File S1

Supporting Information

Derivation of equation (4)

Since $Y_2 = X_0 + X_2$ and $Y_1 = X_0 + X_1$, we have

$$\begin{aligned}
E[Y_2; Y_1 &= i|T=\tau] \\
&= E[X_0 + X_2; X_0 + X_1 = i|T=\tau] \\
&= \sum_j E[X_0 + X_2; X_0 = j; X_1 = i-j|T=\tau] \\
&= \sum_j E[X_0; X_0 = j; X_1 = i-j|T=\tau] + \sum_j E[X_2; X_0 = j; X_1 = i-j|T=\tau] \\
&= \sum_j jP(X_0 = j, X_1 = i-j|T=\tau) \\
&\quad + \sum_j E[X_2|T=\tau]P(X_0 = j, X_1 = i-j|T=\tau) \\
&= \sum_j jP(X_0 = j|T=\tau)P(X_1 = i-j|T=\tau) \\
&\quad + E[X_2|T=\tau]\sum_j P(X_0 = j, X_1 = i-j|T=\tau) \\
&= \sum_j jP(X_0 = j|T=\tau)P(X_1 = i-j|T=\tau) + E[X_2|T=\tau]P(X_0 + X_1 = i|T=\tau)
\end{aligned}$$

Derivation of equations (5) and (6)

If X is a discrete random variable taking values in the non-negative integers $\{0, 1, 2, ...\}$, then we define the probability generating function (pgf) of X as:

$$f_X(s) = E[s^X] = \sum_{x=0}^{\infty} P(X = x)s^x$$

Restating equation (4) in the terms of the probability generating functions, and using independence of $(X_0|T)$, $(X_1|T)$, $(X_2|T)$, we obtain

$$\begin{aligned}
\sum_i E[Y_2, Y_1 &= i|T=\tau]s^i \\
&= \sum_i E[X_2|T=\tau]P(X_0 + X_1 = i|T=\tau)s^i \\
&\quad + \sum_i \sum_j jP(X_0 = j|T=\tau)P(X_1 = i-j|T=\tau)s^{j-1}s^{i-j}s \\
&= E[X_2|T=\tau]f_{X_0+X_1|T=\tau}(s) + s\sum_i f'_{X_0|T=\tau}(s)P(X_1 = i-j|T=\tau)s^{i-j} \\
&= E[X_2|T=\tau]f_{X_0+X_1|T=\tau}(s) + sf'_{X_0|T=\tau}(s)\sum_{i=-\infty}^{\infty} P(X_1 = i = j|T=\tau)s^{i-j} \\
&= E[X_2|T=\tau]f_{X_0+X_1|T=\tau}(s) + sf'_{X_0|T=\tau}(s)\sum_{i-j=-\infty}^{\infty} P(X_1 = i = j|T=\tau)s^{i-j} \\
&\quad (since\ -\infty < j < \infty) \\
&= E[X_2|T=\tau]f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) + sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)
\end{aligned}$$

Thus, equation (5) holds; derivation of equation (6) is similar.

Derivation of equations (7) and (8)

Consider the Poisson Process N with intensity ν, where $N(T) = \#$ {events of mutations occurring in time interval of length T},

$f_{N(T)}(s) = e^{\nu t(s-1)}$ and T is the coalescence time before present. Let X_0, X_1, X_2 denote the incremental changes of allele length of chromosomes 0, 1 and 2. $(X_0|T)$, $(X_1|T)$, $(X_2|T)$ are conditionally independent and $f_{U_k}(s) = \psi_k(s)$, $k = 0, 1, 2$.

Therefore, if $\tau \leq t_0$, it holds $X_0|(T = \tau) = X_{0,1}|(T = \tau) + X_{0,2}|(T = \tau)$, where $X_{0,1}|T = \tau$ is the increment of allele size in the interval $[t, t_0]$ and $X_{0,2}|T = \tau$ is the increment in the interval $[t_0, \tau)$. Since $X_{0,1}|(T = \tau)$ is independent of $X_{0,2}|(T = \tau)$, it holds that pgf $f_{X_0|T=\tau}(s) = f_{X_{0,1}|T=\tau}(s) \cdot f_{X_{0,2}|T=\tau}(s)$, where $f_{X_{0,1}|T=\tau}(s) = f_{N(t-t_0)}(\psi_0(s)) = e^{\nu_0(t-t_0)(\psi_0(s)-1)}$, and $f_{X_{0,2}|T=\tau}(s) = f_{N(t_0-\tau)}(\psi_1(s)) = e^{\nu_1(t_0-\tau)(\psi_1(s)-1)}$.

It follows that $f_{X_0|T=\tau}(s) = exp(\nu_0(t-t_0)(\psi_0(s)-1)+\nu_2(t_0-\tau)(\psi_1(s)-1))$, and $f_{X_1|T=\tau}(s) = f_{N(\tau)}(\psi_1(s)) = exp(\nu_1\tau(\psi_1(s)-1))$. Therefore, both equations (7) and (8) hold for $\tau \leq t_0$. Derivations for $t_0 < \tau \leq t$ or $\tau > t$ are similar.

Derivation of computational expressions for equations (10) and (11)

Expected size of allele drawn from population 2, jointly with size of allele drawn from population 1 being equal to i is equal to $E[Y_2, Y_1 = i] = \int_0^\infty E[y_2, Y_1 = i|T = \tau] f_T(\tau) d\tau$, where $f_T(\tau)$ is given in equation (1). However, T denotes the common ancestor time of chromosomes 1 and 2 sampled at time 0 from populations 1 and 2. Therefore, from Equ. (1), we have

$$f_T(\tau) = \begin{cases} 0 & \tau \leq t_0 \\ \frac{1}{2N_0} e^{-(\tau - t_0)/(2N_0)} & \tau > t_0 \end{cases}$$

If $\tau \leq t_0$, $f_T(\tau) = 0$ and $E[Y_2; Y_1 = i] = 0$.

If $t_0 < \tau \leq t$, from Equ. (5) we obtain $\sum_i E[Y_2, Y_1 = i|T = \tau] s^i = E[X_2|T = \tau] f_{X_0|T=\tau}(s) f_{X_1|T=\tau}(s) + s f'_{X_0|T=\tau}(s) f_{X_1|T=\tau}(s)$

where,

$$\begin{aligned} E[X_2|T &= \tau] f_{X_0|T=\tau}(s) f_{X_1|T=\tau}(s) \\ &= E[X_2|T = \tau] exp(\nu_0(t - t_0)(\psi_0(s) - 1) + \nu_1 t_0(\psi_1(s) - 1)) \\ &= E[X_2|T = \tau] exp(\nu_0(t - t_0)(b_0 s + \frac{d_0}{s} - 1) + \nu_1 t_0(b_1 s + \frac{d_1}{s} - 1) \\ &= E[X_2|T = \tau] exp((\nu_0(t - t_0) b_0 + \nu_1 t_0 b_1) s + \\ &\quad + (\nu_0(t - t_0) d_0 + \nu_1 t_0 d_1)/s - \nu_0(t - t_0) - \nu_1 t_0) \\ &= E[X_2|T = \tau] e^{((\nu_0(t-t_0)b_0 + \nu_1 t_0 b_1)s + (\nu_0(t-t_0)d_0 + \nu_1 t_0 d_1)/s)} e^{-\nu_0(t-t_0) - \nu_1 t_0} \\ &\quad (t_0 < \tau \leq t) \end{aligned}$$

By differentiating $f_{X_2|T=\tau}(s)$ and setting $s = 1$, we obtain

$E[X_2|T = \tau] = (\nu_0(\tau - t_0)(b_0 - d_0) + \nu_2 t_0(b_2 - d_2)) exp(\nu_0(\tau - t_0)(b_0 + d_0 - 1) + \nu_2 t_0(b_2 + d_2 - 1))(t_0 < \tau \leq t)$. We denote $b = \nu_0(t - t_0)b_0 + \nu_1 t_0 b_1$ and $d = \nu_0(t - t_0)d_0 + \nu_1 t_0 d_1$, to obtain $e^{((\nu_0(t-t_0)b_0 + \nu_1 t_0 b_1)s + (\nu_0(t-t_0)d_0 + \nu_1 t_0 d_1)/s)} = e^{(bs + d/s)}$. According to Equ. (9),

$$e^{(bs+d/s)} = \sum_{i \in Z} I_i(2\sqrt{bd})(\frac{b}{d})^{i/2} s^i \quad = \sum_{i \in Z} \beta_i s^i \quad , |s| = 1,$$

where we denote $\beta_i = I_i(2\sqrt{bd})(\frac{b}{d})^{i/2}$, and $I_i = I_{-i}$ is the modified Bessel function of the first type (Abramowitz and Stegun 1972), of integer order i

Thus

$$E(X_2|T=\tau)f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s) = e^{-\nu_0(t-t_0)-\nu_1 t_0}\sum_{i\in Z} E(X_2|T=\tau)\beta_i s^i$$

Furthermore, $sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)$ in equation (5) can be expressed as

$$\begin{aligned}
&sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)\\
&=s\nu_0(t-\tau)\psi_0'(s)\times exp(\nu_0(t-t_0)(\psi_0(s)-1)+\nu_1 t_0(\psi_1(s)-1))\\
&=s\nu_0(t-\tau)(b_0-d_0/s^2)\times e^{((\nu_0(t-t_0)b_0+\nu_1 t_0 b_1)s+(\nu_0(t-t_0)d_0+\nu_1 t_0 d_1)/s)}e^{-\nu_0(t-t_0)-\nu_1 t_0}\\
&=\nu_0(t-\tau)(b_0 s-d_0/s)e^{-\nu_0(t-t_0)-\nu_1 t_0}(\sum_{i\in Z}\beta_i s^i)\\
&=e^{-\nu_0(t-t_0)-\nu_1 t_0}(\sum_{i\in Z}\nu_0(t-\tau)b_0\beta_i s^{i+1}-\sum_{i\in Z}\nu_0(t-\tau)d_0\beta_i s^{i-1})\\
&=e^{-\nu_0(t-t_0)-\nu_1 t_0}(\nu_0(t-\tau)b_0\sum_{(i-1)\in Z}\beta_{i-1}s^i-\nu_0(t-\tau)d_0\sum_{(i+1)\in Z}\beta_{i+1}s^i)\\
&=e^{-\nu_0(t-t_0)-\nu_1 t_0}(\nu_0(t-\tau)b_0\sum_{i\in Z}\beta_{i-1}s^i-\nu_0(t-\tau)d_0\sum_{i\in Z}\beta_{i+1}s^i)
\end{aligned}$$

Therefore, when $t_0<\tau\le t$

$$\begin{aligned}
\sum_i E[Y_2,Y_1=i|T=\tau]s^i &= E[X_2|T=\tau]f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)+sf'_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)\\
&= e^{-\nu_0(t-t_0)-\nu_1 t_0}\sum_{i\in Z}[E[X_2|T=\tau]\beta_i+\nu_0(t-\tau)b_0\beta_{i-1}-\nu_0(t-\tau)d_0\beta_{i+1}]s^i\\
&= e^{-\nu_0(t-t_0)-\nu_1 t_0}\sum_{i\in Z}\{\nu_0(t-\tau)b_0\beta_{i-1}-\nu_0(t-\tau)d_0\beta_{i+1}\\
&\quad+[\nu_0(\tau-t_0)(b_0-d_0)+\nu_2 t_0(b_2-d_2)]e^{\nu_0(\tau-t_0)(b_0+d_0-1)+\nu_2 t_0(b_2+d_2-1)}\beta_i\}s^i
\end{aligned}$$

which yields

$$\begin{aligned}
E[Y_2,Y_1 &= i|T=\tau] = e^{-\nu_0(t-t_0)-\nu_1 t_0}\{\nu_0(t-\tau)b_0\beta_{i-1}-\nu_0(t-\tau)d_0\beta_{i+1}\\
&+[\nu_0(\tau-t_0)(b_0-d_0)+\nu_2 t_0(b_2-d_2)]e^{\nu_0(\tau-t_0)(b_0+d_0-1)+\nu_2 t_0(b_2+d_2-1)}\beta_i\}
\end{aligned}$$

and provides the desired computable form of Equ. (10)

If $\tau>t$, analogous computations yield

$$\begin{aligned}
E[Y_2,Y_1=i] &= \int_t^\infty E[Y_2,Y_1=i|T=\tau]f_T(\tau)d\tau\\
&= E[Y_2,Y_1=i|T=\tau]P[T>\tau]\\
&= e^{-\frac{t-t_0}{2N_0}-\nu_1 t_0-\nu_0(t-t_0)}\beta_i(\nu_0(t-t_0)(b_0-d_0)+\\
&\quad+\nu_2 t_0(b_2-d_2))e^{\nu_0(t-t_0)(b_0+d_0-1)+\nu_2 t_0(b_2+d_2-1)}
\end{aligned}$$

Similarly as derivation of Equ. (10), Equ. (11) can be fully derived from Equ. (2), (6) and (9) for its computable form, where

if $\tau\le t_0$,

$E[Y_1',Y_1=i|T=\tau]=e^{-\nu_0(t-t_0)-\nu_1 t_0}\{[\nu_0(t-t_0)b_0+\nu_1(t_0-\tau)b_1]\beta_{i-1}+\nu_1\tau(b_1-d_1)e^{\nu_1\tau(b_1+d_1-1)}\beta_i-[\nu_0(t-t_0)d_0+\nu_1(t_0-\tau)d_1]\beta_{i+1}\}$

if $t_0<\tau\le t$,

$E[Y_1',Y_1=i|T=\tau]=e^{-\nu_0(t-t_0)-\nu_1 t_0}\{\nu_0(t-\tau)b_0\beta_{i-1}-\nu_0(t-\tau)d_0\beta_{i+1}+[\nu_0(\tau-t_0)(b_0-d_0)+\nu_1 t_0(b_1-d_1)]e^{\nu_0(\tau-t_0)(b_0+d_0-1)+\nu_1 t_0(b_1+d_1-1)}\beta_i\}$

and if $\tau>t$,

$$E[Y_1', Y_1 = i|T = \tau] = e^{-\nu_0(t-t_0)-\nu_1 t_0}[\nu_0(t-t_0)(b_0-d_0)+\nu_1 t_0(b_1-d_1)]e^{\nu_0(t-t_0)(b_0+d_0-1)+\nu_1 t_0(b_1+d_1-1)}\beta_i$$

Derivation of computational expression for equation (12)

The probability generating function of $(Y_1|T) = (X_0|T) + (X_1|T)$ is equal to $f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)$ because $(X_0|T)$ is conditionally independent of $(X_1|T)$.

Therefore,

$$\begin{aligned}\sum_i P(Y_1 = i|T = \tau)s^i &= f_{X_0|T=\tau}(s)f_{X_1|T=\tau}(s)\\ &= exp(\nu_0(t-t_0)(\psi_0(s)-1)+\nu_1 t_0(\psi_1(s)-1))\\ &= e^{-\nu_0(t-t_0)-\nu_1 t_0}\sum_{i\in Z}\beta_i s^i\end{aligned}$$

which yields $P(Y_1 = i) = \int_0^\infty P(Y_1 = i|T = \tau)f_T(\tau)d\tau = e^{-\nu_0(t-t_0)-\nu_1 t_0}\beta_i$, the derived computational expression for Equ. (12).

Derivation of equations (13) and (14)

By definition,

$$\begin{aligned}E\ [Y_2|Y_1 \geq x] &= \frac{E[Y_2, Y_1 \geq x]}{P(Y_1 \geq x)}\\ &= \frac{\sum_j jP(Y_2 = j, Y_1 \geq x)}{\sum_{i\geq x} P(Y_1 = i)}\\ &= \frac{\sum_j j\sum_{i\geq x} P(Y_2 = j, Y_1 = i)}{\sum_{i\geq x} P(Y_1 = i)}\\ &= \frac{\sum_{i\geq x}(\sum_j jP(Y_2 = j, Y_1 = i))}{\sum_{i\geq x} P(Y_1 = i)}\\ &= \frac{\sum_{i\geq x} E[Y_2, Y_1 = i]}{\sum_{i\geq x} P(Y_1 = i)}\end{aligned}$$

Equation (13) has been derived. Similarly, Equ. (14) can be derived.

Derivation of the range for the estimate of t

We denote random variable L as the incremental allele length of a sampled individual, and show the computation of the expectation $E[L]$ and variance$V[L]$ given $t, t_0, \nu_0, \nu_1, b_0$ and b_1, where t and t_0 are expressed in generation units.

Let L_0, L_1 be two random variables that denote the allele length increments in time interval t to t_0 (ancestral population 0) and t_0 to present (cognate population 1), respectively.

Let X_i be the incremental change in the the ith generation of the ancestral population (from t to t_0). We may obtain $P(X_i = 1) = \nu_0 b_0$, $P(X_i = -1) = \nu_0(1-b_0)$, $P(X_i = 0) = 1-\nu_0$ and

$$\begin{aligned}E[X_i] &= 2\nu_0 b_0 - \nu_0\\ V[X_i] &= \nu_0 - \nu_0^2(2b_0-1)^2\end{aligned}$$

Based on the fact that the incremental change of allele length in any generation is independent of that in any other generation, we have

$$\begin{aligned}E[L_0] &= (t-t_0)E[X_i] = (t-t_0)\nu_0(2b_0-1)\\ V[L_0] &= (t-t_0)V[X_i] = (t-t_0)\nu_0[1-\nu_0(2b_0-1)^2]\end{aligned}$$

Similarly we derive $E[L_1]$ and $V[L_1]$ as

$$\begin{aligned} E[L_1] &= t_0\nu_1(2b_1-1) \\ V[L_1] &= t_0\nu_1[1-\nu_1(2b_1-1)^2] \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} E[L] &= E[L_0]+E[L_1] = (t-t_0)\nu_0(2b_0-1)+t_0\nu_1(2b_1-1) \\ V[L] &= V[L_0]+V[L_1] \\ &= (t-t_0)\nu_0[1-\nu_0(2b_0-1)^2]+t_0\nu_1[1-\nu_1(2b_1-1)^2] \end{aligned}$$

Assuming that L is approximately Gaussian and given x as the threshold of allele size discovery, we calculate $P(L \geq x)$ from the cdf of the Gaussian distribution. This helps to determine a realistic range of estimates of t to control the probability of locus discovery.

Derivation of $E[Y_1^{'}; Y_1 = i]$ in the extended model

From the Equ. (16) derived earlier on, we obtain

$$f_T(\tau) = \begin{cases} (\frac{1}{2N_m})exp[-\sum_{k=m+1}^{L}(\frac{t_{k-1}-t_k}{2N_k}) - \frac{\tau-t_m}{2N_m}]; & t < \tau \leq t_{m-1}, \\ exp[-\sum_{k=1}^{L}(\frac{t_{k-1}-t_k}{2N_k}) - \frac{\tau-t_0}{2N_0}]; & \tau > t_0. \end{cases}$$

If $\tau \leq t_0$ and $t_m < \tau \leq t_{m-1}$ $(m = 1, 2, ..., L)$, we obtain

$$\begin{aligned} E[Y_1^{'}; Y_1 &= i|T=\tau] = e^{-\nu_0(t-t_0)-\nu_1 t_0}\{[\nu_0(t-t_0)b_0+\nu_1(t_0-\tau)b_1]\beta_{i-1} \\ &+ \nu_1\tau(b_1-d_1)\beta_i - [\nu_0(t-t_0)d_0+\nu_1(t_0-\tau)d_1]\beta_{i+1}\} \end{aligned}$$

$$E[Y_1^{'}; Y_1 = i] = \sum_{m=1}^{L}\int_{t_m}^{t_{m-1}} E[Y_1^{'}; Y_1 = i|T=\tau](\frac{1}{2N_m})exp[-\sum_{k=m+1}^{L}(\frac{t_{k-1}-t_k}{2N_k}) - \frac{\tau-t_m}{2N_m}]d\tau$$

where β_i is the same as that defined and used in the derivation of Eqs. (10) and (11) in the main text.
If $t_0 < \tau \leq t$, we obtain

$$\begin{aligned} E[Y_1^{'}; Y_1 &= i|T=\tau] = e^{-\nu_0(t-t_0)-\nu_1 t_0}\{\nu_0(t-\tau)b_0\beta_{i-1} \\ &+ [\nu_0(\tau-t_0)(b_0-d_0)+\nu_1 t_0(b_1-d_1)]\beta_i - \nu_0(t-\tau)d_0\beta_{i+1}\} \end{aligned}$$

$$E[Y_1^{'}; Y_1 = i] = \int_{t_0}^{t} E[Y_1^{'}; Y_1 = i|T=\tau](\frac{1}{2N_0})exp[-\sum_{k=1}^{L}(\frac{t_{k-1}-t_k}{2N_k}) - \frac{\tau-t_0}{2N_0}]d\tau$$

If $\tau > t$, we obtain

$$E[Y_1^{'}; Y_1 = i|T=\tau] = e^{-\nu_0(t-t_0)-\nu_1 t_0}[\nu_0(t-t_0)(b_0-d_0)+\nu_1 t_0(b_1-d_1)]\beta_i$$

which is not a function of τ.

Therefore

$$\begin{aligned} E[Y_1^{'}; Y_1 &= i] = \int_t^{\infty} E[Y_1^{'}; Y_1 = i|T=\tau]f_T(\tau)d\tau = E[Y_1^{'}; Y_1 = i|T=\tau]P(T>t) \\ &= e^{-\sum_{k=1}^{L}(\frac{t_{k-1}-t_k}{2N_k}) - \frac{\tau-t_0}{2N_0} - \nu_0(t-t_0)-\nu_1 t_0}[\nu_0(t-t_0)(b_0-d_0)+\nu_1 t_0(b_1-d_1)]\beta_i \end{aligned}$$

With the analytical derivation shown here we are able to compute the allele length difference D with Human population size arbitrarily varied from generation to generation.