Supplementary Information

Robust and scalable inference of population history from hundreds of unphased whole-genomes

Jonathan Terhorst[1], John A. Kamm[1,2] and Yun S. Song[1,2,3,4,*]

[1] Department of Statistics, University of California, Berkeley, CA 94720, USA
[2] Computer Science Division, University of California, Berkeley, CA 94720, USA
[3] Department of Integrative Biology, University of California, Berkeley, CA 94720, USA
[4] Departments of Biology and Mathematics, University of Pennsylvania, PA 19104, USA
[*] To whom correspondence should be addressed.

November 9, 2016

Supplementary Note

S1 Emission Probabilities

In this section we present results and proofs needed to compute the emission probabilities described in the main text. To simplify the equations, we depart slightly from the notation in the main text and now assume that n "undistinguished" + 2 "distinguished" haploids are sampled from a single panmictic population for an overall sample size of $n+2$ haploids. The two distinguished lineages are labeled 1 and 2. The effective size of this population at t coalescent units in the past was $1/\alpha(t)$, and we define the cumulative coalescence intensity function $R(t) := \int_0^t \alpha(s)\,ds$. The random tree describing the ancestry of these samples will be referred to as simply "the coalescent" throughout this section, despite the inhomogeneous rate function. Following standard terminology, we say that a branch of this tree has *size k* if it is ancestral to k sampled individuals at the present. The samples are genotyped at a single biallelic site, and the allelic state (number of derived alleles) at this site is denoted $(a, b) \in \{0, 1, 2\} \times \{0, 1, \dots, n\}$. Finally, we set $a_m \overset{\text{def}}{=} \binom{m}{2}$ throughout.

Let C_{12} denote the (random) time at which 1 and 2 coalesce. Given $C_{12} = \tau$, the *conditioned site frequency spectrum* $\text{CSFS}(\tau)$ is a $3\times n$ matrix whose (a, b)-th entry is the expectation, conditional on $C_{12} = \tau$, for the total branch length subtending a distinguished and b undistinguished lineages. Under neutrality, mutations occur as a rate-$\frac{\theta}{2}$ point process with respect to branch length, and so $\text{CSFS}(\tau)$ is also the leading-order coefficient in the Taylor expansion for the conditional probability of observing (a, b):

$$\mathbb{P}((a,b) \mid C_{12} = \tau) = \frac{\theta}{2} \cdot [\text{CSFS}(\tau)]_{ab} + O(\theta^2),\ 0 < a + b < n + 2.$$

Since the per-site mutation rate is small, i.e. $\frac{\theta}{2} \ll 1$, SMC++ approximates the emission probability by

$$\mathbb{P}((a,b) \mid C_{12} = \tau) \approx \frac{\theta}{2} \cdot [\mathrm{CSFS}(\tau)]_{ab}$$

which is essentially equivalent to assuming at most 1 mutation per site.

To simplify the derivation, we decompose CSFS(τ) into the branch lengths above and below τ and derive each quantity separately:

$$\mathrm{CSFS}(\tau) = \underbrace{\mathrm{CSFS}(\tau \downarrow)}_{\text{below } \tau} + \underbrace{\mathrm{CSFS}(\tau \uparrow)}_{\text{above } \tau}.$$

The strategy in both cases is similar: we first compute the total expected branch length in the coalescent on $n+2$ lineages, and then rely on combinatorial arguments to calculate the probability that a branch of size $2 \le k < n+2$ subtends a sample of (a,b) distinguished and undistinguished lineages.

S1.1 The conditioned coalescent

We start with an explicit generative model for the conditioned coalescent given C_{12}, which will be useful for computing both CSFS($\tau \downarrow$) and CSFS($\tau \uparrow$). The conditioned coalescent given $\{C_{12} = \tau\}$, denoted as $\mathcal{T}_{n+2}$, can be constructed via a sequence of subtrees $\mathcal{T}_2, \mathcal{T}_3, \dots, \mathcal{T}_{n+2}$, where $\mathcal{T}_k$ is the genealogy relating the first k leaves $\{1, 2, \dots, k\}$. $\mathcal{T}_k$ is obtained by recursively adding the kth leaf to the subtree $\mathcal{T}_{k-1}$. Specifically:

1. Initialize $\mathcal{T}_2$ as a cherry (tree with two leaves) with fixed height $A_2 = \tau$.
2. For $k = 3, \dots, n+2$:
   (a) Let $\sigma : \{2, \dots, k-1\} \to \{2, \dots, k-1\}$ be the permutation which sorts $A_2, \dots, A_{k-1}$, in descending order, $A_{\sigma(2)} \ge A_{\sigma(3)} \ge \cdots \ge A_{\sigma(k-1)}$, and set $U_j = A_{\sigma(j)}$. Additionally define $U_1 = \infty$ and $U_k = 0$.

   (b) Sample A_k from an inhomogeneous point process with piecewise intensity function

   $$\mu(t) := j \times \alpha(t), \quad U_{j+1} \le t < U_j.$$

   (c) Select a branch uniformly at random in $\mathcal{T}_{k-1}$ from among those surviving to time A_k, and form $\mathcal{T}_k$ by adding a new external branch (labeled k) to $\mathcal{T}_{k-1}$ which coalesces with the chosen branch at A_k.

To see why $\mathcal{T}_{n+2}$ is distributed as the coalescent conditional on $\{C_{12} = \tau\}$, suppose we replace step 1, and instead sample $\mathcal{T}_2$ as a random cherry whose height is given by an exponential waiting time with inhomogeneous rate $\alpha(t)$. Then it is straightforward to verify that at time t, every pair of lineages is coalescing at rate $\alpha(t)$, i.e. the resulting tree is distributed as the (unconditioned) coalescent.

We note that $\mathcal{T}_{n+2}$ can also be generated backwards in time, as follows. For $t < \tau$, the ancestors of the distinguished pair cannot coalesce with each other, but every other pair coalesces at rate $\alpha(t)$. So the total coalescence rate if there are k lineages is $(a_k - 1)\alpha(t)$. For $t > \tau$, every pair of lineages coalesces at rate $\alpha(t)$, and the total coalescence rate for k lineages is $a_k\alpha(t)$.

S1.2 Computing CSFS($\tau \downarrow$)

In this section we compute the distribution of allelic configurations for a mutation occuring below τ, the time at which the lineages 1 and 2 are conditioned to coalesce. We first fix some notation. We denote by $\xi_{m,k}$ be the partition structure of $\mathcal{T}_m$ when it has $2 \le k \le m$ lineages remaining: $|\xi_{m,k}| = k$. Similarly, $T_{m,k}$ denotes the amount of time in $\mathcal{T}_m$ during which k lineages remain. For $\tau \ge 0$ let $\mathscr{A}_m(\tau)$ denote the "ancestral process", i.e. the pure-death process which gives the number of lineages remaining in $\mathcal{T}_m$ at time τ.

Throughout this section we make repeated use of the event $E_{m,k} \overset{\text{def}}{=} \{\mathscr{A}_m(\tau-) \le k\}$. The truncated times

$$T^{\tau\downarrow}_{m,k} \overset{\text{def}}{=} T_{m,k}\mathbf{1}_{E_{m,k}} \tag{1}$$

record the duration, until 1 and 2 coalesce, that $\mathcal{T}_m$ has k lineages remaining.

Finally, we will need the following important fact concerning the $T_{m,k}$ and $\xi_{m,k,}$, which follows directly from the generative algorithm given above.

Lemma 1. *For $k \le j \le m$, the random variables $T_{m,j}$ and $\xi_{m,j}$ are independent conditional on $E_{m,k}$.*

S1.2.1 Branch lengths

$\mathbb{E}[T^{\tau\downarrow}_{m,k}]$ can be recursively computed by the following theorem:

Theorem 1. $\mathbb{E}[T^{\tau\downarrow}_{m,k}]$ *satisfies the recursion*

$$\left[1 - \tfrac{(k+1)(k-2)}{(m+1)(m-2)}\right]\mathbb{E}T^{\tau\downarrow}_{m,k} = \mathbb{E}T^{\tau\downarrow}_{m-1,k} - \tfrac{(k+2)(k-1)}{(m+1)(m-2)}\mathbb{E}T^{\tau\downarrow}_{m,k+1}$$

with base case $\mathbb{E}T^{\tau\downarrow}_{k,k} = \int_0^\tau e^{-(a_k-1)R(t)}\,dt$.

Proof. Generate $\mathcal{T}_2, \dots, \mathcal{T}_m$ as above. Then

$$T^{\tau\downarrow}_{m-1,k} = T^{\tau\downarrow}_{m,k}\mathbf{1}\{\{m\} \notin \xi_{m,k}\} + T^{\tau\downarrow}_{m,k+1}\mathbf{1}\{\{m\} \in \xi_{m,k+1}\} \tag{2}$$

because if $\mathcal{T}_m$ has j lineages at time t, then $\mathcal{T}_{m-1}$ has $j - \mathbf{1}\{\{m\} \in \xi_{m,j}\}$ lineages at that time. Using (1), we can write the first term in (2) as

$$T^{\tau\downarrow}_{m,k}\mathbf{1}\{\{m\} \notin \xi_{m,k}\} = T_{m,k}\mathbf{1}_{E_{m,k}}\mathbf{1}\{\{m\} \notin \xi_{m,k}\} \tag{3}$$

Taking expectation in (3) and using Lemma 1, we obtain

$$\begin{aligned}\mathbb{E}(T^{\tau\downarrow}_{m,k}\mathbf{1}\{\{m\} \notin \xi_{m,k}\}) &= \mathbb{P}(E_{m,k})\mathbb{E}(T_{m,k} \mid E_{m,k})\mathbb{P}(\{m\} \notin \xi_{m,k} \mid E_{m,k}) \\ &= \mathbb{E}(T^{\tau\downarrow}_{m,k})\mathbb{P}(\{m\} \notin \xi_{m,k} \mid E_{m,k}) \end{aligned} \tag{4}$$

Analogous manipulations of the second term in (2) yield

$$\mathbb{E}T^{\tau\downarrow}_{m-1,k} = \mathbb{E}T^{\tau\downarrow}_{m,k}\mathbb{P}(\{m\} \notin \xi_{m,k} \mid E_{m,k}) + \mathbb{E}T^{\tau\downarrow}_{m,k+1}\mathbb{P}(\{m\} \notin \xi_{m,k+1} \mid E_{m,k+1}). \tag{5}$$

Now, we have

$$\begin{aligned}\mathbb{P}(\{m\} \in \xi_{m,k} \mid E_{m,k}) &= \left(1 - \frac{m-1}{a_m - 1}\right) \cdots \left(1 - \frac{k+1-1}{a_{k+1} - 1}\right) \\ &= \prod_{j=k+1}^{m} \frac{(a_j - 1) - (j-1)}{a_j - 1} \\ &= \frac{(k+1)(k-2)}{(m+1)(m-2)}. \end{aligned} \tag{6}$$

Combining (5) and (6) yields the first part of the claim. The base case follows from direct computation:

$$\begin{aligned}\mathbb{E}T^{\tau\downarrow}_{k,k} &= \tau\mathbb{P}(T_{k,k} > \tau) + \mathbb{E}T_{k,k}\mathbf{1}\{T_{k,k} < \tau\} \\ &= \tau\mathbb{P}(T_{k,k} > \tau) + \int_0^\tau t(a_k - 1)\alpha(t)e^{-(a_k-1)R(t)}\,dt \\ &= \int_0^\tau e^{-(a_k-1)R(t)}\,dt, \end{aligned}$$

where we integrated by parts in the final equality. □

S1.2.2 Block Structure

Next we compute the distribution of partition block structures conditional on 1 and 2 not having coalesced. Following the notation of the preceding subsection, let $\xi_{m,k} = \{L_1, \ldots, L_k\}$ be the coalescent partition of $[m] := \{1, \ldots, m\}$ when $\mathcal{T}_m$ has k lineages. As partitions are unordered, we are free to adopt a labeling, so let L_i contain the smallest numbered leaf which is not in $L_1 \cup \cdots \cup L_{i-1}$. Conditioned on $\{C_{12} = \tau\}$, this is equivalent to $1 \in L_1, 2 \in L_2$ beneath τ. The following results characterizes the block structure of $\mathcal{T}_m$ before 1 and 2 coalesce. In what follows, we use the conditioning notation $\mathbb{P}(\cdot \mid k)$ to signify that they pertain to the distribution on partitions in the coalescent while there are k blocks remaining.

Theorem 2. *The distribution of the block sizes* $|L_1|, \ldots, |L_k|$ *in* $\xi_{m,k}$, *conditional on* $\{1 \in L_1, 2 \in L_2, E_{m,k}\}$, *is given by*

$$\mathbb{P}(|L_1|, \ldots, |L_k| \mid 1 \in L_1, 2 \in L_2, k, E_{m,k}) = \frac{|L_1||L_2|}{\binom{m+1}{k+1}}.$$

Proof. Write

$$\mathbb{P}(|L_1|, \ldots, |L_k| \mid 1 \in L_1, 2 \in L_2, k, E_{m,k}) = \frac{\mathbb{P}(|L_1|, \ldots, |L_k| \mid k, E_{m,k})}{\mathbb{P}(1 \in L_1, 2 \in L_2 \mid k, E_{m,k})}\mathbf{1}\{1 \in L_1, 2 \in L_2\},$$

where in the denominator we are marginalizing over all partitions of $[m]$ into k blocks. A computation resembling equation (6) in the proof of Theorem 1 shows that

$$\mathbb{P}(1 \in L_1, 2 \in L_2 \mid k, E_{m,k}) = \prod_{i=k+1}^{m} \frac{a_i - 2}{a_i - 1} = \frac{(m+1)(k-1)}{(m-1)(k+1)}.$$

Now, the random variables $\xi_{m,k}$ are *not* in general distributed according to the coalescent due to the conditioning of $\mathcal{T}_k$ on $\{C_{12} = \tau\}$. Nevertheless, by the remarks preceding Lemma 1, beneath τ they are equal in distribution to the classical coalescent. Therefore, we may apply a result of Kingman (1982) to conclude that

$$\mathbb{P}(\xi_{m,k} = \{L_1, \dots, L_k\} \mid k, E_{m,k}) = \prod_{i=1}^{k} |L_i|! \times \frac{k!(k-1)!(m-k)!}{m!(m-1)!} \tag{7}$$

Let

$$\mathcal{S}_\xi = \{\zeta \,:\, \zeta_1 \uplus \cdots \uplus \zeta_k = [m], |\zeta_i| = |L_i| \,\forall i, 1 \in \zeta_1, 2 \in \zeta_2\} \tag{8}$$

be the set of partitions which have the same block sizes as $\xi_{m,k}$ and 1 and 2 in different blocks. There are

$$|\mathcal{S}_\xi| = \frac{1}{(k-2)!} \binom{m-2}{|L_1|-1, |L_2|-1, |L_3|, \dots, |L_k|} \tag{9}$$

such partitions, all of which have the probability given in equation (7). Hence,

$$\begin{aligned}
\mathbb{P}(|L_1|, \dots, |L_k| \mid 1 \in L_1, 2 \in L_2, k, E_{m,k}) &= \frac{1}{\mathbb{P}(1 \in L_1, 2 \in L_2 \mid k, E_{m,k})} \sum_{\zeta \in \mathcal{S}_\xi} \mathbb{P}(\zeta \mid k, E_{m,k}) \\
&= \frac{\mathbb{P}(\xi_{m,k} \mid k, E_{m,k})}{\mathbb{P}(1 \in L_1, 2 \in L_2 \mid k, E_{m,k})} |\mathcal{S}_\xi| \\
&= |L_1| \times |L_2| \times \frac{k-1}{m-1} \times \frac{1}{\binom{m}{k}} \times \frac{(n-1)(k+1)}{(m+1)(k-1)} \\
&= \frac{|L_1||L_2|}{\binom{m+1}{k+1}}.
\end{aligned}$$

□

Theorem 3. *Under the conditions of Theorem 2,*

$$\mathbb{P}(|L_i| = \ell_i \mid 1 \in L_1, 2 \in L_2, k+2, E_{m,k+2}) = \frac{\ell_i}{\binom{n+3}{k+3}} \binom{n+2-\ell_i}{k+1},$$

for $i = 1, 2$, while for $3 \le i \le k+2$,

$$\mathbb{P}(|L_i| = \ell_i \mid 1 \in L_1, 2 \in L_2, k+2, E_{m,k+2}) = \frac{1}{\binom{n+3}{k+3}} \binom{n+3-\ell_i}{k+2}.$$

Proof. For the first result, if $k = 0$ then only L_1 and L_2 remain, and the claim follows immediately. Otherwise, using Theorem 2,

$$\mathbb{P}(|L_1| = \ell_1 \mid 1 \in L_1, 2 \in L_2, k+2, E_{m,k+2}) = \frac{\ell_1}{\binom{n+3}{k+3}} \sum_{\ell_2=1}^{n+2-\ell_1-k} \ell_2 \binom{n-\ell_1-\ell_2+1}{k-1},$$

where the binomial term counts the number of k-way compositions of the integer $n+2-\ell_1-\ell_2$. The sum simplifies to

$$\sum_{\ell_2=1}^{n+2-\ell_1-k} \ell_2 \binom{n-\ell_1-\ell_2+1}{k-1} = \sum_{\ell_2=0}^{n-\ell_1} \binom{\ell_2+1}{1}\binom{n-\ell_1-\ell_2}{k-1} = \binom{n+2-\ell_1}{k+1},$$

where the second equality is from Graham et al. (1994, eq. 5.26). By symmetry,

$$\mathbb{P}(|L_2| = \ell \mid 1 \in L_1, 2 \in L_2, k+2, E_{m,k+2}) = \mathbb{P}(|L_1| = \ell \mid 1 \in L_1, 2 \in L_2, k+2, E_{m,k+2}). \tag{10}$$

For the second result, using the same combinatorial identity twice yields

$$\begin{aligned}\mathbb{P}(|L_i| = \ell_i \mid 1 \in L_1, 2 \in L_2, k+2, E_{m,k+2}) &= \frac{1}{\binom{n+3}{k+3}} \sum_{\ell_1=1}^{n+2-\ell_i-k} \ell_1 \sum_{\ell_2=1}^{n+3-\ell_i-\ell_1-k} \ell_2 \binom{n-\ell_1-\ell_2-\ell_i+1}{k-2} \\ &= \frac{1}{\binom{n+3}{k+3}} \sum_{\ell_1=1}^{n+2-\ell_i-k} \ell_1 \binom{n-\ell_1-\ell_i+2}{k} \\ &= \frac{1}{\binom{n+3}{k+3}} \binom{n+3-\ell_i}{k+2},\end{aligned}$$

for all $3 \le i \le k+2$. □

S1.2.3 Computing CSFS($\tau\downarrow$)

Using the preceding results we can compute CSFS($\tau\downarrow$). Let

$$\mathbf{T}^{\tau\downarrow} = \mathbb{E}(T^{\tau\downarrow}_{n+2,2}, \dots, T^{\tau\downarrow}_{n+2,n+2})$$

be the row-vector of expected intercoalescence times below τ and

$$\boldsymbol{\gamma}(\tau) = \mathbb{E}(T^{\tau\downarrow}_{2,2}, \dots, T^{\tau\downarrow}_{n+2,n+2})$$

be the row-vector of first-coalescence times, which can be computed using the base case in Theorem 1. By unwinding the recursion proved in that theorem, we obtain a matrix $\mathbf{B} \in \mathbb{Q}^{(n+1)\times(n+1)}$ such that $\mathbf{T}^{\tau\downarrow} = \boldsymbol{\gamma}(\tau)\mathbf{B}$. Additionally, if $\mathbf{D} = \mathrm{diag}(2, 3, \dots, n+2)$ then $\boldsymbol{\gamma}(\tau)\mathbf{BD}$ is the total expected branch length in each intercoalescence interval.

Next, let $\mathbf{P}_0 \in \mathbb{Q}^{(n+1)\times n}$ and $\mathbf{P}_1 \in \mathbb{Q}^{(n+1)\times(n+1)}$ be matrices whose entries are given by Theorem 3:

$$\begin{aligned}
(\mathbf{P}_0)_{i,j} &= \mathbb{P}(|L_3| = j \mid 1 \in L_1, 2 \in L_2, i+1, E_{n+2,i+1}), \quad 2 \le i \le n+1 \text{ and } 1 \le j \le n-i+1\\
(\mathbf{P}_1)_{i,j} &= \mathbb{P}(|L_1| = j \mid 1 \in L_1, 2 \in L_2, i+1, E_{n+2,i+1}), \quad 1 \le i \le n+1 \text{ and } 1 \le j \le n-i+1\\
(\mathbf{P}_0)_{i,j} &= (\mathbf{P}_1)_{i,j} = 0, \quad \text{elsewhere.}
\end{aligned}$$

The (i, j)th entry of $\mathbf{P}_0$ is the probability that a lineage in level $i+1$ (that is, when $i-1$ undistinguished lineages remain) subtends j undistinguished lineages at present, given that it does not contain 1 or 2. Similarly, the (i, j)th entry of $\mathbf{P}_1$ is the probability that a lineage in level $i+1$ subtends $j-1$ undistinguished lineages, in addition to 1 or 2 (but not both).

Finally, let $\mathbf{E} = 2\cdot\text{diag}(1/2, 1/3, \dots, 1/(n+2))$ be the diagonal matrix whose i-th entry is the probability that a randomly chosen lineage chosen at level $i+1$ contains 1 or 2. We then have that $\mathbf{E}\mathbf{P}_1 \in \mathbb{Q}^{(n+1)\times(n+1)}$ is the matrix whose (i, j)th entry is the probability that a lineage at level $i+1$ subtends a sample $(1, j-1)$. Similarly, $(\mathbf{I}-\mathbf{E})\mathbf{P}_0$ gives the probability of subtending $(0, j)$.

Now consider the $(0, i)$ entry of CSFS($\tau \downarrow$), which is the (conditional) probability of observing a derived allele in i undistinguished lineages, but in neither distinguished lineage. Write $\xi_{m,k} = \{\zeta^{(1)}_{m,k}, \dots, \zeta^{(k)}_{m,k}\}$, where the $\zeta^{(i)}_{m,k}$ are the blocks of $\xi_{m,k}$ randomly ordered. Under the SFS approximation,

$$\begin{aligned}
\frac{2}{\theta}\mathbb{P}((0, j) \mid C_{12} = \tau) &= \sum_{k=2}^{n+2} \mathbb{E}\Big(T^{\tau\downarrow}_{n+2,k} \sum_{\zeta\in\xi_{n+2,k}} \mathbf{1}\{|\zeta| = j, \{1, 2\} \cap \zeta = \emptyset\}\Big)\\
&= \sum_{k=2}^{n+2} k\mathbb{E}\Big(T^{\tau\downarrow}_{n+2,k}\mathbf{1}\{|\zeta^{(1)}_{n+2,k}| = j, \{1, 2\} \cap \zeta^{(1)}_{n+2,k} = \emptyset\}\Big)\\
&= \sum_{k=2}^{n+2} k\mathbb{E}T^{\tau\downarrow}_{n+2,k}\mathbb{P}(|\zeta^{(1)}_{n+2,k}| = j, \{1, 2\} \cap \zeta^{(1)}_{n+2,k} = \emptyset \mid E_{n+2,k})\\
&= \sum_{k=0}^{n} [\boldsymbol{\gamma}(\tau)\mathbf{B}\mathbf{D}]_k[(\mathbf{I}-\mathbf{E})\mathbf{P}_0]_{kj}.
\end{aligned}$$

Here, the third equality holds by the same argument as in equation (4) in the proof of Theorem 1.

We see that the last n columns of the first row of CSFS($\tau \downarrow$) are given by the following matrix-vector product:

$$[\text{CSFS}(\tau \downarrow)]_{0,1:n} = \boldsymbol{\gamma}(\tau)\mathbf{B}\mathbf{D}(\mathbf{I}-\mathbf{E})\mathbf{P}_0. \tag{11}$$

Using similar arguments, the second row can be shown to equal

$$[\text{CSFS}(\tau \downarrow)]_{1,0:n} = \boldsymbol{\gamma}(\tau)\mathbf{B}\mathbf{D}\mathbf{E}\mathbf{P}_1. \tag{12}$$

All other entries of CSFS($\tau \downarrow$) are zero.

S1.2.4 Interval Calculation

The above formulas can be used to compute CSFS($\tau \downarrow$) for a fixed coalescence time τ. Since the hidden states of our model consist of intervals for τ, we must integrate these expressions with respect to the coalescence density of 1 and 2. This is easily accomplished since the integral commutes with the linear transforms expressed above. Specifically, for an interval $[t_1, t_2)$ we integrate the vector of first coalescence times $\boldsymbol{\gamma}(\tau)$ against the conditional coalescence density

$$f_{[t_1,t_2)}(t) = \frac{\alpha(t)e^{-R(t)}}{e^{-R(t_1)} - e^{-R(t_2)}} \tag{13}$$

to obtain $\boldsymbol{\gamma}' = \int_{t_1}^{t_2} f_{[t_1,t_2)}(\tau)\boldsymbol{\gamma}(\tau)\,d\tau$, which can then be used in (11) and (12).

S1.3 Computing CSFS($\tau \uparrow$)

To calculate the CSFS above τ, we adapt the approach of Kamm et al. (2016) for computing the multi-population coalescent. The reader is referred to that paper for motivating details on the method; briefly, the strategy is to compute the unconditioned SFS on $n+1$ lineages beginning at time $\tau+$, and multiply it by the probability that a lineage which has size $k = 1, \dots, n$ at time $\tau+$ has size (a, b) at time 0. The latter quantity is calculated as the transition probability of a certain forward-time Moran model on $n+1$ lineages which is dual to the conditioned coalescent.

Formally, we have

$$[\mathrm{CSFS}(\tau \uparrow)]_{ab} = \sum_{i=1}^{n} [\boldsymbol{v}(\tau)]_i \times \mathbb{P}(i \to (a,b); \tau) \tag{14}$$

where $[\boldsymbol{v}(\tau)]_i$ denotes the total branch length of lineages above time τ that have size $i \in \{1, \dots, n\}$ at time $\tau+$ (i.e. the unnormalized SFS on $n+1$ lineages), and $\mathbb{P}(\cdot; \tau)$ is the aforementioned transition function.

To compute $\boldsymbol{v}(\tau)$, let $\boldsymbol{\varepsilon}(\tau) \in \mathbb{R}^n$ be a row-vector whose ith entry is the expected time to first coalescence in a sample of size $i+1$ beginning at time τ:

$$[\boldsymbol{\varepsilon}(\tau)]_i = \int_{\tau}^{\infty} \alpha(t)e^{-a_{i+1}[R(t)-R(\tau)]}\,dt.$$

Polanski and Kimmel (2003) have derived a matrix $\mathbf{W} \in \mathbb{Q}^{n \times n}$ such that the unnormalized SFS on a sample of size $n+1$ (i.e., the total expected branch length having sizes $1, \dots, n$) equals $\boldsymbol{v}(\tau) = \mathbf{W}\boldsymbol{\varepsilon}(\tau)$.

S1.3.1 Transition probability

Next we calculate the probability that k lineages at time τ have (a, b) descendants at present (time 0). Note that a lineage above τ must subtend either $a = 0$ or $a = 2$ distinguished lineages at present. For $a \in \{0, 2\}$, let $M_a \in \mathbb{Q}^{(n+1)\times(n+1)}$ be the tridiagonal matrix

$$(\mathbf{M}_a)_{i+1,j+1} = \begin{cases} (2-a)i + \frac{1}{2}i(n-i), & j = i-1, \\ a(n-i) + \frac{1}{2}i(n-i), & j = i+1, \\ 0, & \text{otherwise.} \end{cases}$$

and let $\mathbf{Q}_a$ be the rate matrix of a continuous time Markov chain whose embedded jump chain has transition matrix $\mathbf{M}_a$, i.e., its diagonals are the negative row sums of $\mathbf{M}_a$.

The processes defined by $\mathbf{Q}_a$ are Moran models whose state is the number of undistinguished lineages bearing a derived allele as time runs towards the present. The integer a is the number of derived alleles in the two distinguished lineages. The two distinguished lineages copy onto each undistinguished lineage in the opposite allelic class at rate 1; additionally the undistinguished lineages copy between classes at rate $1/2$ as in the usual Moran model. Copying onto either distinguished lineage is disallowed.

One can verify that the coalescent trees embedded in this Moran model are equal in distribution to those of the conditioned coalescent below τ. Accordingly, the transition density function in equation (14) is precisely the matrix exponential

$$\mathbf{M}_a(\tau) := e^{R(\tau)\mathbf{Q}_a}. \tag{15}$$

S1.3.2 Computing CSFS($\tau \uparrow$)

Let $\mathbf{S}_2 := \mathrm{diag}(1, \ldots, n)/(n+1) \in \mathbb{Q}^{n\times n}$ be the matrix whose ith diagonal entry is the probability that i lineages randomly chosen from $n+1$ lineages at $\tau+$ subtend both of the distinguished lineages, and set $\mathbf{S}_0 = \mathbf{I} - \mathbf{S}_2$. If $a = 0$, the probability that a lineage which has size i at time τ has size (a, b) at time 0 is $(\mathbf{S}_0)_{ii} \times [\mathbf{M}_0(\tau)]_{i+1,b+1}$. If $a = 2$, then this probability is $(\mathbf{S}_2)_{ii} \times [\mathbf{M}_2(\tau)]_{i,b+1}$, because $i-1$ mutated undistinguished lineages remain after one of the i blocks "splits" into the two distinguished lineages at time τ. Accordingly, define the submatrices $\mathbf{M}_0'(\tau) = [\mathbf{M}_0(\tau)]_{[2:(n+1),2:(n+1)]}$ and $\mathbf{M}_2'(\tau) = [\mathbf{M}_2(\tau)]_{[1:n,1:n]}$. By (14) and (15) we have

$$\sum_{i=1}^{n} [\boldsymbol{\nu}(\tau)]_i \mathbb{P}(i \to (a,b); \tau) = \sum_{i=1}^{n} (\boldsymbol{\varepsilon}(\tau)^T \mathbf{W}^T)_i (\mathbf{S}_a)_{ii} [\mathbf{M}_a'(\tau)]_{i,b} = [\boldsymbol{\varepsilon}(\tau)^T \mathbf{W}^T \mathbf{S}_a \mathbf{M}_a'(\tau)]_b.$$

S1.3.3 Integrating CSFS($\tau \uparrow$)

To integrate CSFS($\tau \uparrow$) against the conditional coalescence density (13), let $\mathbf{U}_a \cdot e^{R(\tau)\mathbf{D}} \cdot (\mathbf{U}_a^{-1}) = \mathbf{M}_a(\tau)$ be the eigendecomposition of $\mathbf{M}_a$, where matrix $\mathbf{D} = \mathrm{diag}(D_1, \ldots, D_{n+1})$ and $\mathbf{U}_a \in \mathbb{Q}^{(n+1)\times(n+1)}$ can be computed exactly using results from the next section. Let $\mathbf{U}_a'$ and $(\mathbf{U}_a^{-1})'$ denote appropriate submatrices of $\mathbf{U}_a$ and $(\mathbf{U}_a^{-1})$, respectively, that appear in the decomposition of $\mathbf{M}_a'(\tau)$. Defining $\mathbf{X}_a := \mathbf{W}^T \mathbf{S}_a \mathbf{U}_a' \in \mathbb{Q}^{n\times(n+1)}$, we have

$$\int_{t_1}^{t_2} f_{[t_1,t_2)}(\tau) \boldsymbol{\varepsilon}(\tau)^T \mathbf{W}^T \mathbf{S}_a \mathbf{M}_a'(\tau)\, d\tau = \int_{t_1}^{t_2} f_{[t_1,t_2)}(\tau) \boldsymbol{\varepsilon}(\tau)^T \mathbf{X}_a e^{R(\tau)\mathbf{D}}\, d\tau \cdot (\mathbf{U}_a^{-1})'.$$

The integral simplifies to

$$
\begin{aligned}
&\int_{t_1}^{t_2} f_{[t_1,t_2)}(\tau)\boldsymbol{\varepsilon}(\tau)^T \mathbf{X}_a e^{R(\tau)\mathbf{D}}\, d\tau \\
&\qquad = \int_{t_1}^{t_2} f_{[t_1,t_2)}(\tau)\left[\boldsymbol{\varepsilon}(\tau)^T(\mathbf{X}_a)_1 e^{-D_1 R(\tau)}, \dots, \boldsymbol{\varepsilon}(\tau)^T(\mathbf{X}_a)_{n+1} e^{-D_{n+1}R(\tau)}\right]^T d\tau = \mathbf{1}^T(\mathbf{X}_a \circ \mathbf{C}^T),
\end{aligned}
$$

where $(\mathbf{X}_a)_j$ denotes the jth column of $\mathbf{X}_a$, $\mathbf{1}$ is a vector of 1s, $\circ$ denotes the Hadamard product, and matrix $\mathbf{C} \in \mathbb{R}^{(n+1)\times n}$ is defined by

$$
\begin{aligned}
(\mathbf{C})_{ij} &= \frac{1}{e^{-R(t_1)} - e^{-R(t_2)}} \int_{t_1}^{t_2} \alpha(\tau) e^{-(D_i+1)R(\tau)}[\boldsymbol{\varepsilon}(\tau)]_j \, d\tau \\
&= \frac{1}{e^{-R(t_1)} - e^{-R(t_2)}} \int_{t_1}^{t_2} \alpha(\tau) e^{-[D_i+1-a_{j+1}]R(\tau)} \int_{\tau}^{\infty} e^{-a_{j+1}R(t)}\, dt\, d\tau.
\end{aligned}
$$

S1.3.4 Spectral Decomposition of $\mathbf{Q}_a$

We utilize an exact spectral decomposition of $\mathbf{Q}_a$ to perform computations.

Lemma 4. *The eigenvalues of* $\mathbf{Q}_a$ *are* $\{1 - a_m : m = 2, 3, \dots, n+2\}$.

Proof. Let $\mathbf{T}_a = \mathbf{I} + (a_{n+2} - 1)^{-1}\mathbf{Q}_a$ be the discrete time version of $\mathbf{Q}_a$. The eigenvalues of $\mathbf{T}_a$ are known to be $\{G_{ii} : i = 0, \dots, n\}$ where G_{ii} is the probability that i randomly sampled individuals at generation t have distinct parents in the generation $t-1$ (Gladstien, 1978). (Here, the probability measure is with respect to the Markov chain corresponding to $\mathbf{T}_a$.) Let $p_{ni} = \binom{n}{i-2}/\binom{n+2}{i}$ be the probability that the random sample contains both distinguished lineages. We have

$$
\begin{aligned}
G_{ii} &= p_{ni}\left(1 - \frac{a_i - 1}{a_{n+2} - 1}\right) + (1 - p_{ni})\left(1 - \frac{a_i}{a_{n+2} - 1}\right) \\
&= \frac{(n+i+1)(n+2-i)}{n(n+3)}
\end{aligned}
$$

after much simplification. The result now follows from basic spectral theory. □

Using the lemma we obtain the left and right eigenvectors $\mathbf{U}_a, \mathbf{U}_a^{-1} \in \mathbb{Q}^{(n+1)\times(n+1)}$ of $\mathbf{Q}_a$ by solving tridiagonal systems for each eigenvalue.

S2 Locus-Skipping EM Algorithm

In this section we outline the procedure, described in Section , for efficiently computing the posterior expected number of emissions and transitions in long stretches of nonpolymorphic sites. We presume some familiarity with the forward-backward and Baum-Welch algorithms; a good introduction as well as the source of most of the notation in this section may be found in Bishop (2006, §13.2). In particular, the notation in this section is separate from the notation of the rest of the paper (e.g. α, ξ are overloaded to have a different meaning here).

To begin, let $\mathbf{X} = (x_1, \dots, x_L)^T$ be the vector of observed data and let $\mathbf{Z} = (z_1, \dots, z_L)^T$ be the (unobserved) vector of hidden states at each location in the HMM. We suppose that between observations x_k and x_ℓ there is a long stretch of identical monomorphic observations $x_{k+1} = \cdots = x_{\ell-1}$. Let T be the transition matrix and B be a diagonal matrix with entries $B_{ii} = \mathbb{P}(x_{k+1} \mid z_{k+1} = i)$. Define W to be their product $W = TB$.

The forward probabilities at $\ell - 1$, denoted by the vector $\alpha(z_{\ell-1})$, are then

$$\begin{aligned}\alpha(z_{\ell-1}) &= \mathbb{P}(z_{\ell-1}, x_1, x_2, \dots, x_{\ell-1}) \\ &= (W^T)^{\ell-k-1}\alpha(z_k)\end{aligned}$$

which can be computed in a single step from $\alpha(z_k)$, skipping over the positions $\{k+1, k+2, \dots, \ell-2\}$. The locus-skipping transition matrix $(W^T)^{\ell-k-1}$ is efficiently obtained by performing an eigendecomposition of W; see below.

Similarly, the backward probabilities at k, denoted by $\beta(z_k)$, can be computed in a single step from $\beta(z_{\ell-1})$:

$$\begin{aligned}\beta(z_k) &= \mathbb{P}(x_{k+1}, x_{k+2}, \dots, x_L \mid z_k) \\ &= W^{\ell-k-1}\beta(z_{\ell-1}).\end{aligned}$$

To prevent underflow one must rescale the forward and backward probabilites (Bishop, 2006, §13.2.4). To do so, we define $\hat{\alpha}_k \overset{\text{def}}{=} c_k^{-1}\alpha_k$ and $\hat{\beta}_k \overset{\text{def}}{=} \big(\prod_{m=k+1}^{L} c_m\big)^{-1}\beta_k$, where $c_k \overset{\text{def}}{=} \mathbb{P}(x_k \mid x_1, \dots, x_{k-1})$. The $\hat{\alpha}_k$ and $\hat{\beta}_k$ recursions are identical to the unrescaled case, and c_k can be easily computed as the ℓ_1-norm of $W^T\hat{\alpha}_{k-1}$. Note also that $\mathbb{P}(\mathbf{X}) = \prod_{k=1}^{L} c_k$.

The EM algorithm also requires the Baum-Welch quantities $\xi(z_{j-1}, z_j) = \mathbb{P}(z_{j-1}, z_j \mid \mathbf{X})$ and $\gamma(z_j) = \mathbb{P}(z_j \mid \mathbf{X})$ to be summed over all j, including the "skipped" loci j with $k < j < \ell$. To sum these quantities at the skipped loci, we first note

$$\begin{aligned}\hat{\alpha}(z_j) &= \Big(\prod_{i=1}^{j} c_i\Big)^{-1}\alpha(z_j) = \Big(\prod_{i=1}^{j} c_i\Big)^{-1}(W^T)^{j-k}\alpha(z_k) = \Big(\prod_{i=k+1}^{j} c_i\Big)^{-1}(W^T)^{j-k}\hat{\alpha}(z_k) \\ \hat{\beta}(z_j) &= \Big(\prod_{i=j+1}^{L} c_i\Big)^{-1}\beta(z_j) = \Big(\prod_{i=j+1}^{L} c_i\Big)^{-1}W^{\ell-j-1}\beta(z_{\ell-1}) = \Big(\prod_{i=j+1}^{\ell-1} c_i\Big)^{-1}W^{k}\hat{\beta}(z_{\ell-1})\end{aligned}$$

so that

$$\gamma(z_j) = \mathrm{diag}\big(\hat{\alpha}(z_j)\hat{\beta}(z_j))^T\big) = \Big(\prod_{i=k+1}^{\ell-1} c_i\Big)^{-1}\mathrm{diag}\Big((W^T)^{j-k}\hat{\alpha}(z_k)\hat{\beta}(z_{\ell-1})^T(W^T)^{\ell-j-1}\Big) \tag{16}$$

Now let

$$PDP^{-1} = TB \tag{17}$$

be the eigendecomposition of W.[1] The expected number of positions in the interval (k, l) spent at each hidden state is obtained by summing (16):

$$\begin{aligned}\sum_{j=k+1}^{\ell-1} \gamma(z_j) &= \operatorname{diag}\Big[P\Big(\sum_{j=k+1}^{\ell-1} D^{j-k} P^{-1} \hat{\alpha}(z_k) \hat{\beta}(z_{\ell-1})^T P D^{\ell-j-1}\Big) P^{-1}\Big] \\ &= \operatorname{diag}\big(P A P^{-1}\big) \qquad (18)\end{aligned}$$

for $A \overset{\text{def}}{=} \sum_{j=k+1}^{\ell-1} D^{j-k} P^{-1} \hat{\alpha}(z_k) \hat{\beta}(z_{\ell-1})^T P D^{\ell-j-1}$.

It is not difficult to show that

$$\Big(\sum_{i=0}^{m} D^i U D^{m-i}\Big)_{ab} = U_{ab} \frac{D_{aa}^{m+1} - D_{bb}^{m+1}}{D_{aa} - D_{bb}} \overset{\text{def}}{=} U_{ab} Q_{ab} \qquad (19)$$

for diagonal matrix D and arbitrary matrix U. Using (19) with $m = \ell - k - 2$ and setting

$$U \overset{\text{def}}{=} P^{-1} \hat{\alpha}(z_k) \hat{\beta}(z_{\ell-1})^T P$$

we obtain $A = D(U \circ Q)$, where "$\circ$" denotes Hadamard (entrywise) product.

To obtain $\sum_{j=k+1}^{\ell-1} \xi(z_{j-1}, z_j)$ we simply note that

$$\xi(z_{j-1}, z_j) = \mathbb{P}(z_{j-1}, z_j, \mathbf{X}) = (\hat{\alpha}(z_{j-1}) \hat{\beta}(z_j)^T B) \circ T$$

so that

$$\begin{aligned}\sum_{j=k+1}^{\ell-1} \xi(s_{j-1}, s_j) &= \Big[P\Big(\sum_{j=k+1}^{\ell-1} D^{j-k-1} P^{-1} \hat{\alpha}(z_k) \hat{\beta}(z_{\ell-1})^T P D^{\ell-j-1}\Big) P^{-1} B\Big] \circ T \\ &= [P(U \circ Q) P^{-1} B] \circ T. \qquad (20)\end{aligned}$$

S2.1 Complexity

Assuming a total of M hidden states, we see that the computational cost of evaluating equations (18) and (20) is that of an $M \times M$ matrix-matrix multiply which is at most $O(M^3)$. (At each segregating site we employ the standard algorithms, which have lower cost $O(M^2)$.) The spectral decomposition (17) and formation of the auxiliary matrices U and Q all have strictly lower complexity. If we assume that between each of L_p polymorphic sites there is a large stretch of nonpolymorphic sites then the total cost of the algorithm is $O(L_p M^3)$.

[1] Note that in general P and D will not be real matrices, as W is not symmetric. This may seem troublesome since arithmetic over the complex field is a notorious source of inaccuracy in numerical analysis. However, for non-segregating sites we will have $B = I + E_1$ and similarly $T = I + E_2$ for some real-valued perturbation matrices E_i with $\|E_i\| \ll 1$. It follows by the Gershgorin circle theorem that $D = \Re(D) + i\Im(D)$ will have $\|\Im(D)\| \ll 1$, so that D is "almost" real. In practice we have encountered no problems by truncating the imaginary part of D.

S3 Extension to multiple populations

In this section we describe how the above conditioned SFS may be extended to handle multiple populations. We focus on the specific case of a "clean split" model in which two subpopulations are descended from a common ancestral population, with no gene flow occurring more recently than t_S generations ago. This section draws heavily on the ideas used by the program `momi` (Kamm et al., 2016) to calculate the *unconditioned* joint frequency spectrum.

The "clean split" model we analyze is as follows. Population 1 has instantaneous rate of coalescence $\alpha_1(t)$ and cumulative coalescence function $R_1(t) \overset{\text{def}}{=} \int_0^t \alpha_1(s)\,ds$. Going forward in time, at time t_S population 2 splits off from population 1 and follows its own size history with coalescence rate function $\alpha_2(t), 0 \leq t < t_S$. Hence the cumulative coalescence rate function for population 2 is

$$R_2(t) = \begin{cases} \int_0^t \alpha_2(s)\,ds, & t \leq t_S \\ \int_0^{t_S} \alpha_2(s)\,ds + \int_{t_S}^t \alpha_1(s)\,ds, & t > t_S. \end{cases}$$

Analogously to the CSFS described above, the joint conditioned SFS (JCSFS) is a 4-dimensional tensor which gives describes the sampling distribution of segregating sites in $n_1 + n_2 + 2$ sampled lineages, where n_i samples are obtained from population i, conditioned on the coalescence time of a pair of distinguished lineages. The distinguished haploid lineages can either arise in the same subpopulation (the "together" case), in which case phased data is not required, or one can be sampled from each subpopulation (the "apart" case).

For the rest of this section we make use of the following notation. The joint CSFS is denoted

$$\text{JCSFS}(\tau, t_S) \in \mathbb{R}^{(a_1+1)\times(n_1+1)\times(a_2+1)\times(n_2+1)}.$$

Here, τ denotes the conditioned coalescence time of the distinguished pair, t_S is the split time, and $n_i(a_i)$ is the number of (un)distinguished lineages sampled in population i. We require that $a_1 + a_2 = 2$; the "together" case corresponds to $a_1 = 2, a_2 = 0$ and the "apart" case, $a_1 = a_2 = 1$. For brevity, we drop the explicit dependence on τ and t_S from now on. The entry JCSFS$[i, j, k, l]$ therefore gives the probability of observing $i + k$ derived alleles in the distinguished lineages, j derived alleles in subpopulation 1, and l derived alleles in subpopulation 2.

The following quantities will be useful below. The *truncated (C)SFS* is defined to be the total expected branch length occurring beneath the split in each subpopulation. It is obtained mathematically by sending the population size above the split to zero (see (Kamm et al., 2016)). For n lineages we denote this as (C)SFS$_n(\overline{t_S})$. The *ancestral (J)CSFS* is defined to be expected site frequency spectrum of the ancestral population sampled immediately before the split. It is equivalent to computing the standard (J)CSFS using the cumulative rate function $R'(t) \overset{\text{def}}{=} R(t + t_S)$.

S3.1 Case 1: Together

In this case we write $J[i, j, k]$ for the probability of observing $0 \leq i \leq 2$ derived alleles in the distinguished pair, $0 \leq j \leq n_1$ in population 1, and $0 \leq k \leq n_2$ in population 2.

We distinguish two cases, depending on whether the distinguished lineages coalesce before or after the split. In either case, the total expected branch length beneath the split subtending population 2 (and no lineages in population 1) equals the SFS on n_2 lineages truncated to time t_S:

$$\mathrm{JCSFS}[0, 0, k] = \mathrm{SFS}(\overline{t_S})[k].$$

S3.1.1 $\tau < t_S$

If the distinguished lineages coalesce before the split, then the total expected branch length subtending population 1(2) and no samples from the other population is simply the JCSFS(τ) (SFS) truncated at time t_S. (For the additional entries $J[2, n_1, 0]$ and $J[0, 0, n_2]$ see Kamm et al., 2016, Lemma 1).

For the branch length above τ subtending we follow the same approach as in Section S1.3, by computing the ancestral JCSFS and using Moran models to transition forwards in time. That is, the total branch length subtending (i, b_1, b_2) at the present is equal to

$$\mathrm{JCSFS}[i, b_1, b_2] = \sum_{s=1}^{n_1+n_2} \sum_{t=(s-n_2)\vee 0}^{s\wedge(n_1+1)} h(t, n_1 + n_2 + 1, s, n_1 + 1) \cdot \mathcal{A}(s) \cdot \mathbb{P}_{1,i}(s \to b_1) \cdot \mathbb{P}_2(s - t \to b_2).$$

Here, $h(k, M, n, N) = \binom{n}{k}\binom{M-n}{N-k}/\binom{M}{N}$ is the hypegeometric density, and $\mathbb{P}_\alpha(i \to j)$ are transition functions. For $\alpha = 2$, we have

$$\mathbb{P}_2(i \to j) = (e^{t_S \mathbf{M}_{n_2}})_{ij}$$

where $\mathbf{M}_k \in \mathbb{R}^{(k+1)\times(k+1)}$ is the standard Moran rate matrix on k lineages. For $\alpha = (1, 0)$ and $\alpha = (1, 2)$ the process is inhomogeneous since the dynamics change at time τ. Hence,

$$\mathbb{P}_{(1,0)}(i \to j) = \left(e^{(t_S-\tau)\mathbf{M}_{n_1+1}} e^{\tau \mathbf{M}_{(0,n_1)}}\right)_{ij}$$

where $\mathbf{M}_{(0,n_1)}$ is the modified rate matrix described in Section S1.3.1, and similarly for $\alpha = (1, 2)$.

S3.1.2 $\tau \geq t_S$

The total expected branch length in population 1 beneath the split is equal to CSFS($t_S \downarrow$) (see Section S1.2). For the branch length above t_S subtending both populations, we compute ancestral CSFS(τ) on $n_1 + n_2$ lineages (denoted $\mathcal{A}$) and use Moran models to transition to the present:

$$\mathrm{JCSFS}[i, b_1, b_2] = \sum_{s=1}^{n_1+n_2} \sum_{t=(s-n_2)\vee 0}^{s\wedge(n_1+1)} h(t, n_1 + n_2, s, n_1) \cdot \mathcal{A}(i, s) \cdot \mathbb{P}_{1,i}(s \to b_1) \cdot \mathbb{P}_2(s - t \to b_2).$$

In this case $\mathbb{P}_2(i \to j)$ is the same as in the preceding section, and $\mathbb{P}_{(1,j)}(i \to j) = (e^{t_S \mathbf{M}_{(j,n_1)}})_{ij}$ where $\mathbf{M}_{(j,n_1)}$ is defined as in Section S1.3.1.

S3.2 Case 2: Apart

In this case JCSFS is a 4-dimensional tensor which we will index as JCSFS$[i, j, k, \ell]$. One distinguished lineage is sampled from each population, and no coalescences can occur before t_S. To compute the branch length above t_S subtending the present, we again sample the ancestral CSFS $\mathcal{A}$ at time t_S. Let $\gamma \overset{\text{def}}{=} h(t, n_1 + n_2, s, n_1)$ and define the matrix exponentials $T_{ij} \overset{\text{def}}{=} e^{\mathbf{M}_{(j,n_i)}}$ as in the preceding section. We distinguish four subcases.

1. For $i = k = 1$ we have

$$\text{JCSFS}[1, b_1, 1, b_2] = \sum_{s=1}^{n_1+n_2} \sum_{t=(s-n_2)\vee 0}^{s\wedge(n_1+1)} \gamma \cdot \mathcal{A}(2, s) \cdot T_{11}(t, b_1) \cdot T_{21}(s - t, b_2)$$

2. For $i = 1, k = 0$ we have

$$\text{JCSFS}[1, b_1, 0, b_2] = \sum_{s=1}^{n_1+n_2} \sum_{t=(s-n_2)\vee 0}^{s\wedge(n_1+1)} \frac{1}{2} \cdot \gamma \cdot \mathcal{A}(1, s) \cdot T_{11}(t, b_1) \cdot T_{20}(s - t, b_2).$$

   The case $i = 0, k = 1$ in analogous.

3. For $i = k = 0$ we have

$$\text{JCSFS}[0, b_1, 0, b_2] = \sum_{s=1}^{n_1+n_2} \sum_{t=(s-n_2)\vee 0}^{s\wedge(n_1+1)} \gamma \cdot \mathcal{A}(0, s) \cdot T_{10}(t, b_1) \cdot T_{20}(s - t, b_2)$$

   Finally, the branch length subtending population i beneath t_S is given by the truncated SFS on $n_i + 1$ lineages. For $1 \leq j \leq n_i$ the total expected branch length subtending 0 distinguished and j undistinguished lineages beneath t_S equals

$$\text{SFS}_{n_i+1}(\overline{t_S}) \frac{j}{n_i + 1},$$

   while the expected branch length subtending 1 distinguished and $j - 1$ undistinguished is

$$\text{SFS}_{n_i+1}(\overline{t_S}) \frac{n_i + 1 - j}{n_i + 1}.$$

References

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 0387310738.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L. F., Aximu-Petri, A., Prufer, K., de Filippo, C., Meyer, M., Zwyns, N., Salazar-Garcia, D. C., Kuzmin, Y. V., Keates, S. G., Kosintsev, P. A., Razhev, D. I., Richards, M. P., Peristov, N. V., Lachmann, M., Douka, K., Higham, T. F. G., Slatkin, M., Hublin, J.-J., Reich, D., Kelso, J., Viola, T. B., and Paabo, S. 10 2014. Genome sequence of a 45,000-year-old modern human from western siberia. *Nature*, **514,**(7523) 445–449. URL `http://dx.doi.org/10.1038/nature13810`.

Gladstien, K. 1978. The characteristic values and vectors for a class of stochastic matrices arising in genetics. *SIAM Journal on Applied Mathematics*, **34,**(4) 630–642.

Graham, R. L., Knuth, D. E., and Patashnik, O. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, Mass., 2nd ed edition, 1994. ISBN 0201558025.

Kamm, J. A., Terhorst, J., and Song, Y. S. 2016. Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics*.

Kingman, J. F. C. 1982. The coalescent. *Stoch. Process. Appl.*, **13,** 235–248.

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., and Pool, J. E. 2015. The drosophila genome nexus: A population genomic resource of 623 drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics*, **199,** (4) 1229–1241. ISSN 0016-6731. doi: 10.1534/genetics.115.174664. URL `http://www.genetics.org/content/199/4/1229`.

Polanski, A. and Kimmel, M. Sep 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, **165,**(1) 427–436.

Singhal, S., Leffler, E. M., Sannareddy, K., Turner, I., Venn, O., Hooper, D. M., Strand, A. I., Li, Q., Raney, B., Balakrishnan, C. N., Griffith, S. C., McVean, G., and Przeworski, M. 2015. Stable recombination hotspots in birds. *Science*, **350,**(6263) 928–932. ISSN 0036-8075. doi: 10.1126/science.aad0843. URL `http://science.sciencemag.org/content/350/6263/928`.

Supplementary Tables

Supplementary Table 1: Effect of data quality and sample size on the accuracy of posterior decoding a 10Mb region. The first three columns indicate the ratio of recombination to mutation, error rate, and fraction of missing data, respectively. The table entries are the ratio of posterior mean-squared error at the indicated sample size compared to sample size $n = 2$, averaged over ten simulations. (Hence, lower is better.) Entries marked with a "*" are significantly different from 1.0 at the 5% level using a two-sided t-test.

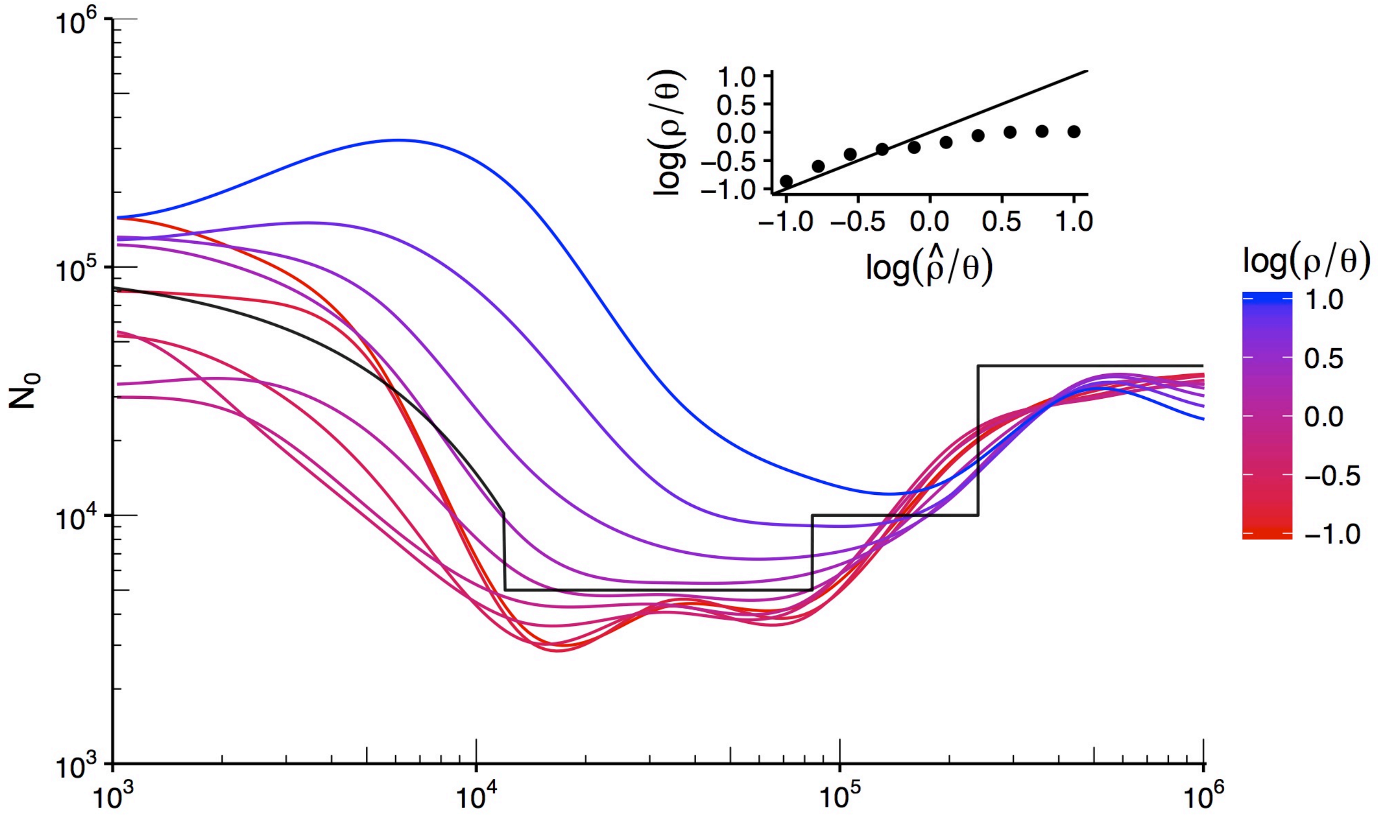
				n	
ρ/μ	$-\log_{10}(\mathrm{err})$	Miss. %	**5**	**10**	**25**
10.0	3	0	0.96*	0.95*	0.95*
		10	0.99*	0.99*	0.99*
		20	0.98*	0.98*	0.98*
	4	0	0.99*	0.99*	0.99*
		10	0.99*	0.99*	0.99*
		20	0.98*	0.98*	0.98*
	5	0	0.99*	0.99*	0.99*
		10	0.99*	0.99*	0.98*
		20	0.98*	0.98*	0.98*
1.0	3	0	0.82*	0.74*	0.68*
		10	0.96	0.93	0.94
		20	0.95*	0.94*	0.94*
	4	0	0.98*	0.97*	0.97*
		10	0.98*	0.98*	0.98*
		20	0.96*	0.95*	0.95*
	5	0	0.92	0.91	0.92
		10	0.97*	0.97*	0.97*
		20	0.95*	0.94*	0.94*
0.1	3	0	0.66*	0.55*	0.44*
		10	1.10	0.82	0.82
		20	0.86*	0.83*	0.81*
	4	0	0.92*	0.91*	0.88*
		10	0.99	0.98	0.99
		20	0.92*	0.85*	0.84
	5	0	1.04	0.99	0.94
		10	0.91	0.92	0.94*
		20	0.88*	0.85*	0.84*

Supplementary Table 2: Description of modern human data sets analyzed.

		Sample Size (n)	
Population	Description	Complete Genomics	1000 Genomes
CEU	Utah residents with European ancestry	12	112
CHB	Han Chinese in Beijing, China	8	198
GIH	Gujarati Indian in Houston, Texas, USA	8	198
JPT	Japanese in Tokyo, Japan	8	200
LWK	Luhya in Webuye, Kenya	8	190
MKK	Maasai in Kinyawa, Kenya	8	0
TSI	Toscans in Italy	10	212
YRI	Yoruba in Ibadan, Nigeria	14	202

Supplementary Table 3: Description of other data sets analyzed.

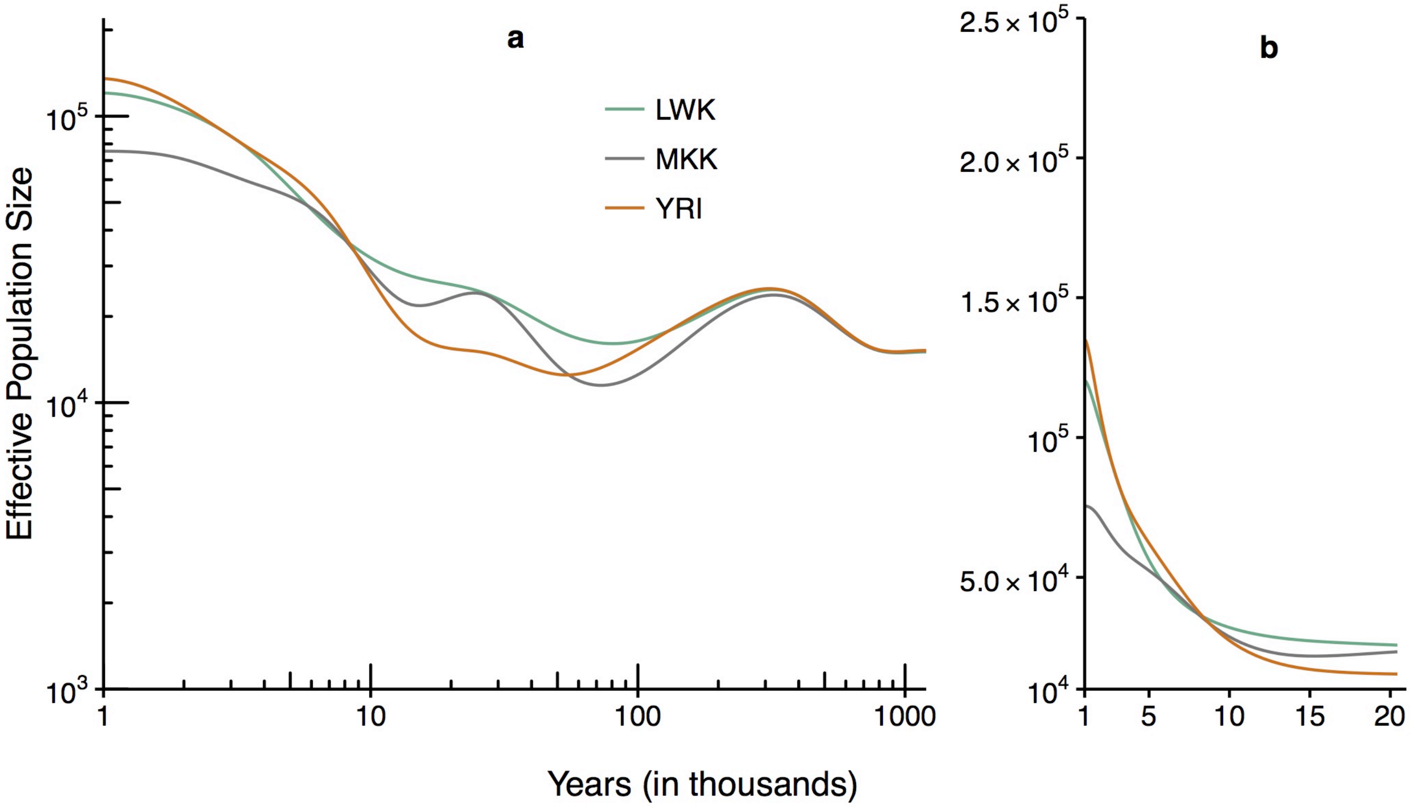
Organism	Sample Size (n)	Source
Long-tailed finch	40	(Singhal et al., 2015)
Zebra finch	40	*ibid.*
D. melanogaster	197	(Lack et al., 2015)
Ust'-Ishim (human)	2	(Fu et al., 2014)



Supplementary Figure 1

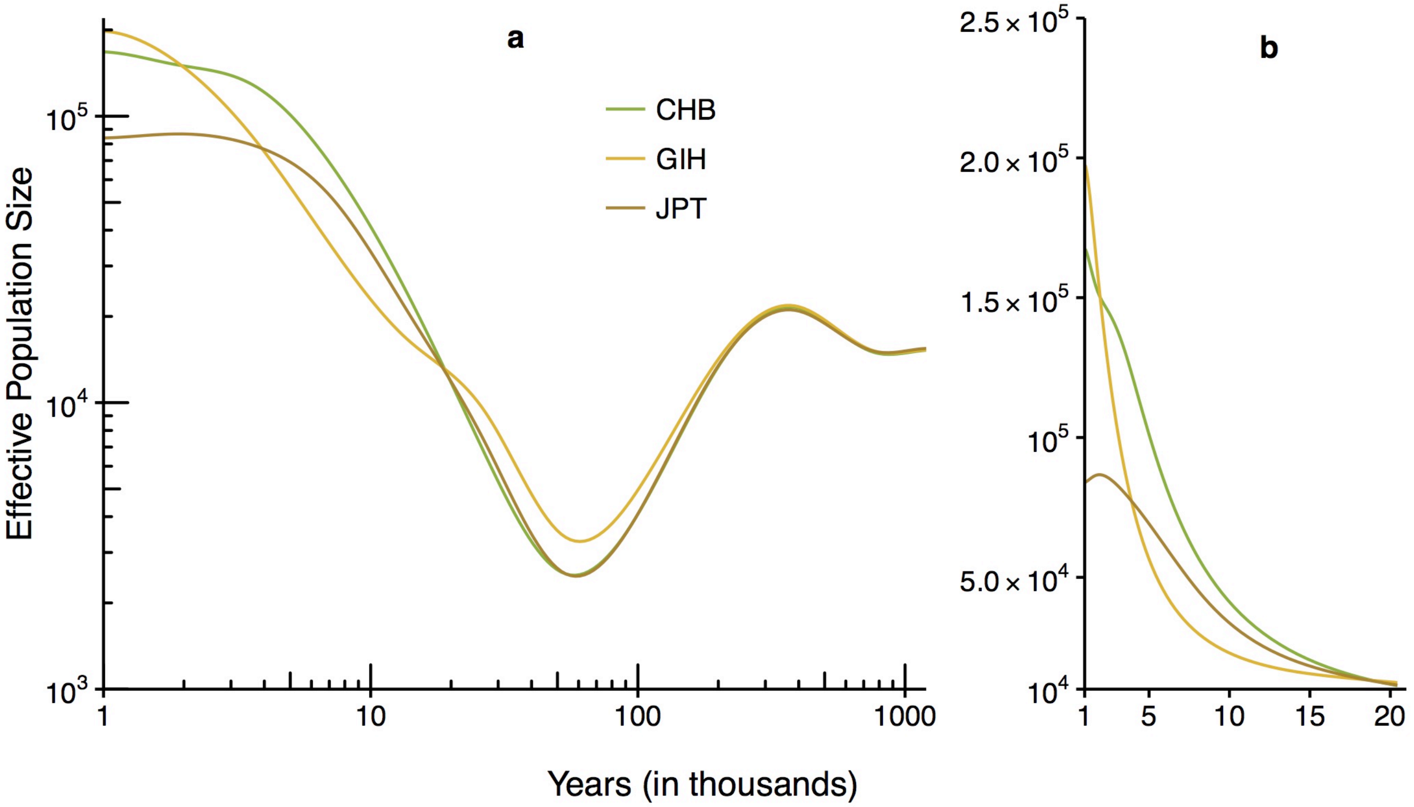
Results of demographic inference when $\hat{\rho}$ is not known.

Each step plot represents inference on a single simulated data set with sample size *n* = 50. The colors of the estimated size histories indicate the ratio of recombination to mutation used in each simulation, which was not known to SMC++ during model fitting. The ratio ranged from 1:10 (black) to 10:1 (light blue). The true demography used for simulation is indicated in bold black. The nested scatterplot compares the true versus estimated ratio of recombination to mutation rates. The mutation rate θ/2 was assumed to be known. SMC++ is able to fairly accurately estimate the recombination rate over two orders of magnitude with respect to the mutation rate, and is most accurate when the mutation and recombination rates are approximately equal.



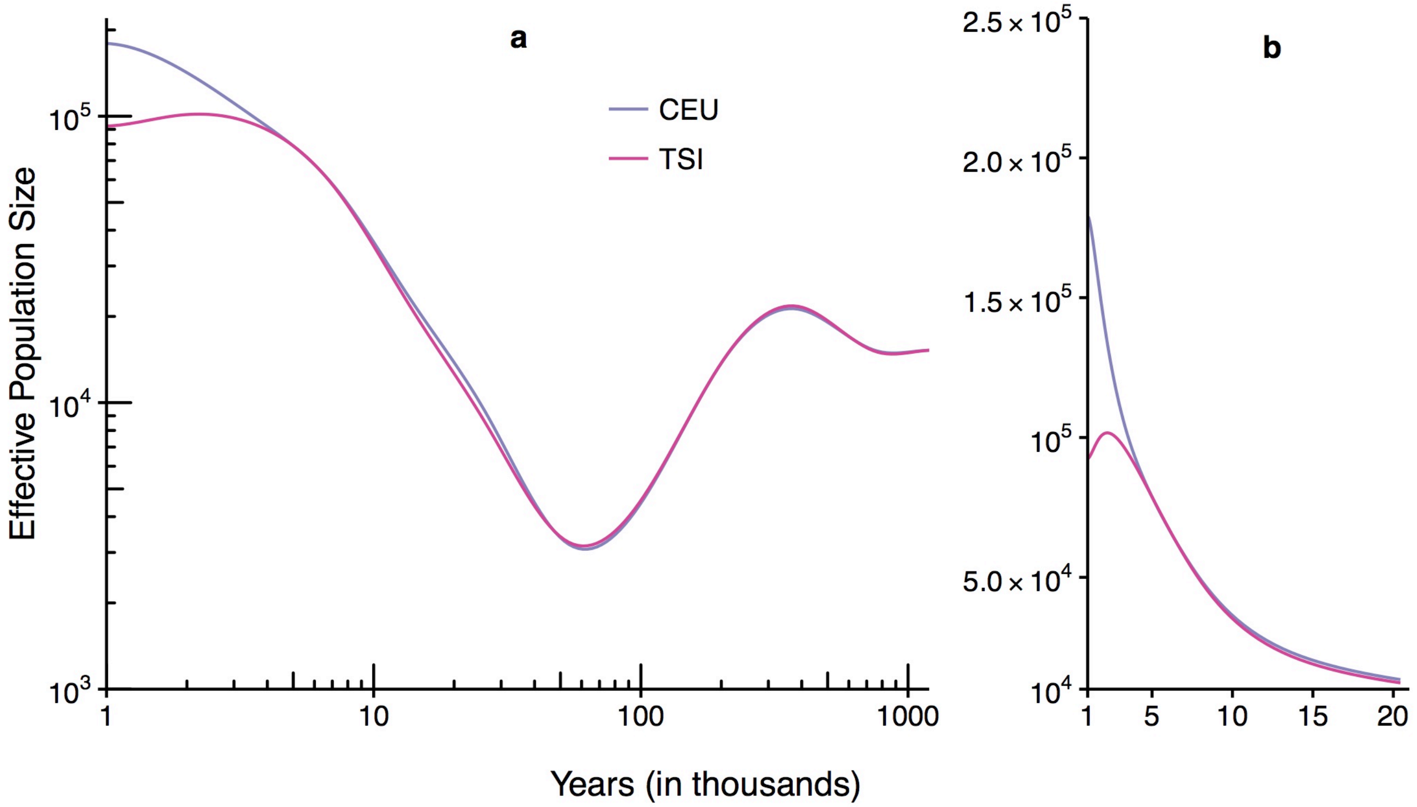
Supplementary Figure 2

Results of demographic inference across three African subpopulations.



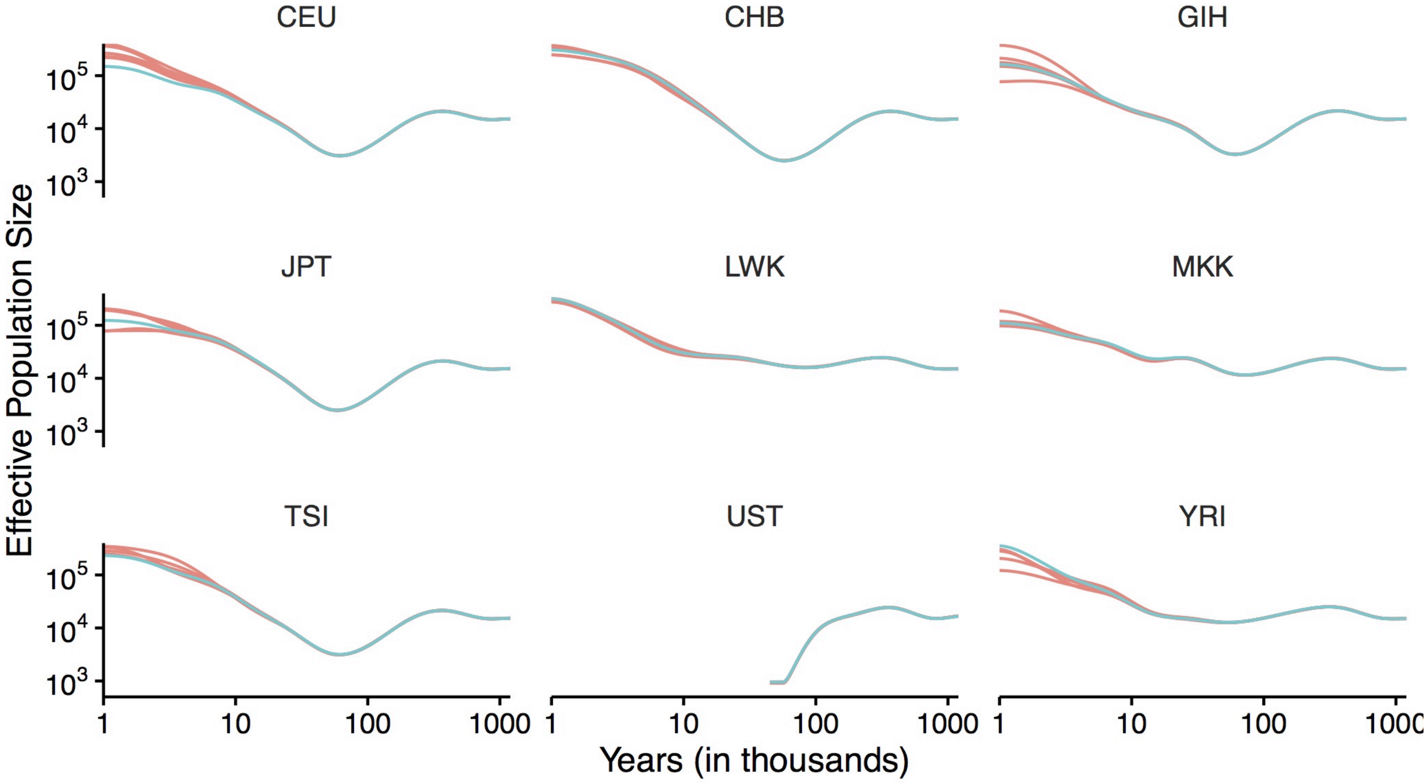
Supplementary Figure 3

Results of demographic inference across three Asian subpopulations.



Supplementary Figure 4

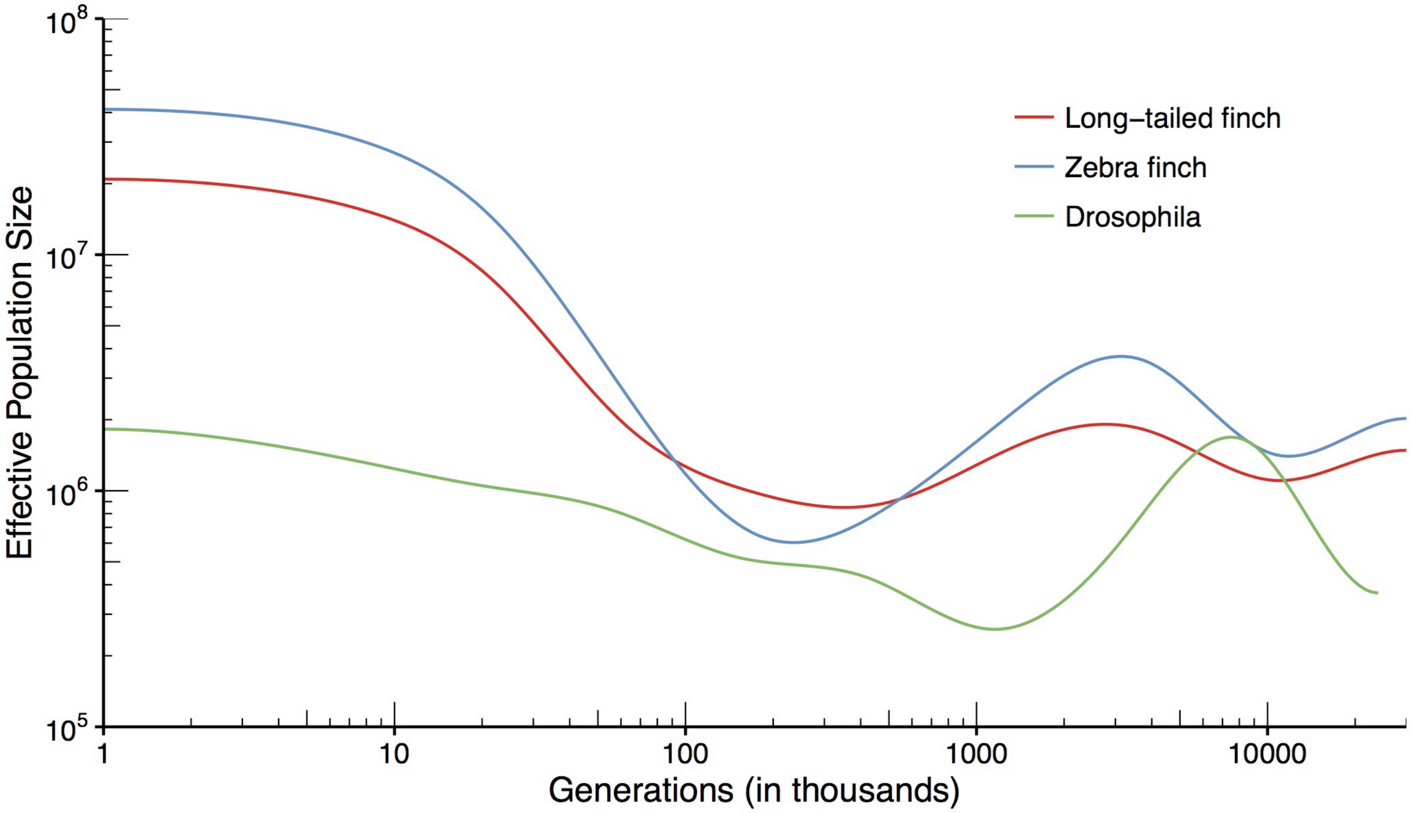
Results of demographic inference across two European subpopulations.



Supplementary Figure 5

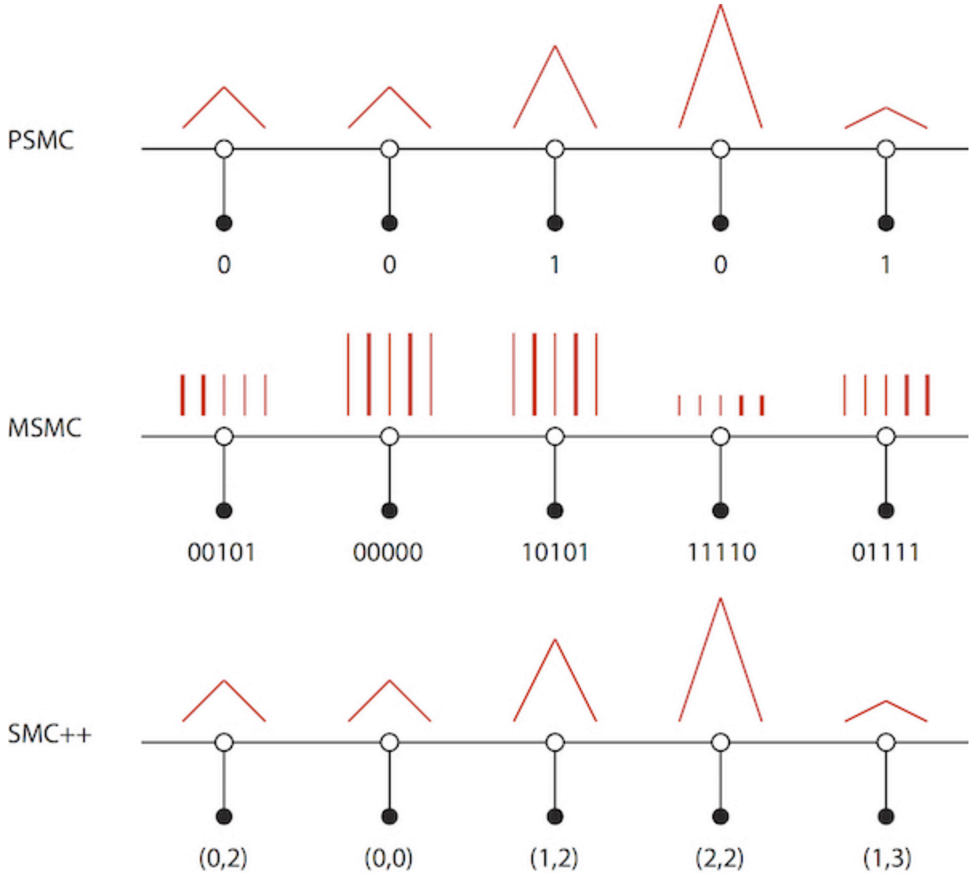
Sensitivity analysis for human demographic inference.

Blue lines are reproduced from Figure **5**. Red lines represent the result of randomly downsampling the data to contain 90% of the original set of chromosomes and re-running the analysis.



Supplementary Figure 6

Results of analyzing non-human species, in generations.



Supplementary Figure 7

Schematic of the differences between PSMC, MSMC, and SMC++.

The HMM used in PSMC tracks the hidden TRMCA of a pair of haploid lineages, and emits binary symbols based on the heterozygosity of this pair at each block of sites. MSMC tracks the hidden time to first coalescence among several haploid lineages, as well as the identity (denoted by the bolded bars) of the two lineages that coalesce first. It considers as emissions the allelic state of all lineages in the sample. SMC++, like PSMC, tracks the TMRCA in only a pair of individuals, and emits 2-tuples whose distribution is given by the conditioned SFS (Section S1).