In our analyses of datasets, we considered two different mutational models, namely: K80 and F84 models.

**K80 model:**

Kimura’s two-parameter, K80 model allows for two parameters, different rates for transitions (changes involving purine-to-purine or pyrimidine-to-pyrimidine) and transversions (purine-to-pyrimidine and vice-versa) i.e., it distinguishes between the rate of transitions and transversions using the $\kappa$ parameter. In addition, the nucleotide frequencies are the same i.e $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$. The three types of changes are:

\[
P(t)_{ii} = \frac{1}{4} + \frac{3}{4} e^{-\mu t},
\]
\[
P(t)_{ij, \text{Transition}} = \frac{1}{4} + \frac{1}{4} e^{-\mu t} + \frac{1}{2} e^{-\mu t \left(\frac{\kappa + 1}{2}\right)},
\]
\[
P(t)_{ij, \text{Transversion}} = \frac{1}{4} - \frac{1}{4} e^{-\mu t},
\]

where $P(t)_{ii}$ is the probability of a nucleotide to remain the same during the evolutionary time $t$ and $P(t)_{ij}$ is the probability of a nucleotide changing from initial state $i$ to final state $j$ during the evolutionary time $t$.  

*to whom correspondence should be addressed
F84 model:

F84 model distinguishes between the rate of transitions and transversions using the $\kappa$ parameter and in addition, it allows unequal nucleotide frequencies i.e., $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$. Thus, both $\kappa$ and the nucleotide frequencies are to be estimated when F84 model is considered. The three types of changes are:

\[
P(t)_{jj} = \pi_j + \pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\mu t} + \left( \frac{\Pi_j - \pi_j}{\Pi_j} - 1 \right) e^{-\mu t (\kappa + 1)},
\]

\[
P(t)_{ij,\text{Transition}} = \pi_j + \pi_j \left( \frac{1}{\Pi_j} - 1 \right) e^{-\mu t} - \left( \frac{\pi_j}{\Pi_j} - 1 \right) e^{-\mu t (\kappa + 1)},
\]

\[
P(t)_{ij,\text{Transversion}} = \pi_j (1 - e^{-\mu t}),
\]

where $\Pi_j = \pi_A + \pi_G$ if $j$ is a purine (A or G) and $\Pi_j = \pi_C + \pi_T$ if $j$ is a pyrimidine (C or T).

Note S3

Here, we present the prior distributions for all the unknown parameters in the Bayesian model that we considered.

Prior density of $\Theta$:

$\Theta$ is a positive parameter and we impose a uniform distribution in the interval $[0, \Theta_{\text{max}}]$, i.e.,

\[
\Theta \sim \mathcal{U}(0, \Theta_{\text{max}}),
\]

with the probability density function (PDF) given as

\[
p(\Theta) = \begin{cases} \frac{1}{\Theta_{\text{max}}}, & \text{for } \Theta \in [0, \Theta_{\text{max}}] \\ 0, & \text{otherwise.} \end{cases}
\]

Table 1 shows the values of $\Theta_{\text{max}}$ considered in the experiments with simulated datasets and the mitochondrial DNA sequence data (mtDNA data). When analyzing real biological datasets, results obtained from some other estimation approaches can provide knowledge about the range of $\Theta_{\text{max}}$. 
Simulated data: $\Theta = 0.01$ & $\Theta_{\text{max}}$ for the prior distribution
Simulated data: $\Theta = 0.10$ & 1.00
Simulated data: $\Theta = 0.50$ & 1.00
mtDNA data & 1.00

Table 1: Chosen values of $\Theta_{\text{max}}$ for all datasets.

Prior density of $\lambda$:

$\lambda$ is the set of unknown parameters associated with each mutational model and as stated earlier, the mutational models considered are K80 and F84 models. K80 model has only one unknown parameter i.e., the transition-transversion ratio $\kappa$ while the F80 model has two unknown parameters i.e., $\kappa$ and the nucleotide frequencies $\pi$.

Prior density of $\kappa$:

Like $\Theta$, $\kappa$ is a positive parameter and we impose a uniform distribution in the interval $[0, \kappa_{\text{max}}]$, i.e.,

$$\kappa \sim U(0, \kappa_{\text{max}}),$$

and the PDF is given as

$$p(\kappa) = \begin{cases} \frac{1}{\kappa_{\text{max}}}, & \text{for } \kappa \in [0, \kappa_{\text{max}}] \\ 0, & \text{otherwise.} \end{cases}$$

Table 2 shows the values of $\kappa_{\text{max}}$ considered in the experiments with simulated datasets and the mtDNA dataset.

| $\kappa_{\text{max}}$ for the prior distribution |
|--------------------------|
| Simulated data: $\kappa = 2$ & 10
| mtDNA data & 200

Table 2: In simulated datasets, true $\kappa = 2$ and $\kappa_{\text{max}} = 10$. 
Prior density of $\pi$:

Here, we impose Dirichlet distribution on the nucleotide frequencies $\pi$ i.e.,

$$\pi \sim \text{Dir}(\alpha),$$

and the PDF is given by

$$p(\pi|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \pi_i^{\alpha_i - 1}, \quad i \in \{A,C,G,T\},$$

where $\Gamma(\cdot)$ is the gamma function, $\alpha = (\alpha_A, \alpha_C, \alpha_G, \alpha_T)$ are the concentration parameters and $\pi_A + \pi_C + \pi_G + \pi_T = 1$. In our experiments, we chose $\alpha = \beta_0 \pi_{\text{emp}}$, where $\pi_{\text{emp}}$ is the empirical value of the nucleotide frequencies calculated from the data and $\beta_0$, a scalar value, chosen to be 10, is used to control the peakedness of the prior distribution of $\pi$ around $\pi_{\text{emp}}$ [1].

Prior density of the genealogy:

Algorithm 1 describes the procedure to sample from the distribution of the possible genealogies of $m$ copies in a population with effective population size $N_e$.

Algorithm 1 The coalescent: Simulation of a random genealogy.
1: Start with $m$ lineages.
2: Set $k = m$ and $T = 0$.
3: Draw a random quantity $t_k$ from an exponential distribution with expectation $4N_e/(k(k-1))$.
4: Randomly pick two of the $k$ copies of the gene, without replacement.
5: Create a node of the genealogical tree which is the immediate common ancestor of these two copies, and which existed $T + t_k$ generations before the present.
6: Set $T = T + t_k$.
7: Replace these two copies by this common ancestor and set $k = k - 1$.
8: If $k = 1$ we are done. Otherwise return to step 3.

Note S4

Obtaining the weight at time $t$ from the weights at time $t-1$

In the main text, it was shown that the set of weighted samples $\{\mathcal{H}_{t-1}^n, w_{t-1}^n\}_{n=1}^N$ approximates the joint target distribution $\tilde{\pi}_{t-1}$. After propagating the sam-
where normalized weights at time \( t \) are calculated as follows:

\[
\tilde{w}^n_t \propto \frac{\tilde{\pi}_t(\mathcal{H}^n_{t,d})}{q_t(\mathcal{H}^n_{t,d})} = \frac{\pi_t(\mathcal{H}^n_t) \prod_{d=1}^{T} L_d(H^n_d, H^n_t)}{q_1(\mathcal{H}^n_t) \prod_{r=2}^{T} \mathcal{K}_r(H^n_{r-1}, H^n_r)} = \frac{\pi_t(\mathcal{H}^n_t) \mathcal{L}_t-1(H^n_t, H^n_{t-1}) \prod_{d=1}^{T} L_d(H^n_d, H^n_t)}{q_1(\mathcal{H}^n_t) \mathcal{K}_t(H^n_{t-1}, H^n_t)} \]

\[
\tilde{w}^n_t \propto \tilde{w}^n_{t-1} \frac{\pi_t(\mathcal{H}^n_t) \mathcal{L}_t-1(\mathcal{H}^n_t, \mathcal{H}^n_{t-1})}{\pi_{t-1}(\mathcal{H}^n_{t-1}) \mathcal{K}_t(\mathcal{H}^n_{t-1}, \mathcal{H}^n_t)} ,
\]

from the definitions of \( \pi_t \) and \( \pi_{t-1} \) in the main text and noticing that \( Z_t \) and \( Z_{t-1} \) are constants with respect to \( \mathcal{H}^n_t \) and \( \mathcal{H}^n_{t-1} \), then

\[
\tilde{w}^n_t \propto \tilde{w}^n_{t-1} \frac{\Psi_t(\mathcal{H}^n_t) \mathcal{L}_t-1(\mathcal{H}^n_t, \mathcal{H}^n_{t-1})}{\Psi_{t-1}(\mathcal{H}^n_{t-1}) \mathcal{K}_t(\mathcal{H}^n_{t-1}, \mathcal{H}^n_t)} = \tilde{w}^n_{t-1} W_t(\mathcal{H}^n_{t-1}, \mathcal{H}^n_t), \quad n = 1, ..., N,
\]

where \( \{\tilde{w}^n_{t-1}\}_{n=1}^N \) are the unnormalized weights at time \( t-1 \) and \( \{W_t(\mathcal{H}^n_{t-1}, \mathcal{H}^n_t)\}_{n=1}^N \) the unnormalized incremental weights, calculated as

\[
W_t(\mathcal{H}^n_{t-1}, \mathcal{H}^n_t) = \frac{\Psi_t(\mathcal{H}^n_t) \mathcal{L}_t-1(\mathcal{H}^n_t, \theta^n_{t-1})}{\Psi_{t-1}(\mathcal{H}^n_{t-1}) \mathcal{K}_t(\theta^n_{t-1}, \mathcal{H}^n_t)}, \quad n = 1, ..., N.
\]

**Target Distributions, Forward and Backward Kernels Specification Population Parameters Estimation**

Here, we specify the exact form of the sequence of target distributions \( \{\pi_t\}_{t=1}^T \), the forward kernels, \( \{\mathcal{K}_t\}_{t=2}^T \) and the backward kernels \( \{\mathcal{L}_{t-1}\}_{t=2}^T \).

**Sequence of target distributions and forward kernels:**

Since we are interested in the likelihood tempered target sequence we sample initially from the prior distribution \( \pi_1 = p(\mathcal{H}) \) and gradually introduce the effect of the likelihood in order to obtain an approximation of the posterior distribution \( p(\mathcal{H} | \mathcal{D}) \), i.e, \( \pi_T \) at \( t = T \). Although, there are a few choices of \( \mathcal{K}_t \) that have been proposed, we employ MCMC kernels of invariant distribution
\( \pi_t, K_t \). There is more freedom in specifying the forward kernels in SMC since kernels do not need to be reversible or even Markov. As such, we will employ a hybrid Gibbs and Metropolis-Hastings (M-H) samplers that allows local moves in order to successively move all the parameters of interest in \( \mathcal{H} \) i.e., \( \Theta, \Upsilon \) and \( \lambda \). The set of proposal distributions for the genealogy and other parameters are discussed below.

**Sequence of backward kernels:**

To obtain a good performance of the SMC samplers algorithm, the backward kernel \( L_t \) is often optimized with respect to the forward kernel \( K_t \). Thus, if an MCMC kernel is used as forward kernel, then the following \( L_t \) is employed:

\[
L_{t-1}(\mathcal{H}_t, \mathcal{H}_{t-1}) = \frac{\pi_t(\mathcal{H}_{t-1})K_t(\mathcal{H}_t, \mathcal{H}_{t-1})}{\pi_t(\mathcal{H}_t)},
\]

a good approximation of the optimal backward kernel when the discrepancy between \( \pi_t \) and \( \pi_{t-1} \) is small. Thus, since the chosen \( L_t \) is the reversal Markov kernel associated with \( K_t \), then the unnormalized incremental weights become:

\[
W_t(\mathcal{H}^n_{t-1}, \mathcal{H}^n_t) = \frac{\Psi_t(\mathcal{H}^n_t)\pi_t(\mathcal{H}^n_{t-1})}{\Psi_{t-1}(\mathcal{H}^n_{t-1})\pi_t(\mathcal{H}^n_t)} = \frac{p(\mathcal{H}^n_t)p(D|\mathcal{H}^n_t)^e_t p(\mathcal{H}^n_{t-1}) p(D|\mathcal{H}^n_{t-1})^{e_t}}{p(\mathcal{H}^n_{t-1})p(D|\mathcal{H}^n_{t-1})^{e_t-1} p(\mathcal{H}^n_t)p(D|\mathcal{H}^n_t)^e_t} = p(D|\mathcal{H}^n_{t-1})^{(e_t - e_{t-1})}, \ n = 1, ..., N.
\]

**Resampling Procedure**

Resampling is performed when the ESS is significantly less than the number of samples, discarding the ineffective samples and then multiply the effective ones. In all our experiments, we performed resampling when the ESS is less than \( N/10 \). The resampling procedure is briefly summarized as follows:

- Interpret each weight \( w^n_t \) as the probability of obtaining the sample index \( n \) in the set \( \{ \mathcal{H}^n_t : n = 1, ..., N \} \).
- Draw \( N \) samples from the discrete probability distribution and replace the old sample set with this new one.
- Set all weights to the constant value \( w^n_t = 1/N \).
Note S5

Here, we present the hybrid Gibbs and Metropolis-Hastings algorithm of invariant distribution $\pi_t$ employed to move the particles in the SMC algorithm presented in the main text. In the algorithm, $V$ is the total number of parameters and $R_{MCMC}$ is the chain length for each sample (particle).

**Algorithm 2** Hybrid Gibbs and M-H Kernel $K(\cdot; \cdot)$ for the $n$-th sample

1: Initialization Set $H^0 = [\sigma_1^0, ..., \sigma_V^0] = H^t_{t-1} = [\sigma_{1,t-1}^n, ..., \sigma_{V,t-1}^n]

2: for $r = 1, ..., R_{MCMC}$ do

3: for $v = 1, ..., V$ do

4: (a) Sample $\sigma_v^* \sim h_v(\sigma_v^*|\sigma_v^{r-1})$, where $h_v(\cdot|\cdot)$ is the proposal distribution for the $v^{th}$ parameter (See below for details).

5: (b) Compute the acceptance ratio:

6: $\alpha = \min \left\{ 1, \frac{Pr(D|H^*)^{t}p(H^*)}{Pr(D|H^{r-1})^{t}p(H^{r-1})} \frac{h_v(\sigma_v^{r-1}|\sigma_v^*)}{h_v(\sigma_v^*|\sigma_v^{r-1})} \right\}$,

7: where $H^* = [\sigma_1^*, ..., \sigma_v^*, \sigma_{v+1}^*, ..., \sigma_V^*]$, and $H^{r-1} = [\sigma_1^{r-1}, ..., \sigma_v^{r-1}, \sigma_{v+1}^{r-1}, ..., \sigma_V^{r-1}]$

8: (c) Sample a random variate $u \sim U(0, 1)$

9: if $u \leq \alpha$ then

10: $\sigma_v^* = \sigma_v^*$

11: else

12: $\sigma_v^* = \sigma_v^{r-1}$

13: end if

14: end for

15: end for

16: Set the new particle value at time $t$ as $H^n_t = [\sigma_1^{R_{MCMC}}, ..., \sigma_V^{R_{MCMC}}]$.

Proposal distributions:

Next, we present the proposal distributions $h_v(\cdot|\cdot)$ in Algorithm 2 for all the parameters in $H$.

Proposal distribution for $\Theta$:

The update algorithm here is similar to the one described in [2]. Let the previous value in the Markov chain be $\Theta_{r-1}$; a proposal value, $\Theta^*$ is drawn
from a uniform distribution in the interval $[\Theta_{r-1} - \Delta_\Theta, \Theta_{r-1} + \Delta_\Theta]$, i.e.,

$$\Theta^* \sim U(\Theta_{r-1} - \Delta_\Theta, \Theta_{r-1} + \Delta_\Theta),$$

where $\Delta_\Theta$ is chosen such that $\Delta_\Theta < \Theta_{max}$ (note: $\Theta_{max}$ is defined in the prior distribution.). Thus, if $\Theta^* < 0$, set $\Theta^* = -\Theta^*$, and if $\Theta^* > \Theta_{max}$, set $\Theta^* = 2\Theta_{max} - \Theta^*$.

**Proposal distribution for $\kappa$:**

Similar to the update algorithm for $\Theta$, let the previous value in the Markov chain be $\kappa_{r-1}$; a proposal value, $\kappa^*$ is drawn from a uniform distribution in the interval $[\kappa_{r-1} - \Delta_\kappa, \kappa_{r-1} + \Delta_\kappa]$, i.e.,

$$\kappa^* \sim U(\kappa_{r-1} - \Delta_\kappa, \kappa_{r-1} + \Delta_\kappa),$$

where $\Delta_\kappa$ is chosen such that $\Delta_\kappa < \kappa_{max}$ ($\kappa_{max}$ is defined in the prior distribution for $\kappa$). Therefore, if $\kappa^* < 0$, set $\kappa^* = -\kappa^*$, and if $\kappa^* > \kappa_{max}$, set $\kappa^* = 2\kappa_{max} - \kappa^*$.

**Proposal distribution for $\pi$:**

Given that the previous value of $\pi$ in the Markov chain is denoted by $\pi_{r-1}$, a proposal value $\pi^*$ is drawn as follows:

$$\pi^* \sim \text{Dir}(\beta_1 \pi_{r-1}),$$

$\text{Dir}(\cdot)$ is a Dirichlet distribution with concentration parameters given as $\beta_1 \pi_{r-1} = (\beta_1 \pi_{A,r-1}, \beta_1 \pi_{C,r-1}, \beta_2 \pi_{G,r-1}, \beta_1 \pi_{T,r-1})$. Similar to $\beta_0$, $\beta_1$ is used to control the peakedness of the transition kernel around the previously sampled value $\pi_{r-1}$. $\beta_1 = 100$ in all the experiments.

**Proposal distribution for genealogy:**

The conditional coalescent proposal for the genealogy is well described in [2, 3, 4] and the algorithms can easily be implemented.

**References**

[1] Erkkilä, T., et al. (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics*, 26, 2571–2577.
[2] Nielsen, R., Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

[3] Felsenstein, J., et al. (1999) Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. *Lecture Notes-Monograph Series*, 163–185.

[4] Stephens, M. (2001) Inference under the coalescent. *Handbook of statistical genetics*. 