta	0.0002	0.0063	0.0537	0.0623	0.1453	0.1702
	Beta with spikes	0.0005	0.0027	0.0198	0.0257	0.0629	0.1096
	Kim Tree	0.0010	0.0037	0.0761	0.0910	0.2006	0.2254
Scenario II	Beta	0.0026	0.0241	0.1508	0.2947	0.8582	0.8753
	Beta with spikes	0.0015	0.0250	0.0922	0.1056	0.2536	0.4073
	Kim Tree	0.0015	0.0063	0.1184	0.2134	0.5735	0.6544

The table shows the summary of the distribution of the absolute normalized difference ($|1-\tau_{est}/\tau|$) between the inferred (τ_{est}) and true (τ) scaled branch lengths, for the two simulation scenarios and beta, beta with spikes and Kim Tree. For beta and beta with spikes, we used $T = 30$ and $K = 25$ and $K = 20$ for scenarios I and II, respectively.

et al. 1995) implemented in `SciPy` (Jones *et al.* 2001). For this, we treat the DAF at the root as a nuisance parameter assumed to be a beta distribution and estimated the two shape parameters.

To estimate the scaled branch lengths, we fixed the number of generations on each branch, estimated the population size and then reported the resulting scaled time. As presented in the parameter scaling section, if the population size is large enough, this approach should provide similar estimates independent of the chosen number of generations per branch.

For the tree depicted in Figure 2 in the main text, we set the total height (number of generations from the root to the present) to a given T and the generations per branches to be $t_{5\to4} = t_{4\to1} = t_{4\to2} = T/2$ and $t_{5\to3} = T$. We simulated data using two different scenarios (Table 1 in the main text).

A comparison between beta, beta with spikes and Kim Tree (Gautier and Vitalis 2013) is reported in the main text (Figure 3). Table S1 contains the summary of the quality of the estimates for all three methods for both simulation scenarios. Here, we discuss in more details the effect of the chosen height T and number of bins K.

Figure S2 (A and B) illustrates the quality of the estimates from beta and beta with spikes, for different tree heights T and number of bins K. One of the trends that is clear in the figure is that beta with spikes has a lower variance than beta in the estimated branches lengths. This is probably a result of the variability between the different simulated data sets of the number of sites that are close to being fixed or lost. This should have a stronger effect on the beta than the beta with spikes, as these sites require accurate probabilities close to the boundaries.

For scenario I, Figure S2 A indicates that using $K = 25$ bins is enough to obtain a good approximation for the likelihood, as, for a fixed tree height T, the beta with spikes has similar performance for $K = 25$ and larger values of K. The lower $K = 10$ decreases the quality of the estimates just slightly for the beta with spikes, but the effect is more noticeable in the quality of the beta approximation. The different behavior of the beta and beta with spikes when comparing $K = 10$ and $K = 25$ might indicate that the likelihood approximation is more robust to the number of bins, provided that the boundary probabilities are treated separately (as in the case of the beta with spikes). For values of K larger than 10, the beta distribution provides worse and worse estimates with an increased number of bins. This is most likely due to the more fine grained bins increasing the importance of accurately modeling the boundary probabilities, rendering worse results from the beta approximation.

We generally observe the same trends for both simulation scenarios (Figure S2 A and B), with the noticeable differences that: the average performance of beta and beta with spikes is lower for scenario II than scenario I; and the beta distribution has a surprisingly good performance for $K = 5$ bins under scenario II. However, the likelihood of the inferred branches under the beta distribution with $K = 5$ is approximately 30,000 units lower than the one under $K = 20$, indicating a much lower support for the branches inferred using $K = 5$.

The beta approximation provides just as good estimates regardless of the tree height T

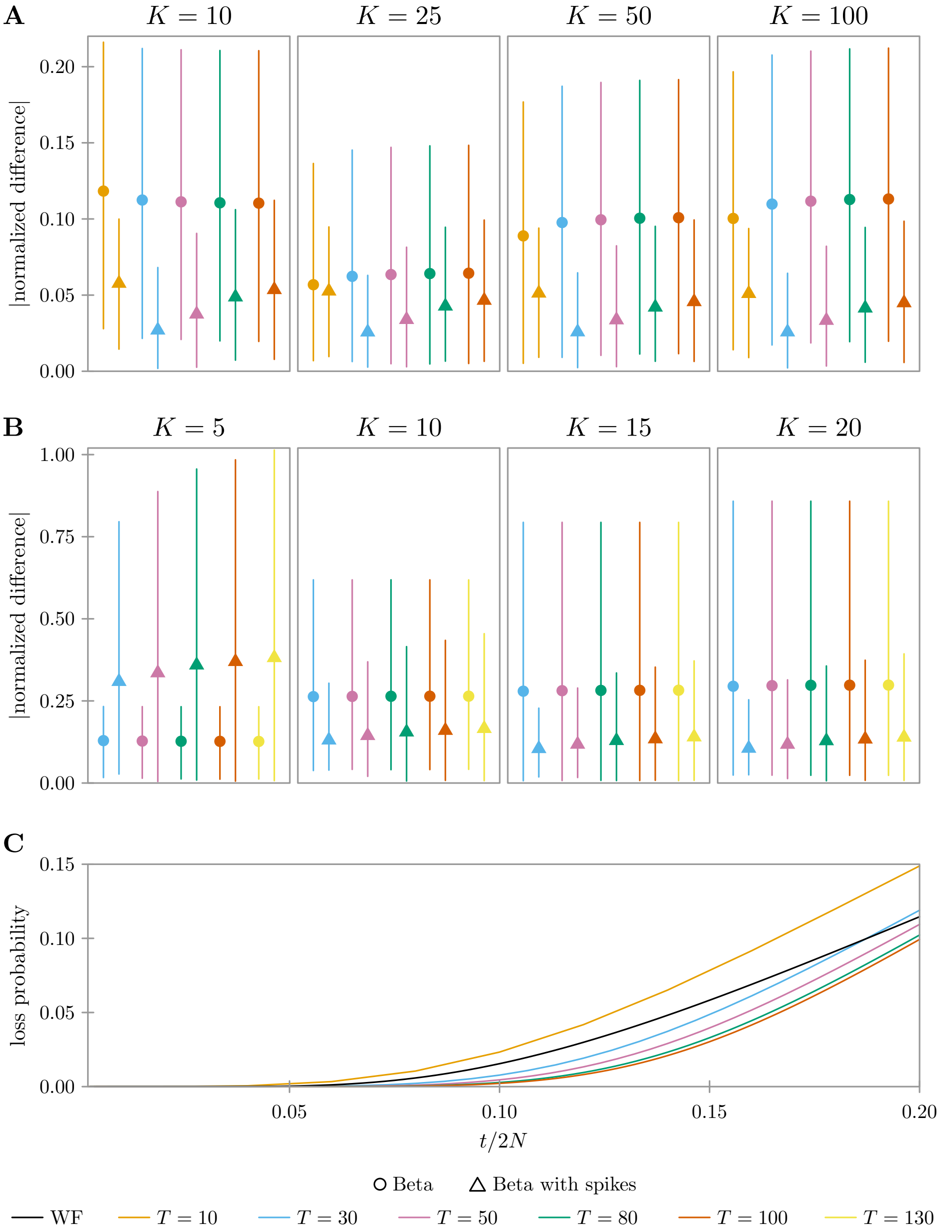

Figure S2: Effect of tree height T and number of bins K. Absolute value of the normalized difference between the estimated branch length τ_{est} and the true τ, given by $|1 - \tau_{est}/\tau|$, for beta (circle) and beta with spikes (triangle), for scenarios I (A) and II (B). The plot indicates the mean over all 50 replicates for all four branch lengths, together with the 5th and 95th percentiles as error bars. (C) Loss probability as calculated from the Wright-Fisher (black) using $2N = 200$, initial frequency $x_0 = 0.2$ and generation times t up to $t/2N = 0.2$, and beta with spikes using different maximum generation times T and corresponding population sizes $2N$ such that $T/2N = 0.2$.

used, while the beta with spikes is more sensitive to the tree height. In this case, different tree heights would correspond to different scaling and Δ, which is essentially a time step in a time discretization. The taller the tree, the more iterations are used in the recursion, allowing for more errors to accumulate from one iteration to the next. On the other hand, a tree that is too short leads to less accurate branch length inference. Here, a tree height of $T = 30$ provided the best inference for both simulation scenarios, which contained trees with different true heights (Table 1 in main text), indicating that this height might be a good general choice regardless of the true underlying tree. Figure S2 C shows the effect of T on the loss probability, illustrating that $T = 30$ leads to the most accurate approximation of the loss probability.

For the results reported in Figure 3 and Table S1, we used $T = 30$, $K = 25$ and $K = 20$ for scenarios I and II, respectively. As scenario II was built to generate chimpanzee-like data, we also used $T = 30$ and $K = 20$ for the results on the chimpanzee exome data reported in Figure 4 and Table 2.

We note here that the likelihoods reported in the main text in Table 2 were numerically maximized over the root DAF, while the branch lengths were kept constant.

LITERATURE CITED

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu, 1995 A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing *16*(5): 1190–1208.

Erdélyi, A., W. Magnus, F. Oberhettinger, and F. G. Tricomi, 1953 *Higher transcendental functions*, Volume 1. McGraw-Hill New York.

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

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

Jones, E., T. Oliphant, P. Peterson, et al., 2001 SciPy: Open source scientific tools for Python. [Online; accessed 2014-04-03].

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.