Supporting Information

The coalescent model

The coalescent is a stochastic model for the genealogical tree of a sample of $n$ sequences. It has become a standard tool in population genetics, and it is frequently used for simulating genomic data (see e.g. [9, 7]). The coalescent model provides an approximation of the genealogy of a sample containing $n$ genes, for a large class of models in populations genetics for which the population size is constant [3]. A coalescent event occurs when two lineages find a common ancestor. In a sample of $n \geq 3$ sequences, the first coalescence event corresponds to the first time when a pair of sequences finds a common ancestor. After the first coalescence event, the two sequences that coalesced are replaced by their common ancestor so that the initial sample of $n$ sequences is represented by $n - 1$ ancestral sequences. The second coalescence event corresponds to the first time when a pair of sequences, chosen among the $n - 1$ ancestral sequences, coalesces. This process is repeated until a unique common ancestor is found. In the coalescent, time is measured backward and in units of $N$ generations where $N$ is the diploid population size. This rescaling of time means that one unit of time corresponds to $N$ generations. The inter-coalescence times $Y_j$, $2 \leq j \leq n$, measure the amount of time during which the sample has $j$ ancestors. These times form independent random variables distributed according to the exponential distribution of rate $j(j - 1)/2$, for $2 \leq j \leq n$, and their joint distribution can then be
computed as a product distribution.

In varying population size models, the joint distribution of the $Y_j$’s is no longer a product of exponential distributions. A description of the distribution of the inter-coalescence times, $Y_j$, in a population whose size evolves deterministically, is given in [1]. For a large class of deterministic demographic models, Griffiths and Tavaré [1] showed that, for $2 \leq j \leq n$, we have

$$
\mathbb{P}(Y_j > t | Y_n + \ldots + Y_{j+1} = s) = \exp \left( -\frac{j(j - 1)}{2} \int_s^{s+t} \lambda(u) du \right), \quad s, t \geq 0
$$

where $\lambda(t) = N/N(t)$, $N$ is the present population size and $N(t)$ is the population size $t$ units of time before present. In this context, time is measured in units of the present population size $N$. For simulation, it is useful to consider the process $A^\nu(.)$ that counts the number of distinct ancestors of the sample at any time. The above equation implies that the coalescent in a population whose size evolves deterministically can be obtained from the standard coalescent with a deterministic change of time

$$
A^\nu(t) = A(\Lambda(t)), \quad t > 0,
$$

where $A(.)$ is the process that counts the number of ancestors in a constant-size model and

$$
\Lambda(t) = \int_0^t \lambda(s) \, ds.
$$

In the population expansion model, a population of constant size $N_A$ started to grow $t_0$ years ago to reach $N = N(0)$ individuals at the present time. When time is counted in units of $N$ generations, the ratio of the population
size, \( t \) units of time ago, to the present population size is then given by
\[
\frac{N}{N(t)} = \begin{cases} 
\alpha, & t \geq \nu_0 \\
\alpha^{t/\nu_0}, & 0 \leq t < \nu_0.
\end{cases}
\]
Assuming 20 years for the generation time, we set \( \nu_0 = t_0/20N \).

**Mutational models and summary statistics**

In coalescent models, we generally assume that mutations occur according to a Poisson process of rate \( \theta/2 \) where \( \theta = 2N\mu \), with \( \mu \) denoting the mutation rate per gene per generation and \( N \) denoting the diploid population size. It means that for a branch of length \( \ell \) in the coalescent, the number of mutations on that branch is given by a Poisson distribution of mean \( \theta\ell/2 \), independent on mutations occurring in other branches.

**Infinitely-many-sites model**

The infinitely-many-sites model assumes that every mutation affects a site that has not been hit by mutation before. All sites in the DNA sequence are assumed to be completely linked so that no recombination event occurred within the sequence. The number of segregating sites \( S_n \) corresponds to the number of mutations in the coalescent tree. The distribution of \( S_n \) depends on the coalescent tree through its total length \( L_n \) only
\[
L_n = \sum_{j=2}^{n} jY_j. \tag{2}
\]
Conditional on \( L_n \), the number of segregating sites \( S_n \) has a Poisson distribution of mean \( \theta L_n/2 \) where \( L_n \) is given by equation (2). In the constant population size model, \( L_n \) is a sum of exponentially distributed random variables of rate \( (j - 1)/2, \ j = 2, \ldots, n. \)
Single-step mutation model in an expanding population

In the population expansion model, \( n \) individuals are surveyed at unlinked microsatellite markers. At a microsatellite locus, a motif of two to four nucleotides may repeat itself several times, and the number of repeats is usually variable within the population. The simplest model of mutation is the single-step model [4] which assumes that a mutation may be, with equal probability, a deletion or an insertion of a single motif. Looking at the number of repeats of the motif, the single-step mutation model of [4] can be viewed as a random walk for which the +1 and the −1 steps are equally likely. Here we assume that the microsatellite markers are independent, meaning that they are located either on the same chromosome but separated by a large physical distance, or that they are located on distinct chromosomes. Simulating replicates from this model requires generating independent coalescent genealogies on which mutations are superimposed randomly.

To capture the pattern of genetic diversity, we computed 7 summary statistics. First, we summarized the amount of genetic diversity by the mean (over the loci) of the variance in the number of repeats (up to a multiplicative constant of 2) and the mean (over the loci) of the heterozygositites. These two measures of the amount of genetic diversity are known to be strongly correlated to the ancestral population size [6]. At each locus, estimates of the variance and the homozygosity (one minus the heterozygosity) can be computed as follows

\[
V = \sum_{i \neq j} (X_i - X_j)^2 = \frac{2}{n - 1} \sum_{k=1}^{n} (X_k - \bar{X})^2
\] (3)
and

\[ P_0 = \frac{n \sum_{a \in A} p_a^2 - 1}{n - 1} \]  

(4)

where \( X_k \) is the number of repeats for individual \( k \), \( \hat{X} = \sum_{k=1}^{n} X_k/n \), \( A \) is the set of alleles represented in the sample and \( p_a \) is the frequency of alleles of size \( a \) [2]. There has been a substantial amount of work in the populations genetics community to find statistics sensitive to the intensity of the demographic expansion for microsatellite data. Here, we used the two imbalance indices [2]

\[ \log \beta_1 = \log \bar{V} - \log [(\bar{P}_O - 2 - 1)/2], \]

and

\[ \log \beta_2 = \frac{1}{n_{\text{loci}}} \sum_{l=1}^{n_{\text{loci}}} (\log V)_l - (\log P_0)_l \]

where \( \bar{V} \) and \( \bar{P}_O - 2 \) are averages over the loci of the single locus estimates given by equations (3) and (4), the subscript \( l \) runs over all loci, and \( n_{\text{loci}} \) denotes the total number of loci. We also considered the interlocus \( g \) statistic of [5], and the expansion index of [10] which are sensitive to the genomic signature of a demographic expansion. The expression of the \( g \) statistic is given by

\[ g = \frac{\text{Observed Var}(V)}{\frac{4}{3} V^2 + \frac{1}{3} V}, \]

where Observed Var(\( V \)) is the variance over the loci of the single locus values of \( V \) (equation (3)). For the single-step mutation model, the expansion index of [10] is given by

\[ S = 1 - \frac{\bar{Q} - \bar{V} / 2}{5 \bar{V}^2}, \]
where $Q$ is the single locus estimate of the fourth moment of the $X_k$'s and $\bar{Q}$ denotes the average over the loci of the single-locus estimates. We also introduced a new summary statistic whose computation is based on an observation of Shriver et al. [8]. They studied the $S_K$ distribution where $S_K$ is defined as the proportion of pairwise comparisons that differ by $K$ repeat units. The $S_K$ distribution was noticed to have its peak at the value 0 for recent expansion, and the peak shifts to the value 1 for older expansions. We then defined a new index by

$$s = P_1 - P_0,$$

and we averaged this index over all the loci.
Figure 1: Example 2. The posterior quantiles of the ratio $\alpha$ of the ancestral population to the present population size under the exponentially growing population model. The quantiles are plotted against the tolerance rate for the NCH and the ANCH methods (left) and LocL ABC method (right). For each algorithm, the curves represent the 0.025, 0.25, 0.5, 0.75, and 0.975 quantiles of the posterior distribution. These values correspond to the median over 100 replicates. The ground truth value is represented by the blue line.
Figure 2: Example 3. The boxplot of the posterior quantiles for $\theta_1$ as a function of the tolerance rate. The different ABC methods were run 100 times each using 2,000 simulations at each run.
Figure 3: Example 3. The boxplot of the posterior quantiles for $\theta_2$ as a function of the tolerance rate. The different ABC methods were run 100 times each using 2,000 simulations at each run.
References


