SUPPLEMENTARY INFORMATION

Scaling up Bayesian population phylogenomics through virtual dimension reduction

Flouri, T., Jiao, X., Huang, J., Rannala, B., and Yang Z.

Supplementary Note 1. Improved MCMC algorithms

We have introduced the following improvements to the MCMC proposals in BPP [1–4].

Step 1. Coalescent or migration time move on gene trees. The Gene-tree node-age move is used under the MSC, MSC-I, and MSC-M models while the migration-time move is used in MSC-M. Loop over loci and for each locus over the gene-tree node ages. Use a sliding window based on the Bactrian-Laplace kernel [5] to propose a new age t^*, and apply the Metropolis-Hastings (M-H) step of acceptance/rejection using the gene-tree density $p(G|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varpi})$ calculated at the current values of $\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varpi}$. As each gene-tree node age update is a small step, unlikely to greatly affect the conditionals of $\boldsymbol{\theta}$ or $\boldsymbol{\varpi}$, there does not seem to be a need to sample $\boldsymbol{\theta}$ or $\boldsymbol{\varpi}$ in each node-age move. After all node ages on all gene trees are updated, sample $\boldsymbol{\theta}, \boldsymbol{\varpi}$ from their (approximate) conditionals. In BPP, this step is parallelised by loci.

Step 2. Gene-tree SPR or simulation. Under MSC and MSC-I, loop over loci and for each locus over branches to make deterministic changes to the gene tree and apply the M-H step. After all gene trees are updated, sample $\boldsymbol{\theta}, \boldsymbol{\varpi}$ from their conditionals. As in the gene-tree node-age move, we consider changes to each gene tree as small moves.

Under the MSC-M model, we generate gene-tree topologies by simulating the backward-in-time process of coalescent and migration. Loop over loci and for each locus i, simulate the SPR subtrees (resulting in new gene tree G_i^*) using the current values of $\boldsymbol{\theta}$ and $\boldsymbol{\varpi}$, and applying the M-H step using the gene tree density $p(G|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varpi})$. After the loop over all loci, sample $\boldsymbol{\theta}$ and $\boldsymbol{\varpi}$ from their conditional distributions.

Step 4. Rubber-band move for updating τ_X for species-tree node X under MSC, MSC-I, and MSC-M. Determine the bounds (τ_l, τ_u) and propose τ_X by applying the Bactrian-Laplace move on $\log \tau_X$, with appropriate reflection. Apply changes to gene-tree node ages in affected populations according to eqs. A7 and A8 of ref. [1]. The deterministic modifications to node ages on gene trees favor small changes to the gene trees and the sequence likelihood. Sample $\boldsymbol{\theta}$ and $\boldsymbol{\varpi}$ from their (approximate) conditionals given the proposed gene trees. Apply the M-H step using the gene-tree density $p(G|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varpi})$. Note that under the inverse-gamma prior for θ and gamma prior for $\boldsymbol{\varpi}$, this move is equivalent to integrating out $\boldsymbol{\theta}$ and $\boldsymbol{\varpi}$. Under the gamma prior for θ, the conditional for θ is intractable and is approximated using a gamma or inverse-gamma distribution.

Step 5. Mixing or rescaling move under MSC, MSC-I, and MSC-M). We rescale all species split times on the species tree, and coalescent times and migration times on all gene trees by the same factor that is close to 1. This is a 1-D move in a high-dimensional space. The scale factor is generated from a sliding window on the log scale based on the Bactrian-Laplace kernel [6, p.225]. Sample $\boldsymbol{\theta}$ and $\boldsymbol{\varpi}$ from their conditionals given the gene trees with newly proposed coalescent times. Apply the M-H step of acceptance/rejection. We note that the conditional of the introgression probability φ depends on the numbers of lineages taking the hybridization route at each hybrid node [3], but not the coalescent times. As a result, there is no need to sample φs from their conditionals in this step.

Step 6. Updating $\boldsymbol{\theta}$ and $\boldsymbol{\varpi}$ under MSC-M or $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ under MSC-I. We sample those parameters from their conditionals or approximate conditionals given the gene trees. Our default option is to allow a small proportion (10%) for the sliding-window move, in addition to the (Metropolised) Gibbs move sampling from the conditionals.

Supplementary Note 2. Approximate gamma and inverse-gamma conditionals for the expected waiting time for a Poisson event

Suppose we use data of k events observed over time period T for a Poisson process to infer the expected waiting time $\mu = 1/\lambda$ where λ is the Poisson rate. While the inverse-gamma is a conjugate prior on μ (equivalent to a gamma prior on λ), a gamma prior with its lighter tail, $\mu \sim \mathrm{G}(a, b)$, may be more appropriate in certain applications. The gamma prior for μ is not conjugate, and neither the posterior nor the conditional is tractable. However, we may use the gamma or inverse-gamma distributions as an approximation.

For data of k events over time period T, the likelihood is

$$L(\mu) = p(k|\mu) = \mu^{-k}\, \mathrm{e}^{-T/\mu}\,. \tag{S1}$$

The maximum likelihood estimate (MLE) is given by the sample mean, T/k. Here we treat k as data although one could treat T as data as well.

Under the gamma prior $\mu \sim \mathrm{G}(a, b)$, the posterior or conditional is

$$p(\mu|k) = \frac{1}{z}\mu^{a-1-k}\, \mathrm{e}^{-b\mu - T/\mu}, \tag{S2}$$

where $z = \int_0^\infty \mu^{a-1-k}\, \mathrm{e}^{-b\mu - T/\mu}\, \mathrm{d}\mu$, the normalizing constant, is in general not needed in the algorithm.

To fit an approximate distribution such as the gamma $\mu \sim \mathrm{G}(a^*, b^*)$, one should ideally optimise a^* and b^*, with a, b, k, and T given, by minimizing the Kullback-Leibler (KL) divergence between the true conditional and the fitted gamma distribution

$$D = \int_0^\infty p(\mu|k) \log \frac{p(\mu|k)}{g(\mu; a^*, b^*)}\, \mathrm{d}\mu \tag{S3}$$

or, equivalently, by maximizing

$$\int_0^\infty p(\mu|k) \log g(\mu; a^*, b^*)\, \mathrm{d}\mu. \tag{S4}$$

This is a 2-D optimization but calculation of the objective function involves numerical calculation of a 1-D integral. Instead we use a method-of-moments–like approach, which appears to be adequate for our purpose.

Let $\ell(\mu) = \log p(\mu|k)$. Setting the first derivative to zero,

$$\ell' = \frac{a-1-k}{\mu} - b + \frac{T}{\mu^2} = 0$$

gives the conditional mode,

$$\bar{\mu} = \frac{(a-1-k) + \sqrt{(a-1-k)^2 + 4bT}}{2b}.$$

This always exists, and is between the prior mode $(a-1)/b$ and the sample mean T/k if $a > 1$. Also $\ell'' = -\frac{a-1-k}{\mu^2} - \frac{2T}{\mu^3} < 0$ at $\hat{\mu}$ so $\hat{\mu}$ is indeed a mode.

Next we match the mode $m = \hat{\mu}$ and the approximate variance $v = -1/\ell''$ with those of the fitted gamma distribution. From

$$m = (a^* - 1)/b^*, \quad v = a^*/b^{*2},$$

we get

$$\begin{aligned} a^* &= 1 + \frac{m^2}{2v} + \sqrt{\frac{m^2}{v} + \frac{m^4}{4v^2}}, \\ b^* &= \frac{a^* - 1}{m}. \end{aligned} \tag{S5}$$

Here a^*, b^* are functions of a, b in the prior and k, T in the data. It seems better to treat $\hat{\mu}$ as the mode rather than the mean of the gamma distribution. Another option is to use $b^* = m/v$, which achieves better fit than $b^* = (a^* - 1)/m$ in some tests (fig. S1**a**). All those forms are nearly equivalent in large datasets (with large k and T). In case of no data ($k = T = 0$), we set $a^* = a, b^* = b$.

If we fit the inverse-gamma distribution $\mathrm{IG}(\alpha^*, \beta^*)$ to the mode $m = \hat{\mu}$ and the variance v, we have

$$m = \frac{\beta^*}{\alpha^* + 1}, \quad v = \frac{\beta^{*2}}{(\alpha^* - 1)^2(\alpha^* - 2)},$$

from which we get α^* to be the root of the following cubic equation

$$z^3 - (4+d)z^2 + (5-2d)z - (d+2) = 0, \tag{S6}$$

while

$$\beta^* = m(\alpha^* + 1),$$

where $d = m^2/v$. In case of no data ($k = T = 0$), we match the first two moments between the gamma prior and the inverse-gamma conditional to get

$$\begin{aligned} \alpha^* &= a + 2, \\ \beta^* &= a(a+1)/b. \end{aligned} \tag{S7}$$

For example $\mathrm{G}(2, 200)$ translates into $\mathrm{IG}(4, 0.03)$.

We can use the KL divergence D, calculated at the estimates $(a^*, b^*$ for the gamma or α^*, β^* for the inverse-gamma), to assess the quality of approximation (fig. S1**a**). Some example calculations are shown in figure S1**b**, with the gamma prior $\mu \sim \mathrm{G}(a, b)$ with $a = 2, b = 200$, and data of $k = 2$ or 10 events observed over time period T, where T is fixed at $c \times a/b \cdot k$ so that the MLE T/k is $c = 0.1, 1$, or 10 times the prior mean.

Note that when the amount of data increases (i.e., when $k \to \infty$, $T \to \infty$, with T/k held constant), the true conditional will be dominated by the likelihood and approaches a Gaussian density, and both the gamma and inverse-gamma approximations will also approach the same Gaussian density. Thus both approximations are expected to be good in large datasets. This is confirmed by the observation that the acceptance rate $P_{\text{jump}} \approx 100\%$ for the θ updates using approximate conditionals.

We note that the gamma approximate conditional, which has a light tail, may lead to rejection of the Gibbs proposal, causing convergence issues if the chain starts from a poor starting point in the tail (fig. S2). As an example, suppose the posterior mode is around 0.01, but the chain has started using a value sampled from the prior $\mathrm{G}(2, 10)$, with mean 0.1. Because of its light tail, the gamma proposal may lead to rejection of the approximate Gibbs move. For fast convergence with poor starting values, the approximate inverse-gamma conditional is thus better than the approximate gamma conditional. When the chain has reached stationarity and is visiting states around the posterior mode, both approximate conditionals are effective.

When applied to algorithms under the MSC models in BPP, $\mu = \theta_j/2$ is the expected waiting time for coalescence between two sequences in population j and we assign the prior $\theta_j \sim \mathrm{G}(a, b)$. The total number of coalescent events (k) and the total waiting time (T) on the gene trees constitute the 'data'. We then use the gamma and inverse-gamma to approximate the conditional distribution of θ given the gene trees, leading to improved mixing of the MCMC algorithm. The approximate gamma conditional for θ_j is then $\mathrm{G}(a^*, b^*)$, with a^*, b^* given by eq. S5 and with $k = \sum_i c_{ji}$ coalescent events over the total time period for coalescence, $T = \sum_i \frac{2}{h_i} C_{ji}$, where c_{ji} and C_{ji} are defined in eq. 4. Similarly the approximate inverse-gamma conditional is $\mathrm{IG}(\alpha^*, \beta^*)$, given by eq. S6.

Supplementary Note 3. Calculation of posterior distributions and posterior summaries for parameters $\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\varpi}$ using their conditionals

The statistical problem considered here is as follows. We have an MCMC sample, $\{x_1, x_2, \cdots, x_N\}$, from an algorithm operating on X with the variable Y integrated out. Given each $x_i, i = 1, \cdots, N$, the conditional distribution of Y is analytically tractable. In our context, X represents species split times ($\boldsymbol{\tau}$) and gene trees (G) at all loci, while Y represents θs, φs, or ϖs, which are integrated out in the MCMC (or sampled from their conditionals), with the conditional distribution of $Y|X$ known (gamma, inverse-gamma, or beta). Here we consider the case of $Y|x_i \sim \mathrm{G}(\alpha_i, \beta_i)$, with $\alpha_i > 1, \beta_i > 0$. Let $u_i = \mathbb{E}(Y|x_i) = \alpha_i/\beta_i$ be the conditional mean and $v_i = \mathbb{V}(Y|x_i) = \alpha_i/\beta_i^2$ the conditional variance. Note that $\alpha_i, \beta_i, u_i, v_i$ are all functions of x_i. For other conditional distributions, these are defined accordingly.

The marginal mean and variance of Y are defined as

$$\mu_y = \iint y p(x) p(y|x)\, \mathrm{d}y\, \mathrm{d}x,$$
$$\sigma_y^2 = \iint (y-\mu_y)^2 p(x) p(y|x)\, \mathrm{d}y\, \mathrm{d}x. \tag{S8}$$

We use the mean of the conditional distribution (u_i) to estimate μ_y,

$$\tilde{u} = \frac{1}{N}\sum_{i=1}^{N} u_i = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(Y|x_i), \tag{S9}$$

where $u_i = \mathbb{E}(Y|x_i)$. Here we are using an MCMC sample of X to estimate μ_y, treating μ_y as the expectation of a deterministic function of X: $\mu_y = \mathbb{E}_X(u)$ where $u = u(X) = \mathbb{E}(Y|X)$. The function u has the variance

$$\sigma_u^2 = \int (u(X)-\mu_y)^2 p(x)\, \mathrm{d}x, \tag{S10}$$

which can be estimated by

$$\hat{\sigma}_u^2 = \frac{1}{N}\sum_i (u_i - \tilde{u})^2. \tag{S11}$$

The estimator of eq. S9 has the variance $\mathbb{V}(\tilde{u})$, with

$$\lim_{N\to\infty} N\,\mathbb{V}(\tilde{u}) = \sigma_u^2\big[1 + 2(\rho_1^u + \rho_2^u + \cdots)\big], \tag{S12}$$

where $\rho_k^u = \mathrm{corr}(u_i, u_{i+k})$ is the lag-k auto-correlation. As the x_i are a dependent sample, the u_i are a dependent sample also. $\mathbb{V}(\tilde{u})$ can be calculated from the u_i using the initial positive sequence method [7].

Next we consider estimating μ_y using an MCMC sample of Y: $\{y_1, y_2, \cdots, y_N\}$, with y_i to be a random sample from the conditional $Y|x_i \sim \mathrm{G}(\alpha_i, \beta_i)$. The estimator

$$\tilde{y} = \frac{1}{N}\sum_i y_i \tag{S13}$$

has the variance $\mathbb{V}(\tilde{y})$, with

$$\lim_{N\to\infty} N\,\mathbb{V}(\tilde{y}) = \sigma_y^2\big[1 + 2(\rho_1^y + \rho_2^y + \cdots)\big], \tag{S14}$$

where $\rho_k^y = \mathrm{corr}(y_i, y_{i+k})$. Since $y_i \sim p(y|x_i)$ and the x_i are correlated, the y_i are correlated as well. Note that the sample mean ($\bar{y}$) from an independent sample of the same size has the variance σ_y^2/N, so that the efficiency of $\tilde{y}$ is defined as $E_{\tilde{y}} = 1/[1+2(\rho_1^y+\rho_2^y+\cdots)]$ (eq. 15).

Note that

$$\sigma_y^2 = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)) = \mathbb{E}(v) + \sigma_u^2, \tag{S15}$$

where $\mathbb{E}(v)$ can be estimated by

$$\tilde{v} = \frac{1}{N}\sum_i v_i. \tag{S16}$$

We have the relative efficiency of $\tilde{u}$ relative to $\tilde{y}$, both considered estimators of μ_y, to be

$$E_{\tilde{u},\tilde{y}} = \lim_{N\to\infty}\frac{\mathbb{V}(\tilde{y})}{\mathbb{V}(\tilde{u})} = \frac{\sigma_y^2}{\sigma_u^2}\times\frac{1+2\sum_{k\ge1}\rho_k^y}{1+2\sum_{k\ge1}\rho_k^u}. \tag{S17}$$

Note that

$$\begin{aligned}
\mathbb{E}(y_i y_{i+k}) &= \iint y_i y_{i+k} p(y_i, y_{i+k})\, \mathrm{d}y_i\, \mathrm{d}y_{i+k} \\
&= \iiiint y_i y_{i+k} p(y_i|x_i) p(y_{i+k}|x_{i+k}) p(x_i, x_{i+k})\, \mathrm{d}y_i\, \mathrm{d}y_{i+k}\, \mathrm{d}x_i\, \mathrm{d}x_{i+k} \\
&= \iint \Big[\int y_i p(y_i|x_i)\, \mathrm{d}y_i\Big]\Big[\int y_{i+k} p(y_{i+k}|x_{i+k})\, \mathrm{d}y_{i+k}\Big] \\
&\quad \times p(x_i, x_{i+k})\, \mathrm{d}x_i\, \mathrm{d}x_{i+k} \\
&= \iint u_i u_{i+k} p(x_i, x_{i+k})\, \mathrm{d}x_i\, \mathrm{d}x_{i+k} \\
&= \mathbb{E}(u_i u_{i+k}).
\end{aligned} \tag{S18}$$

Note that $u_i = \mathbb{E}(Y|x_i) = \int y_i p(y_i|x_i)\, \mathrm{d}y_i$.

Thus from $\rho_k^y = (\mathbb{E}(y_i y_{i+k}) - \mu_y^2)/\sigma_y^2$, we have

$$\rho_k^u = \frac{\mathbb{E}(u_i u_{i+k}) - \mu_y^2}{\sigma_u^2} = \frac{\sigma_y^2}{\sigma_u^2}\rho_k^y. \tag{S19}$$

Let $c = \sigma_y^2/\sigma_u^2 > 1$. Then $\rho_k^u = c\rho_k^y$ for all $k \ge 1$, and

$$E_{\tilde{u},\tilde{y}} = \frac{c + 2c\sum_{k\ge1}\rho_k^y}{1+2c\sum_{k\ge1}\rho_k^y} > 1. \tag{S20}$$

Then we have

$$E_{\tilde{u},\tilde{y}} \equiv \lim_{N\to\infty}\frac{\mathbb{V}(\tilde{y})}{\mathbb{V}(\tilde{u})} = \frac{1}{1-\big(1-\frac{1}{c}\big)E_{\tilde{y}}}, \tag{S21}$$

while the efficiency of $\tilde{u}$ (relative to the sample mean $\bar{y}$ for an independent sample of Y) is

$$E_{\tilde{u}} = E_{\tilde{u},\tilde{y}} \cdot E_{\tilde{y}}. \tag{S22}$$

Note that $E_{\tilde{u},\tilde{y}}$ is a monotonically increasing function of c and $E_{\tilde{y}}$. It may be noted that c is determined by the inference problem or the posterior/target distribution, while $E_{\tilde{y}}$ is affected by the MCMC proposal algorithm such as $q(x'|x)$. First we consider $E_{\tilde{y}}$. *The more efficient $\tilde{y}$ already is, the greater will the improvement be by switching to $\tilde{u}$.*

We note a few special cases. When $E_{\tilde{y}} \approx 0$ (i.e., when the x_i are strongly positively correlated with $\rho_k^y \approx 1$), $E_{\tilde{u},\tilde{y}} \approx 1$ (which is the minimum). When $E_{\tilde{y}} = 1$ (i.e., when the x_i are an independent sample, so that the y_i are independent as well, with $\rho_k^y = 0$), $E_{\tilde{u},\tilde{y}} = c$. When $E_{\tilde{y}} > 1$ or in other words when $2\sum_{k\ge1}\rho_k^y < 0$, $\tilde{y}$ is already 'super-efficient' and we have $E_{\tilde{u},\tilde{y}} > c$. Note that $E_{\tilde{y},\tilde{u}} = 1/E_{\tilde{u},\tilde{y}} = 1 - (1-\frac{1}{c})E_{\tilde{y}}$ is a linear decreasing function of $E_{\tilde{y}}$, taking values 1, $\frac{1}{c}$, and 0, at $E_{\tilde{y}} = 0, 1$, and $\frac{c}{c-1}$, respectively. $E_{\tilde{y}}$ has a maximum, $\frac{c}{c-1}$, achieved when $E_{\tilde{u}} \to \infty$ in which case $\rho_1^u \to -1$ and $\rho_1^y \to -1/c$.

Consider for instance two algorithms for sampling from $X \sim \mathrm{G}(10, 100)$ and $Y|X \sim \mathrm{IG}(2+10, 0.02+x)$. We have $\mathbb{E}(x) = 0.1$, $\mathbb{V}(x) = 0.001$, $u = \frac{0.02+x}{11}$, $v = \mathbb{V}(Y|X) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = \frac{(0.02+x)^2}{1210}$, $\sigma_u^2 = \mathbb{V}(u) = \mathbb{V}(x)/121 = 0.001/121$. Also $\mathbb{E}(v) = \frac{1}{1210}\mathbb{E}(0.02+$

$x)^2 = \frac{1}{1210}[\mathbb{V}(x) + \mathbb{E}(x)^2 + 0.04\,\mathbb{E}(x) + 0.0004] = \frac{1}{1210}[0.001 + 0.01 + 0.004 + 0.0004] = \frac{0.00154}{121}$. From $\sigma_y^2 = \sigma_u^2 + \mathbb{E}(v)$, we have $c = \sigma_y^2/\sigma_u^2 = \frac{0.00254}{121} \Big/ \frac{0.001}{121} = 2.54$.

In the first algorithm, we use a sliding-window on $\log(x)$ to update x, and sample y from $p(y|x)$. The estimated efficiency for $\tilde{y}$ is $E_{\tilde{y}} = 0.43$, while $\tilde{u}$ based on the sample of $u = \frac{0.02+x}{11}$ gives $E_{\tilde{u}} = 0.58$, with an improvement of $E_{\tilde{u},\tilde{y}} = 1.35$, in agreement with Eq. S21.

In the second algorithm we use the mirror move to update $\log(x)$, with y sampled from $p(y|x)$ [8]. Then $E_{\tilde{y}} = 1.13$, while the efficiency for the u sample is $E_{\tilde{u}} = 3.58$, with an improvement of $E_{\tilde{u},\tilde{y}} = 3.17$, again in agreement with Eq. S21.

The efficiency measures $E_{\tilde{y}}$ and $E_{\tilde{u}}$ are included in the BPP output. The approximation using the approximate conditional the gamma prior on θ is more accurate in larger datasets.

Algorithm for generating the marginal summary of Y. Note that the conditional distribution sampled during the MCMC, that is α_i, β_i or u_i, v_i, allow us to estimate all of the following quantities:

- $\mu_X = \mathbb{E}(Y|X)$ (the mean of Y in the target distribution),
- σ_y^2 (the variance of Y in the target distribution),
- σ_u^2 (the variance of the conditional mean u),
- $\mathbb{E}(v)$ (the mean of the conditional variance),
- $E_{\tilde{y}}$ (the efficiency of the sample mean $\tilde{y}$ from the MCMC sample),
- $E_{\tilde{u}}$ (the efficiency of the sample mean $\tilde{u}$ from the MCMC sample),
- $E_{\tilde{u},\tilde{y}}$ (the relative efficiency of eq. S21),
- $E_{\bar{u}}$ (the efficiency of the sample mean $\bar{u}$ from an i.i.d. sample; BPP does not generate such a sample).

Our algorithm works as follows. Loop through the MCMC sample (α_i, β_i), which can be used to calculate (u_i, v_i). The average of v_i is an estimate of $\mathbb{E}(v)$. Collect the u_i into a vector and apply the initial positive sequence algorithm to calculate $\mathbb{E}(u) \equiv u_y$, $\mathbb{V}(u) \equiv \sigma_u^2$, and apply the initial positive sequence algorithm to calculate

$$E_{\bar{u}}/E_{\tilde{u}} = 1 + 2(\rho_1^u + \rho_2^u + \cdots) \equiv \mathcal{T}_u,$$

where $\mathcal{T}_u$ is the *integrated autocorrelation time*. Then we have $\sigma_y^2 = \mathbb{E}(v) + \sigma_u^2$ from eq. S15, and $c = \sigma_y^2/\sigma_u^2$. Furthermore

$$\frac{E_{\bar{y}}}{E_{\tilde{y}}} = \frac{1}{E_{\tilde{y}}} \equiv \mathcal{T}_y = 1 + 2(\rho_1^y + \rho_2^y + \cdots) = 1 + (\mathcal{T}_u - 1)/c,$$

since $\rho_k^u = c\rho_k^y$ for all $k \ge 1$. In other words,

$$E_{\tilde{y}} = \frac{1}{1 + (\mathcal{T}_u - 1)/c}. \tag{S23}$$

Then we use eq. S21 to calculate the relative efficiency $E_{\tilde{u},\tilde{y}}$, with

$$E_{\tilde{u}} = E_{\tilde{u},\tilde{y}} E_{\tilde{y}} = c/\mathcal{T}_u. \tag{S24}$$

Finally if we had an i.i.d. sample of u, the efficiency of the sample mean $\bar{u}$ for estimating $\mu_y \equiv \mathbb{E}(Y)$, would be

$$E_{\bar{u}} = E_{\tilde{u}} \cdot \mathcal{T}_u = c. \tag{S25}$$

Algorithm for generating the posterior mean of $Y = Y_1Y_2$. Under the MSC-M model, the population migration rate, $M_{AB} = \varpi\theta_B/4$, is a function of ϖ and θ_B, the (approximate) conditionals of which are available (gamma for ϖ and inverse-gamma or gamma for θ). We would like to estimate the posterior mean of M_{AB}.

Let $Y = Y_1Y_2$. Note that Y_1 and Y_2 are independent given X. Let $\mu = \mathbb{E}(Y|X) = \mathbb{E}(Y_1|X)\,\mathbb{E}(Y_2|X)$ and $v = \mathbb{V}(Y|X) = \mathbb{E}(Y_1^2|X)\,\mathbb{E}(Y_2^2|X) - [\mathbb{E}(Y_1|X)\,\mathbb{E}(Y_2|X)]^2$. Let $u_{i1} = \mathbb{E}(Y_1|x_i)$, $v_{i1} = \mathbb{V}(Y_1|x_i)$ be the conditional mean and variance for Y_1 and define u_{i2}, v_{i2} similarly for Y_2. Loop through the MCMC sample to calculate

$$\begin{aligned} u_i &= u_{i1}u_{i2}, \\ v_i &= (v_{i1} + u_{i1}^2)(v_{i2} + u_{i2}^2) - (u_{i1}u_{i2})^2. \end{aligned} \tag{S26}$$

Like the algorithm above, the average of v_i is used to estimate $\mathbb{E}(v)$. Collect the u_i into a vector and use the procedure above to calculate the estimate and the efficiency based on Y and u.

The same principles can be used to calculate other functions of parameters that have been integrated out, such as XY, Y_1Y_2, XY_1Y_2, etc. This is not pursued here, since these are not needed in the current models implemented in BPP.

Next we estimate the PDF for Y and calculate the 95% HPD CI for the marginal mean $\mathbb{E}(Y)$, by using the fact that the HPD interval is the shortest. We assume that the density is single-moded. When the posterior is multimodal, the HPD CI may consist of multiple disconnected intervals, and modifications to the algorithm may be necessary [9]. Our algorithm works as follows.

Step 1. Determine the range for the variable, $(y_{\min}, y_{\max})$, and define K bins of equal width $\Delta = (y_{\max} - y_{\min})/K$, with the midpoint of the kth bin to be $y_k = y_{\min} + (k - \frac{1}{2})\Delta$. We used $y_{\min} = \max\{0, \hat{\mu}_y - 4\hat{\sigma}_y\}$ and $y_{\max} = \hat{\mu}_y + 4\hat{\sigma}_y$, where $\hat{\mu}_y$ and $\hat{\sigma}_y$ are the estimated mean and SD of Y.

Step 2. Calculate the empirical PDF for y by looping through the N samples $\{x_1, \cdots, x_N\}$, and for each sample adding the probability to each bin. For each x_i, calculate α_i, β_i for the conditional. Then the increment for the sample is

$$p_k = \frac{1}{N}\sum_{i=1}^{N} g(y_k; \alpha_i, \beta_i)\Delta, \quad k = 1, \cdots, K, \tag{S27}$$

where $g(y_k; \alpha_i, \beta_i)$ is the gamma or inverse-gamma density with parameters α_i, β_i at the midpoint y_k for the k-th bin. A more accurate approach may be to use the gamma or inverse-gamma CDF or incomplete gamma function. This is not pursued here.

Step 3. Scan all intervals with coverage $1 - \alpha$ and the shortest is the HPD interval. An alternative approach may be to slide downwards a horizontal line from the

mode of the distribution (the maximum of p_k) until the interval has the correct coverage $(1-\alpha)$.

Processing $N = 10^6$ samples using $K = 1000$ bins took 3-4 seconds on a laptop, mostly spent on step 2.

This theory for calculating the PDF and the HPD CI for Y should apply to functions of parameters that have been integrated out. Again we consider $Y = Y_1Y_2$ (which is useful to generating posterior summaries for the population migration rate $M_{AB} = \varpi\theta_B/4$ under the MSC-M model).

Let $Y = Y_1Y_2$ and $\mu_i = \mathbb{E}(Y_1|x_i)\,\mathbb{E}(Y_2|x_i)$, since Y_1 and Y_2 are independent given X. We have $\tilde{\mu} = \frac{1}{N}\sum \mu_i$. We then follow the same procedure above to calculate the efficiency $(E_{\tilde{\mu}})$ of $\tilde{\mu}$ for estimating $\mathbb{E}(Y)$, relative to an iid sample of Y. In calculating the PDF (and HPD CI) for Y, we partition Y_1 and Y_2 each into K bins, with the probability in each bin given by p_{1k}, p_{2k}, for $k = 1, \cdots, K$. The k_1k_2-th grid on the Y_1-Y_2 plane makes a contribution of $p_{1k_1}p_{2k_2}$ to the bin including $Y = y_{1k_1}y_{2k_2}$. One should calculate the index for the bin without the need for a search.

The four vectors of parameters (a_i, b_i for Y_1 and α_i, β_i for Y_2, say) are sufficient, and there is no need for the Y sample, to calculate $\tilde{\mu}, E_{\tilde{\mu}}, E_{\tilde{y}}$, and the HDP CI for $\mathbb{E}(Y)$.

A numerical example. Suppose $Y|X \sim \text{Exp}(X)$ and $X \sim \text{G}(3, 1)$. Then the parameters for the conditionals $a_i = 1$ (a constant) and $b_i \sim \text{G}(3, 1)$. The marginal PDF of Y is $p(y) = \frac{3}{(y+1)^4}$, which is a monotonically decreasing function for $y > 0$. The CDF is $F(y) = 1 - \frac{1}{(y+1)^3}$. Thus $\mathbb{E}(Y) = \frac{1}{2}$, $\mathbb{V}(Y) = \frac{3}{4}$ (SD = 0.8660), and the $100p\%$ quantile is $(1-p)^{-1/3}-1$. The 95% HPD interval of Y is $(0, (1-0.95)^{-1/3} - 1) = (0, 1.7144)$ while the 95% equal-tail interval is $((1-0.025)^{-1/3} - 1, (1-0.975)^{-1/3} - 1) = (0.008475, 2.4200)$.

We applied the algorithm to 10^6 samples of a_i and b_i (table S2). The accuracy depends on the range and the number of bins used, but is overall acceptable, so that more sophisticated procedures such as kernel-density smoothing is not needed.

References

[1] Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656 (2003).

[2] Rannala, B. & Yang, Z. Efficient bayesian species tree inference under the multispecies coalescent. *Syst. Biol.* **66**, 823–842 (2017).

[3] Flouri, T., Jiao, X., Rannala, B. & Yang, Z. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* **37**, 1211–1223 (2020).

[4] Flouri, T., Jiao, X., Huang, J., Rannala, B. & Yang, Z. Efficient Bayesian inference under the multispecies coalescent with migration. *Proc. Nat. Acad. Sci. U.S.A.* **120**, e2310708120 (2023).

[5] Yang, Z. & Rodríguez, C. E. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19307–19312 (2013).

[6] Yang, Z. *Molecular Evolution: A Statistical Approach* (Oxford University Press, Oxford, England, 2014).

[7] Geyer, C. Practical Markov chain Monte Carlo. *Stat. Sci.* **7**, 473–511 (1992).

[8] Thawornwattana, Y., Dalquen, D. & Yang, Z. Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Analysis* **13**, 1033–1059 (2018).

[9] Chen, M.-H. & Shao, Q.-M. Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Computat. Graph. Stat.* **8**, 69–92 (1999).

[10] Fontaine, M. C. *et al.* Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524 (2015).

[11] Neafsey, D. E. *et al.* Mosquito genomics. highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347**, 1258522 (2015).

Table S1: BPP settings for MCMC algorithms tested using empirical datasets

Algorithm	Priors on θ and ϖ (or ϖ)	Moves for updating θ and φ (or ϖ)	Rubber-band & scaler	Command-line options
The MSC-I model				

A1 (I-IG-int)	IG prior on θ, integrating out θ
	`thetaprior = 3 0.04 int`
	`phiprior = 1 1`
		None		

A2 (I-IG-gIG)	IG prior on θ, estimating θ and φ
	`thetaprior = 3 0.04`
	`phiprior = 1 1`
		Gibbs (IG for θ and beta for φ)	Sample θ from IG, φ from beta	`--theta-slide-prob=0`
				`--phi-slide-prob=0`

A3 (I-IG-slide)	ibid	Sliding window	Don't sample θ or φ	`--theta-slide-prob=1`
				`--no-rb-theta-update`
				`--no-mix-theta-update`
				`--phi-slide-prob=1`

A4 (I-G-gG)	G prior on θ, estimating θ and φ
	`thetaprior = gamma 2 100`
	`phiprior = 1 1`
		Metropolised Gibbs on θ (G) and Gibbs on φ (beta)	Sample θ from G, φ from beta	`--theta-slide-prob=0`
				`--theta-prop=mg_gamma`
				`--phi-slide-prob=0`

A5 (I-G-gIG)	ibid	Metropolised Gibbs on θ (IG) and Gibbs on φ (beta)	Sample θ from IG, φ from beta	`--theta-slide-prob=0`
				`--theta-prop=mg_invg`
				`--phi-slide-prob=0`

A6 (I-G-slide)	ibid	Sliding window	Don't sample θ or φ	same as for A3
The MSC-M model				

B4 (M-G-gG)	G priors on θ and ϖ, estimating θ and ϖ
	`thetaprior = gamma 2 100`
	`wprior = 2 1`
		Metropolised Gibbs on θ (G) and Gibbs on ϖ (G)	Sample θs from G and ϖ from G	`--theta-slide-prob=0`
				`--theta-prop=mg_gamma`

B5 (M-G-gIG)	ibid	Metropolised Gibbs on θ (IG) and Gibbs on ϖ (G)	Sample θs from IG and ϖ from G	`--theta-slide-prob=0` `--theta-prop=mg_invg`

B6 (M-G-slide)	ibid	Sliding window	Don't sample θ or ϖ	`--theta-slide-prob=1`
				`--no-rb-theta-update`
				`--no-rb-w-update`
				`--no-mix-theta-update`
				`--no-mix-w-update`
				`--wrate-slide-prob=0`

Note.— The key to algorithm is in the format 'model-prior-algorithm', with model to be 'I' for MSC-I and 'M' for MSC-M; with prior for θs to be 'IG' for inverse-gamma and 'G' for gamma (beta for φ in MSC-I and gamma for ϖ in MSC-M are always used); and algorithm to be 'slide' for sliding-windows, 'it' for integrating out θs, 'gIG' for (Metropolised) Gibbs based on the inverse-gamma conditional for θs, 'gG' for Metropolised Gibbs based on the gamma conditional for θs. Under the MSC-M model, inverse-gamma priors on θs are expected to produce similar results to gamma priors and are not included in our test. The default option for the rubber-band and mixing (scaler) moves in BPP is to sample θ and ϖ (or φ), and command-line switches are used to turn off the sampling. Also by default, θ, φ and ϖ are updated using a mixture of sliding windows and (Metropolised) Gibbs sampling.

Table S2: Estimation of the marginal distribution of Y using the conditionals under different settings.

Range	nbins	mean	SD	95% HPD CI	95% equal-tail CI
Truth	–	0.5000	0.8660	(0.00000, 1.7144)	(0.00848, 2.4200)
$(0, \hat{\mu} + 3\hat{\sigma})$	10^3	0.5001	0.8654	(0.00155, 1.9027)	(0.00774, 2.4198)
$(0, \hat{\mu} + 4\hat{\sigma})$	10^3	0.5001	0.8654	(0.00198, 1.9670)	(0.00990, 2.4186)
$(0, \hat{\mu} + 5\hat{\sigma})$	10^3	0.5001	0.8654	(0.00241, 2.0346)	(0.00724, 2.4160)
$(0, \hat{\mu} + 3\hat{\sigma})$	10^4	0.5001	0.8654	(0.00015, 1.7325)	(0.00852, 2.4221)
$(0, \hat{\mu} + 4\hat{\sigma})$	10^4	0.5001	0.8654	(0.00020, 1.7370)	(0.00852, 2.4220)
$(0, \hat{\mu} + 5\hat{\sigma})$	10^4	0.5001	0.8654	(0.00024, 1.7419)	(0.00845, 2.4220)

Note.– The MCMC algorithm samples from $X \sim \mathrm{G}(3, 1)$, while the conditional is $Y|X \sim \mathrm{Exp}(X)$.

Table S3: Posterior means (and 95% HPD CIs) and relative mixing efficiency of MCMC algorithms applied to the *Baobab* dataset

	MSC-I			MSC-M	
	Mean (CI)	$E_{\mathrm{A2}}/E_{\mathrm{A3}}$	$E_{\mathrm{A5}}/E_{\mathrm{A6}}$	Mean (CI)	$E_{\mathrm{B4}}/E_{\mathrm{B6}}$
τ_r	0.0145 (0.0140, 0.0150)	1.23	0.92	0.0146 (0.0140, 0.0151)	1.98
τ_a	0.0103 (0.0093, 0.0113)	0.80	0.58	0.0063 (0.0058, 0.0068)	1.29
τ_b	0.0052 (0.0050, 0.0054)	0.76	0.69	0.0052 (0.0050, 0.0055)	2.39
τ_c	0.0043 (0.0041, 0.0045)	1.41	1.62	0.0041 (0.0039, 0.0043)	1.70
τ_d	0.0013 (0.0011, 0.0014)	1.24	1.43	0.0013 (0.0011, 0.0014)	1.35
τ_e	0.0032 (0.0030, 0.0034)	0.92	1.05	0.0033 (0.0031, 0.0034)	1.71
τ_f	0.0009 (0.0008, 0.0010)	1.12	0.75	0.0009 (0.0008, 0.0010)	1.12
τ_g	0.0006 (0.0005, 0.0007)	2.79	2.48	0.0006 (0.0005, 0.0007)	2.33
τ_y	0.0049 (0.0047, 0.0051)	0.79	0.50	–	–
θ_{Age}	0.0033 (0.0029, 0.0038)	0.55 (0.57)	0.30 (0.31)	0.0033 (0.0029, 0.0037)	2.32 (2.37)
θ_{Adi}	0.0136 (0.0130, 0.0143)	2.18 (2.26)	0.73 (0.75)	0.0140 (0.0133, 0.0146)	2.76 (2.86)
θ_{Aru}	0.0059 (0.0055, 0.0064)	1.44 (1.58)	1.53 (1.72)	0.0060 (0.0055, 0.0064)	1.89 (2.03)
θ_{Asu}	0.0017 (0.0014, 0.0020)	2.08 (2.14)	2.09 (2.14)	0.0017 (0.0014, 0.0020)	2.06 (2.11)
θ_{Aga}	0.0041 (0.0036, 0.0045)	1.89 (1.93)	1.77 (1.80)	0.0041 (0.0036, 0.0046)	1.85 (1.89)
θ_{Aza}	0.0061 (0.0052, 0.0070)	3.63 (3.85)	2.99 (3.16)	0.0061 (0.0052, 0.0071)	3.01 (3.17)
θ_{Ape}	0.0029 (0.0025, 0.0033)	4.15 (4.31)	4.15 (4.30)	0.0029 (0.0025, 0.0033)	3.56 (3.68)
θ_{Ama}	0.0044 (0.0040, 0.0049)	1.55 (1.65)	1.64 (1.73)	0.0044 (0.0039, 0.0049)	1.64 (1.74)
θ_{Smi}	0.0083 (0.0074, 0.0093)	2.40 (2.96)	1.88 (2.20)	0.0084 (0.0074, 0.0094)	2.18 (2.66)
θ_r	0.0354 (0.0325, 0.0384)	1.49 (1.53)	1.11 (1.12)	0.0351 (0.0322, 0.0380)	1.72 (1.77)
θ_a	0.0227 (0.0194, 0.0261)	0.44 (0.44)	0.25 (0.26)	0.0278 (0.0256, 0.0299)	2.01 (2.01)
θ_b	0.0244 (0.0223, 0.0265)	0.72 (0.72)	0.67 (0.67)	0.0309 (0.0227, 0.0400)	2.73 (2.73)
θ_c	0.0228 (0.0174, 0.0282)	0.88 (0.88)	1.14 (1.14)	0.0235 (0.0178, 0.0291)	1.89 (1.89)
θ_d	0.0120 (0.0106, 0.0134)	2.73 (2.77)	1.66 (1.68)	0.0117 (0.0103, 0.0131)	1.55 (1.56)
θ_e	0.0291 (0.0230, 0.0354)	1.52 (1.52)	1.23 (1.24)	0.0266 (0.0198, 0.0336)	3.08 (3.08)
θ_f	0.0150 (0.0138, 0.0162)	3.35 (3.41)	1.87 (1.89)	0.0150 (0.0139, 0.0162)	2.86 (2.91)
θ_g	0.0374 (0.0210, 0.0568)	10.48 (10.60)	6.80 (7.03)	0.0344 (0.0210, 0.0495)	5.64 (5.82)
φ_{xy}	0.428 (0.352, 0.506)	0.71 (0.71)	0.44 (0.44)	–	–
ϖ_{xy}	–	–	–	20.0 (4.0, 37.0)	1.16 (1.16)

Note.— Relative efficiency is measured by the ratio of variances for the two algorithms for estimating the posterior mean of the parameter. For θ, φ and ϖ, the conditionals can be used to provide more efficient estimate ($\tilde{\mu}$ instead of $\tilde{y}$ in SI text C), with efficiency shown in parentheses. For example, under MSC-I, the estimate of posterior mean of θ_g from algorithm A2 (I-IG-gIG) is 10.48 times as efficient as that from A3 (I-IG-slide), and the estimate from A2 based on the conditional mean ($\tilde{u}$) is 10.60 times as efficient. See table S1 for details of the algorithms.

Table S4: Posterior means (and 95% HPD CIs) and relative mixing efficiency of MCMC algorithms applied to the *Anopheles* dataset

	MSC-I			MSC-M	
	Mean (CI)	E_{A2}/E_{A3}	E_{A5}/E_{A6}	Mean (CI)	E_{B4}/E_{B6}
τ_o	0.0155 (0.0153, 0.0157)	1.94	1.81	0.0155 (0.0153, 0.0157)	1.99
τ_g	0.0093 (0.0090, 0.0096)	1.01	1.16	–	–
τ_a	0.0149 (0.0136, 0.0156)	1.17	1.32	0.0126 (0.0123, 0.0128)	4.51
τ_c	0.0133 (0.0130, 0.0135)	5.94	5.76	0.0125 (0.0122, 0.0127)	5.54
τ_d	0.0095 (0.0093, 0.0096)	2.60	2.32	0.0103 (0.0102, 0.0105)	9.53
τ_e	0.0064 (0.0063, 0.0065)	1.07	1.62	–	–
τ_b	0.0027 (0.0026, 0.0029)	2.08	2.18	0.0036 (0.0035, 0.0037)	9.68
θ_G	0.0575 (0.0481, 0.0675)	16.62 (17.81)	15.72 (16.92)	0.0544 (0.0472, 0.0622)	12.31 (13.34)
θ_C	0.0367 (0.0319, 0.0416)	5.03 (5.46)	5.13 (5.55)	0.0361 (0.0322, 0.0403)	5.38 (5.91)
θ_R	0.0047 (0.0045, 0.0049)	30.32 (35.59)	31.92 (37.81)	0.0047 (0.0045, 0.0049)	22.06 (26.18)
θ_L	0.0030 (0.0028, 0.0031)	38.38 (42.35)	44.20 (49.08)	0.0029 (0.0028, 0.0031)	48.60 (57.01)
θ_A	0.0093 (0.0090, 0.0097)	4.96 (5.02)	2.88 (2.90)	0.0087 (0.0084, 0.0090)	4.29 (4.31)
θ_Q	0.0100 (0.0096, 0.0104)	7.35 (9.12)	6.69 (8.24)	0.0103 (0.0099, 0.0107)	6.03 (8.25)
θ_o	0.0143 (0.0137, 0.0149)	1.56 (1.57)	1.73 (1.73)	0.0142 (0.0137, 0.0148)	3.11 (3.14)
θ_a	0.0148 (0.0032, 0.0282)	7.33 (7.34)	6.93 (6.93)	0.0186 (0.0165, 0.0208)	4.68 (4.68)
θ_c	0.0131 (0.0074, 0.0178)	2.60 (2.60)	10.16 (10.17)	0.0353 (0.0041, 0.0743)	50.87 (51.00)
θ_d	0.0124 (0.0115, 0.0134)	3.05 (3.06)	2.47 (2.47)	0.0069 (0.0060, 0.0078)	9.57 (9.58)
θ_b	0.0135 (0.0127, 0.0142)	1.54 (1.55)	2.83 (2.84)	0.0056 (0.0052, 0.0059)	7.62 (7.64)
$\varphi_{R\to Q}$	0.0164 (0.0085, 0.0246)	1.06 (1.07)	0.96 (0.97)	–	–
$\varphi_{A\to GC}$	0.9634 (0.9508, 0.9751)	1.25 (1.25)	1.48 (1.48)	–	–
$\varpi_{R\to Q}$	–	–	–	0.1370 (0.0073, 0.2912)	1.04 (1.04)
$\varpi_{A\to GC}$	–	–	–	2032.0 (190.6, 214.7)	9.08 (9.10)

Note.— See notes for table S3.

Table S5: Posterior means (and 95% HPD CIs) and relative mixing efficiency of MCMC algorithms applied to the *Heliconius* dataset

	MSC-I			MSC-M	
	Mean (CI)	E_{A2}/E_{A3}	E_{A5}/E_{A6}	Mean (CI)	E_{B4}/E_{B6}
τ_r	0.0118 (0.0116, 0.0119)	1.20	1.22	0.0116 (0.0114, 0.0117)	2.40
τ_s	0.0054 (0.0050, 0.0057)	2.50	1.06	0.0010 (0.0008, 0.0012)	17.02
τ_y	0.0002 (0.0001, 0.0003)	3.30	1.34	–	–
θ_H	0.0134 (0.0129, 0.0138)	2.29 (3.54)	2.63 (4.11)	0.0131 (0.0127, 0.0136)	8.74 (12.55)
θ_C	0.0438 (0.0401, 0.0475)	3.15 (3.15)	1.31 (1.31)	0.0408 (0.0328, 0.0494)	6.89 (7.21)
θ_M	0.0008 (0.0006, 0.0010)	4.10 (4.10)	1.54 (1.54)	0.0026 (0.0021, 0.0031)	19.32 (19.33)
θ_r	0.0122 (0.0118, 0.0127)	1.42 (1.44)	1.48 (1.49)	0.0124 (0.0119, 0.0128)	1.81 (1.82)
θ_s	0.0183 (0.0173, 0.0194)	2.28 (2.28)	0.97 (0.97)	0.0342 (0.0327, 0.0357)	15.82 (15.87)
φ_{cm}	0.4648 (0.4294, 0.4998)	1.10 (1.10)	1.01 (1.01)	–	–
ϖ_{cm}	–	–	–	2.2526 (0.0483, 5.3840)	1.40 (1.40)

Note.— See notes for table S3.

Table S6: *Adansonia* baobabs accession numbers (based on ref. [41, table S1])

tAXON	Sample ID	Accession #
Adansonia digitata L.	Adi001	UW11 (UWBG)
Adansonia digitata L.	Adi002	UW2291 (UWBG)
Adansonia digitata L.	Adi003	GenBank: KU145771
Adansonia grandidieri Baill.	Aga001	97-B002010-1 (GBDBG)
Adansonia grandidieri Baill.	Aga002	03-B000192-1 (GBDBG)
Adansonia gregorii F.Muell.	Age001	North Western Australia, D.A.Baum
Adansonia perrieri Capuron	Ape001	92-B000060-1 (GBDBG)
Adansonia perrieri Capuron	Ape009	Karimi-2014-09 (WIS)
Adansonia madagascariensis Baill.	Ama006	Karimi-2014-006 (WIS)
Adansonia madagascariensis Baill.	Ama018	Karimi-2014-018 (WIS)
Adansonia rubrostipa Jum. & Perr.	Aru001	D.A.Baum 313 (MO)
Adansonia rubrostipa Jum. & Perr.	Aru127	Karimi-2014-127 (WIS)
Adansonia suarezensis H.Perrier	Asu001	UW11 (GBDBG)
Adansonia za Baill.	Aza037	Karimi-2014-37 (WIS)
Adansonia za Baill.	Aza135	Karimi-2014-135 (WIS)
Bombax ceiba L.	Bce020	UW10 (UWBG)
Pseudobombax croizatii A.Robyns.	Pcr070	UW1255 (UWBG)
Scleronema micrantha Ducke	Smi165	GenBank: AF111735

Table S7: Accession numbers for genomic sequences from 12 samples of *Anopheles* mosquitoes

Species	Sample ID	Accession #
Six genomes from ref. [10, table S2]		BioSample #
Anopheles gambiae	40.2	SAMN02899205
Anopheles coluzzii	C27.2	SAMN02899195
Anopheles arabiensis	4080	SAMN03083367
Anopheles quadriannulatus	72	SAMN01760635
Anopheles merus	Mpug686i	SAMN01760628
Anopheles melas	CM1001067	SAMN01760621
Six genomes from ref. [11, table S4]		GenBank Assembly
Anopheles gambiae	AgamS1	GCA_000150785.1
Anopheles coluzzii	AcolM1	GCA_000150765.1
Anopheles arabiensis	AaraD1	GCA_000349185.1
Anopheles quadriannulatus	AquaS1	GCA_000349065.1
Anopheles merus	AmerM1	GCA_000473845.1
Anopheles melas	AmelC1	GCA_000473525.1

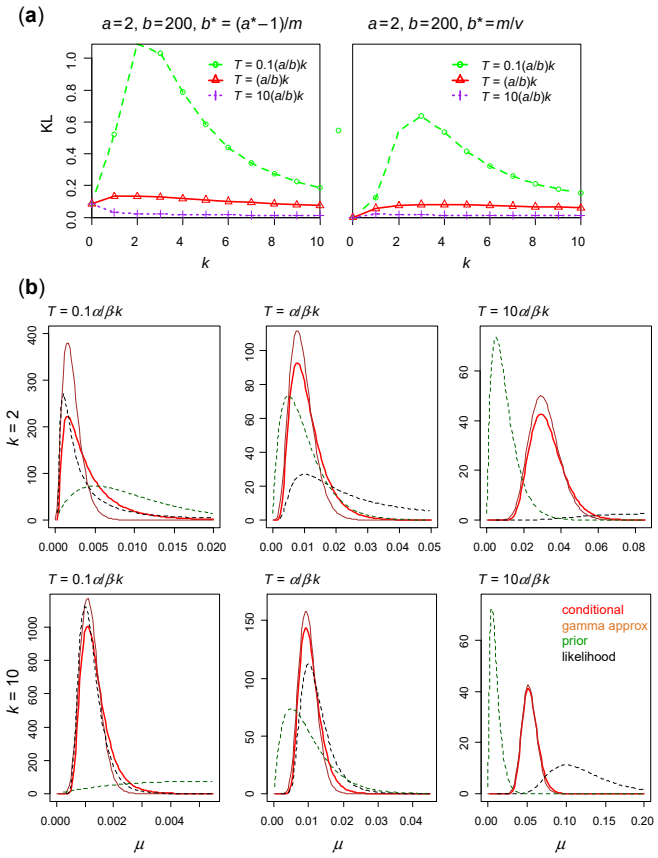

Fig. S1: (**a**) KL divergence between the true conditional and the gamma approximation. (**b**) The conditional for Poisson waiting time (μ) under the gamma prior $\mu \sim \mathrm{G}(a, b)$ with $a = 2, b = 200$, and approximate gamma conditional, when the data consist of $k = 2$ or 10 events observed over time period T. Time T is fixed at 0.1×, 1×, 10× $a/b \cdot k$ so that the sample mean T/k is 0.1, 1, or 10 times the prior mean. The likelihood (eq. S1) is the density for $\mathrm{IG}(k-1, T)$.

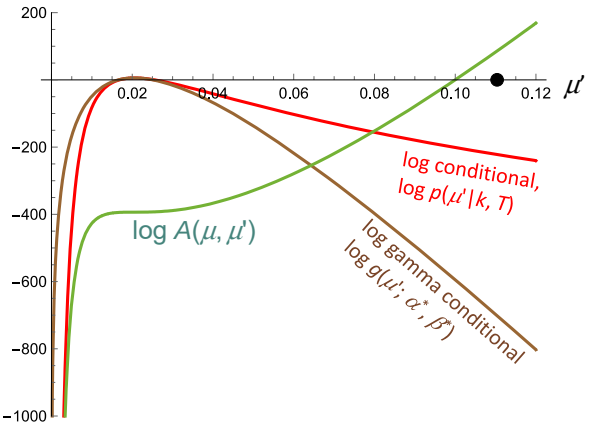

Fig. S2: Gamma approximate conditional with a light tail may lead to rejection of the approximate Gibbs proposal causing convergence difficulties when the chain is far out in the tail. Here the prior is $\mathrm{G}(a, b)$ with $a = 2, b = 20$ with mean 0.1. The true conditional is given by eq. S2, with $k = 155, T = 12.71167$, with the mode around 0.02 (The data were generated with the true parameter $\mu = 0.02$). The approximate gamma conditional is $\mathrm{G}(a^*, b^*)$ with $a^* = 263.8, b^* = 12792$ (eq. S5). The current value is $\mu = 0.10924$ (the solid dot), as the chain started with an initial value sampled from the prior. With the approximate Gibbs proposal, the proposed value μ' is around 0.02, near the mode of the conditional. The gamma conditional has a lighter tail than the true conditional, so that $\frac{g(\mu';a^*,b^*)}{g(\mu;a^*,b^*)} \gg \frac{p(\mu'|k,T)}{p(\mu|k,T)}$ for such values of μ and μ'. As a result the proposal is rejected because the acceptance ratio $A(\mu, \mu') = \frac{g(\mu;a^*,b^*)}{g(\mu';a^*,b^*)} \cdot \frac{p(\mu'|k,T)}{p(\mu|k,T)} \ll 1$ (or $\log A \ll 0$).

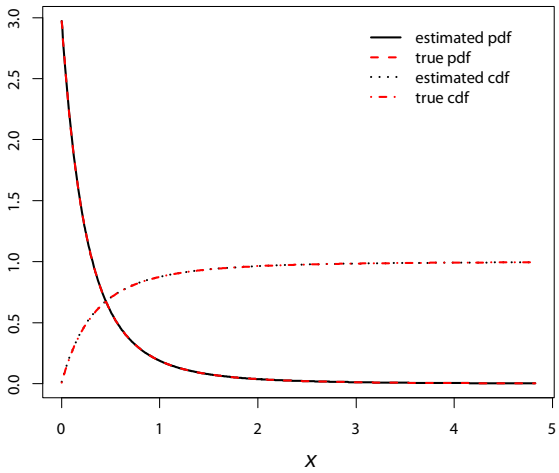

Fig. S3: Approximation to the marginal CDF and PDF by using the conditionals sampled during the MCMC.