

Linkage Disequilibrium Under Recurrent Bottlenecks

E. Schaper,^{*,1} A. Eriksson,[†] M. Rafajlovic,^{*} S. Sagitov,[‡] and B. Mehlig^{*,2}

^{*}Department of Physics, University of Gothenburg, SE-41296 Gothenburg, Sweden, [†]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom, and [‡]Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-41296 Gothenburg, Sweden

ABSTRACT To model deviations from selectively neutral genetic variation caused by different forms of selection, it is necessary to first understand patterns of neutral variation. Best understood is neutral genetic variation at a single locus. But, as is well known, additional insights can be gained by investigating multiple loci. The resulting patterns reflect the degree of association (linkage) between loci and provide information about the underlying multilocus gene genealogies. The statistical properties of two-locus gene genealogies have been intensively studied for populations of constant size, as well as for simple demographic histories such as exponential population growth and single bottlenecks. By contrast, the combined effect of recombination and sustained demographic fluctuations is poorly understood. Addressing this issue, we study a two-locus Wright–Fisher model of a population subject to recurrent bottlenecks. We derive coalescent approximations for the covariance of the times to the most recent common ancestor at two loci in samples of two chromosomes. This covariance reflects the degree of association and thus linkage disequilibrium between these loci. We find, first, that an effective population-size approximation describes the numerically observed association between two loci provided that recombination occurs either much faster or much more slowly than the population-size fluctuations. Second, when recombination occurs frequently between but rarely within bottlenecks, we observe that the association of gene histories becomes independent of physical distance over a certain range of distances. Third, we show that in this case, a commonly used measure of linkage disequilibrium, σ_d^2 (closely related to \hat{r}^2), fails to capture the long-range association between two loci. The reason is that constituent terms, each reflecting the long-range association, cancel. Fourth, we analyze a limiting case in which the long-range association can be described in terms of a Xi coalescent allowing for simultaneous multiple mergers of ancestral lines.

GENETIC variation at a single neutral locus has been investigated in great detail for population models under different demographic processes, such as population expansions, single bottlenecks, or genetic hitchhiking caused by nearby selective sweeps (see, for example, Eriksson *et al.* 2008 for a review of such models). Biological populations exhibit abundance fluctuations on both short and long timescales, caused by, *e.g.*, environmental and ecological changes. Such size fluctuations in the form of repeated bottlenecks are characteristic of populations expanding into new territories. Examples include the human out-of-Africa scenario (Ramachandran *et al.* 2005; Liu *et al.* 2006), the accompanying expansion of the parasite *Plasmodium falciparum* causing

severe malaria (Tanabe *et al.* 2010), and the recolonization by the marine snail *Littorina saxatilis* of Sweden's west coast archipelago (Johannesson 2003). Genetic variation in populations subject to bottlenecks is now routinely investigated in the laboratory. England *et al.* (2003), for example, have studied genetic variation in *Drosophila melanogaster* populations subjected to bottlenecks of different durations and strengths. Last but not least, bottlenecks can also be due to environmental fluctuations. Pujolar *et al.* (2011), for example, have investigated populations of *Salmo marmoratus* subject to weather-induced recurrent bottlenecks.

It is common practice to accommodate such fluctuations in the theory by using an effective population size instead of the census population size. See Ewens (1982) for a review of different measures of the effective population size.

Recent research has highlighted the importance of two competing timescales in the context of such effective population-size approximations: the timescale of the population-size fluctuations and the coalescent timescale (which reflects the time to the most recent common

Copyright © 2012 by the Genetics Society of America
doi: 10.1534/genetics.111.134437

Manuscript received September 5, 2011; accepted for publication October 27, 2011

¹Present addresses: Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland, and Computer Science Department, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland.

²Corresponding author: University of Gothenburg, SE-41296 Gothenburg, Sweden.
E-mail: mehlig@fy.chalmers.se

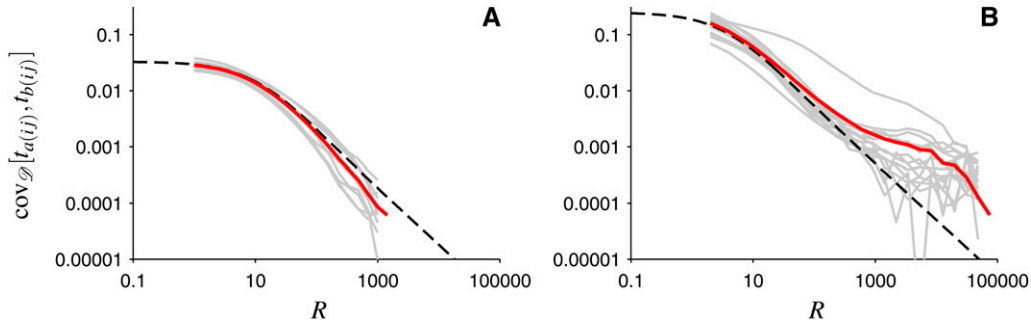


Figure 1 Covariance of the times to the MRCA at two loci, in a sample of two chromosomes in a population subject to repeated bottlenecks (details in *Model* section). (A) Rapid population-size fluctuations. Wright–Fisher simulations are shown for 10 random sequences of bottlenecks with $p = 10^{-3}$, $q = 10^{-3}$, $N_0 = 10^5$, and $N_B = 10^4$ (gray lines). Each gray line is obtained by first generating a random sequence of bottlenecks and then averaging over an ensemble of 1000 gene genealogies. The red line shows the covariance averaged over demographic histories. The dashed line shows the result of the effective population-size approximation (Equation 2). (B) Same as in A, but for the case of severe reductions of population size during bottlenecks. Wright–Fisher simulations are shown for 15 randomly generated sequences of bottlenecks, with parameters $p = 10^{-5}$, $q = 2 \times 10^{-2}$, $N_0 = 10^6$, and $N_B = 5 \times 10^2$. Averages are over 100 gene genealogies for each demographic history.

sequence of bottlenecks and then averaging over an ensemble of 1000 gene genealogies. The red line shows the covariance averaged over demographic histories. The dashed line shows the result of the effective population-size approximation (Equation 2). (B) Same as in A, but for the case of severe reductions of population size during bottlenecks. Wright–Fisher simulations are shown for 15 randomly generated sequences of bottlenecks, with parameters $p = 10^{-5}$, $q = 2 \times 10^{-2}$, $N_0 = 10^6$, and $N_B = 5 \times 10^2$. Averages are over 100 gene genealogies for each demographic history.

ancestor, MRCA). When demographic fluctuations are much slower than the coalescent timescale, they can be ignored and the effective population size can be approximated by the initial population size (Sjödin *et al.* 2005). In the opposite case of rapid demographic fluctuations, it has been argued (Wright 1938; Crow and Kimura 1970) that genetic variation is well described in terms of a population with effective population size N_{eff} , given by the harmonic average of the population size:

$$N_{\text{eff}} = \lim_{T \rightarrow \infty} \left(\frac{1}{T} \sum_{\tau=0}^T \frac{1}{N_{\tau}} \right)^{-1}. \quad (1)$$

Here N_{τ} is the population size in generation τ . See Jagers and Sagitov (2004), Sjödin *et al.* (2005), and Wakeley and Sargsyan (2009) for recent developments of the concept of an effective population size. In conclusion, for both fast and slow demographic fluctuations, the statistical properties of single-locus gene genealogies agree with those of the constant population-size model. By contrast, when both timescales are of the same order, it has been shown (Kaj and Krone 2003; Nordborg and Krone 2002; Sjödin *et al.* 2005; Eriksson *et al.* 2010) that the distribution of total branch lengths in samples of single-locus gene genealogies does not in general agree with that predicted by the standard coalescent approximation. But Eriksson *et al.* (2010) have shown how to compute moments of the distribution of the total branch lengths of gene genealogies at a single locus, conditional on a given demographic history (see also Zivkovic and Wiehe 2008). In summary, the effect of population-size fluctuations upon genetic variation at a single locus is well understood.

But how do population-size fluctuations affect multilocus patterns of genetic variation on the same chromosome? Such patterns are influenced by recombination. Genetic recombination introduces a new timescale that is inversely proportional to r , the probability of recombination between a pair of loci per generation. As is well known, genetic recombination plays an important role in shaping empirically observed multilocus patterns of genetic variation in biological populations. Measures of linkage disequilibrium quan-

tify the degree of association of genetic variation at pairs of loci on the same chromosome. Common measures of linkage disequilibrium, such as r^2 (Hill and Robertson 1968), and its approximation σ_d^2 (Ohta and Kimura 1971; McVean 2002), depend upon the allelic frequencies at two loci. These measures are thus closely related to the covariance of the times (*i.e.*, the number of generations) to the MRCA of the underlying gene genealogies (McVean 2002).

Figure 1 shows this covariance as a function of genetic distance between the two loci. The results shown were obtained by computer simulations of the Wright–Fisher dynamics (Fisher 1930, 1999; Wright 1931) of a population experiencing recurrent bottlenecks, with random durations of and random separations between bottlenecks. Details of the model are given in the *Model* section. In Figure 1, gray lines show how the covariance of the times to the MRCA for two chromosomes at two loci (averaged over all pairs of loci the same distance apart) depends on genetic distance. Each gray line corresponds to a single realization of the sequence of bottlenecks. The red lines in Figure 1 are the averages of the covariances within each panel. Figure 1, A and B, shows results for two different sets of parameters of the model. In Figure 1A, the bottlenecks happen frequently and have short durations. In this case, the single-locus properties are expected to be in good agreement with those of a population with effective population size given by Equation 1. For such populations, the coalescent approximation predicts

$$\text{cov} [t_{a(ij)}, t_{b(ij)}] = x_{\text{eff}}^2 \frac{R x_{\text{eff}} + 18}{(R x_{\text{eff}})^2 + 13 R x_{\text{eff}} + 18}. \quad (2)$$

(Griffiths 1981; Hudson 1983, 1990). Here $t_{a(ij)}$ and $t_{b(ij)}$ denote the times to the MRCA of two loci (called a and b) in a sample of two chromosomes (denoted by i and j). In Equation 2, $R = 2N_0r$ is a scaled recombination rate, and $x_{\text{eff}} = N_{\text{eff}}/N_0$ is the effective population size relative to the population size at the present time, denoted by N_0 . Units of time are chosen so that $\tau = \lfloor tN_0 \rfloor$, and $\lfloor tN_0 \rfloor$ is the largest integer not larger than tN_0 . In Figure 1 effective population-size approximations, according to Equation 2, are shown as

dashed lines. In Figure 1A we observe good agreement between Equation 2 and the average covariance of the times to the MRCA obtained from computer simulations of the model. However, in Figure 1B, Equation 2 agrees with the simulated covariance only for short genetic distances. For large genetic distances, by contrast, Figure 1B shows that the covariance decreases much more slowly than expected according to Equation 2. Thus, the results shown in Figure 1B imply long-range association between two loci.

The examples shown in Figure 1 raise many questions. What are the conditions for the effective population-size approximation to be valid in the multilocus case? Why does it fail when these conditions are not met? How significant are deviations of the exact result from the effective population-size approximation? Why does long-range association between two loci appear in some cases? How large are fluctuations around the covariance of the coalescent times, averaged over an ensemble of gene genealogies and over different demographic histories? What is the significance of the fluctuations around such averages for data analysis?

The aim of this article is to provide answers to the above questions by computing the covariance of the times to the MRCA for a pair of chromosomes under a model of recurrent bottlenecks introduced below in the *Model* section. Our analysis enables us to qualitatively and quantitatively determine the effects of fluctuating population size on the two-locus statistics in terms of the timescales of population-size fluctuations, of coalescence, and of recombination. Using both analytical (*Covariance of the Times to the MRCA*) and numerical approaches (*Comparison of the Coalescent Calculations to the Wright–Fisher Simulations*), we estimate the range of validity of the effective population-size approximation for the two-locus case. We find that the effective population-size approximation inevitably fails for large recombination rates: the failure is sometimes minor (as in the case shown in Figure 1A) and sometimes significant (as in the case shown in Figure 1B). By taking different limits of the parameters of the model, we provide both a qualitative and a quantitative understanding of how the effective population-size approximation may fail in predicting the long-range association between two loci. We demonstrate in *The Effect of Recurrent Bottlenecks upon σ_d^2* that the long-range association has a surprisingly small effect upon σ_d^2 . In the *Discussion* we review our results. In particular, in the limit where bottlenecks correspond to severe reductions of population size during bottlenecks, we show that gene genealogies in our model can be described in terms of the so-called Xi coalescent (Schweinsberg 2000; Möhle and Sagitov 2001; Sagitov 2003), which allows for simultaneous multiple coalescent events (called multiple mergers below).

In summary we describe in this article how sustained population-size variations in the form of sequences of bottlenecks influence two-locus patterns of genetic variation. We conclude this Introduction by briefly commenting on related models. The effect of recurrent bottlenecks on single-locus statistics was investigated by Sjödin *et al.*

(2005) and also by Eriksson *et al.* (2010). The model introduced by Sargsyan and Wakeley (2008) contains a recurrent bottleneck model for a single locus as a special case. Sargsyan and Wakeley (2008) show under which conditions single-locus gene genealogies in their model can be approximated by either Kingman's coalescent or the Xi coalescent. Eldon and Wakeley (2008), finally, have analyzed two-locus gene genealogies under a population model allowing for skewed reproduction. We discuss the connection between the results of Eldon and Wakeley (2008) and our results in the *Discussion* (see Equation 14).

Model

Using a Wright–Fisher model of a population of gametes, we trace the ancestry of two loci on a pair of gametes backward in time, until the MRCA of both loci is found. In each generation τ we perform two steps. In the first step, each gamete is independently subjected to recombination such that, with probability r , the two loci segregate onto two different gametes (corresponding to different parents). We assume that r is approximately proportional to the physical distance between the given loci (see McPeck and Speed 1995 for a review of the general relation between recombination rate and physical distance). Because of this step, the ancestors of the original loci may be spread over up to four different gametes in any given generation in the past (if recombination leads to a gamete where neither locus is ancestral to the original sample, that gamete is dropped from further consideration). In the second step, the parent of each gamete is chosen randomly from the gametes in the parental generation. Whenever a pair of gametes have the same parent, the ancestries of these gametes are the same, and the number of gametes to be traced is reduced.

To investigate the effect of population-size fluctuations on genetic variation, we consider a model of recurrent bottlenecks in which the population size can take one of two values, N_0 or N_B . Here N_B denotes the population size in the bottlenecks, and N_0 stands for the population size between bottlenecks. We write $N_B = xN_0$, with $0 < x < 1$. The probability of changing the population size, going one generation back in time, depends on the population size at the current time. The switching probabilities are denoted by p and q when the current population sizes are N_0 and N_B , respectively. Hence, the expected durations of the high and the low population-size phases are $1/p$ and $1/q$ generations, respectively. The population size in the first generation is taken to be N_0 . For a single locus, as mentioned above, such a model has been investigated by Sjödin *et al.* (2005) and Eriksson *et al.* (2010); see also Sargsyan and Wakeley (2008).

Figure 2, A and B, illustrates population-size fluctuations in this recurrent bottleneck model. In Figure 2, C and D, examples of gene genealogies of two loci (called a and b) in a sample of two chromosomes are shown. In Figure 2, C and D, generations with low population size (N_B) are marked yellow, and otherwise the population size is N_0 .

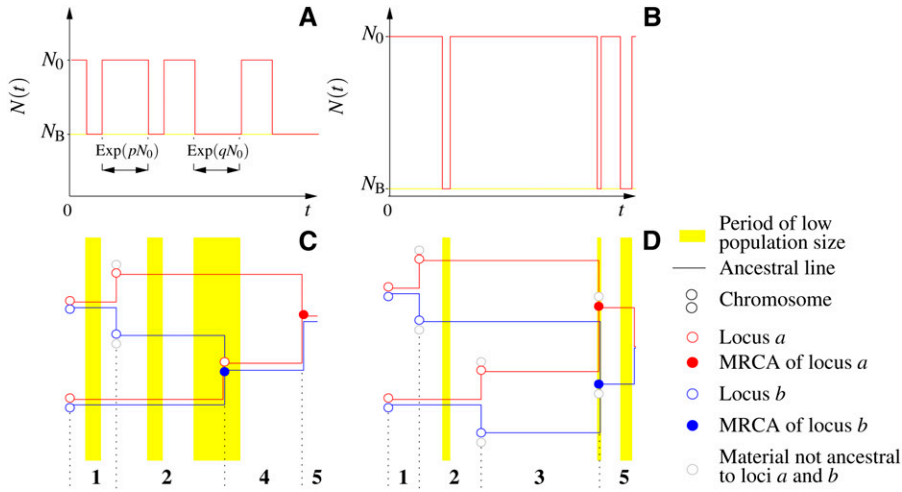


Figure 2 (A and B) Two realizations of the population-size curve, $N(t)$, backward in time ($t = 0$ denotes the present time). Initially, the population size is N_0 . Going backward in time, the population size randomly jumps between two values, N_0 and N_B , with transition rates pN_0 (from N_0 to N_B) and qN_0 (from N_B to N_0). (C and D) Examples of ancestral histories of two loci (blue and red open circles correspond to two loci, called a and b) subject to genetic recombination in a sample of two chromosomes. Times at which the population was subject to a bottleneck are shaded yellow. Two joint circles depict two loci in the same chromosome. The numbers $1, \dots, 5$ denote the possible states of the system (they are explained in detail in Figure 3). Gray circles denote genetic material not ancestral to the sampled loci. Blue and red solid circles indicate that the corresponding loci have

found their most recent common ancestor. Note that bottlenecks can host more than one coalescent event. In the case of severe reductions of population size during bottlenecks, such multiple coalescences appear as simultaneous multiple mergers on the timescale of the gene genealogy. An example is shown in D: an almost instantaneous transition from state 3 to state 5.

Each chromosome is represented by a pair of lines (red and blue lines correspond to loci a and b , respectively). In generations where a common ancestor is found for a pair of ancestral lines or recombination between two loci occurs, the chromosomes are represented by circles instead of lines. MRCAs are shown as solid circles. In some cases recombination causes the ancestry of one locus to become associated with a chromosome that lacks direct descendants in the sample (open gray circles). The ancestries of such segments of DNA are irrelevant to the gene genealogy of the sample, and these ancestral lines are not traced further.

Covariance of the Times to the MRCA

Consider a sample of two chromosomes (denoted by i and j) and two loci (called a and b). As in the Introduction, the time to the MRCA at locus a is denoted by $t_{a(ij)}$ [and $t_{b(ij)}$ at locus b]. Consider $\text{cov}_{\mathcal{D}} [t_{a(ij)}, t_{b(ij)}]$, the covariance conditional on a particular demographic history \mathcal{D} . Taking the average of the conditional covariance over random demographic histories, we have

$$\begin{aligned} & \langle \text{cov}_{\mathcal{D}} [t_{a(ij)}, t_{b(ij)}] \rangle \\ &= \langle \langle t_{a(ij)} t_{b(ij)} | \mathcal{D} \rangle - \langle t_{a(ij)} | \mathcal{D} \rangle \langle t_{b(ij)} | \mathcal{D} \rangle \rangle \quad (3) \\ &= \langle t_{a(ij)} t_{b(ij)} \rangle - \langle \langle t_{a(ij)} | \mathcal{D} \rangle^2 \rangle. \end{aligned}$$

Here $\langle \dots | \mathcal{D} \rangle$ denotes the expectation conditional on a particular demographic history \mathcal{D} . In the second equality we have used that the expected times to the MRCA are the same for both loci. Note that the averaged conditional covariance is not the same as the unconditional covariance of the times to the MRCA for the full process. The latter is given by

$$\langle t_{a(ij)} t_{b(ij)} \rangle - \langle t_{a(ij)} \rangle^2 \equiv \langle \text{cov}_{\mathcal{D}} [t_{a(ij)}, t_{b(ij)}] \rangle + \text{var} [\langle t_{a(ij)} | \mathcal{D} \rangle]. \quad (4)$$

We remark that biological data correspond to a particular individual realization \mathcal{D} . The additional term in Equation 4 reflects the variance over different demographic histories of the average time to the MRCA at a single locus and is thus irrelevant to analyzing the covariance of the times to the MRCA in a particular empirical data set.

We now derive an approximate expression for $\langle \text{cov}_{\mathcal{D}} [t_{a(ij)}, t_{b(ij)}] \rangle$, Equation 3. The method we employ requires the typical timescales of coalescence, of population-size fluctuations, and of recombination to be large. We require that

$$N_0 \gg 1, N_B \gg 1, p \ll 1, q \ll 1, \text{ and } r \ll 1. \quad (5)$$

The first two conditions allow us to employ the standard coalescent approximation (Kingman 1982). The first four conditions allow us to neglect events involving simultaneous changes of population size and coalescence. Similarly, as a consequence of combining the first two conditions with the fifth one, we may omit events involving simultaneous coalescence and recombination. Finally, combining conditions three to five allows us to neglect simultaneous occurrences of population-size changes and recombination. Assuming large population sizes allows us to approximate the discrete generation index $\tau = 0, 1, 2, \dots$ by a continuous-time variable t . As mentioned in the Introduction, we choose the units of time such that $\tau = \lfloor tN_0 \rfloor$. We denote the population size at time t by $N(t)$.

Under the conditions summarized above, we find the following expression for the term $\langle \langle t_{a(ij)} | \mathcal{D} \rangle \langle t_{b(ij)} | \mathcal{D} \rangle \rangle \equiv \langle \langle t_{a(ij)} | \mathcal{D} \rangle^2 \rangle$ occurring in Equation 3:

$$\langle \langle t_{a(ij)} | \mathcal{D} \rangle^2 \rangle = \frac{\lambda_B (2x\lambda + \lambda_B + 3) + x\lambda(x\lambda + x + 2) + 2}{(\lambda + \lambda_B + 1)(\lambda + \lambda_B + 2)}. \quad (6)$$

This equation follows from Equation 20 in Eriksson *et al.* (2010). The parameters λ and λ_B in Equation 6 are given by $\lambda = pN_0$ and $\lambda_B = qN_B$.

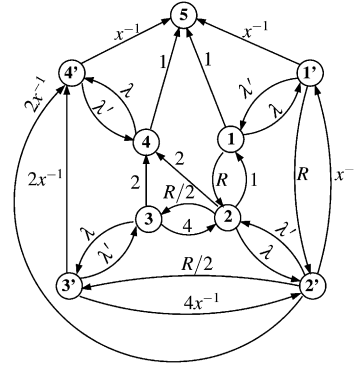
To evaluate the remaining term in Equation 3, $\langle t_{a(ij)}t_{b(ij)} \rangle$, we adapt the method described by Eriksson and Mehlig (2004) for calculating the covariance of the times to the MRCA at two loci. As can be seen in Figure 2, C and D, there are only a small number of possible combinations of ancestral lines in gene genealogies of two loci for two chromosomes. Thus, we can write down a Markov process for how states of the ancestral lines change along a gene genealogy. The corresponding graph is shown in Figure 3, where the vertices represent states (combinations of ancestral lines), and the edges represent transitions between the states (the transition rates are shown along the edges). The vertices labeled by a prime correspond to states with population size N_B (bottlenecks). The expected value $\langle t_{a(ij)}t_{b(ij)} \rangle$ is determined by a subgraph of the graph shown in Figure 3, consisting of the vertices 1, 2, 3, 1', 2', and 3'. Let \mathbf{M} be the corresponding 6×6 transition matrix. Its off-diagonal elements M_{mn} , $m \neq n$, are given by the transition rates w_{mn} . The diagonal elements M_{nn} are equal to the negative sum of the rates of all edges leaving node n in the graph, $M_{nn} = -\sum_{m \neq n} M_{mn}$. In terms of \mathbf{M} we can write

$$\langle t_{a(ij)}t_{b(ij)} \rangle = \int_0^\infty dt_1 t_1^2 \mathbf{u} e^{\mathbf{M} t_1} \mathbf{v} + \int_0^\infty dt_1 \int_{t_1}^\infty dt_2 t_1 t_2 \mathbf{c} e^{\mathbf{K}(t_2-t_1)} \mathbf{Q} e^{\mathbf{M} t_1} \mathbf{v} \quad (7)$$

(Eriksson and Mehlig 2004). The vectors \mathbf{u} , \mathbf{v} , and \mathbf{c} and the matrices \mathbf{Q} and \mathbf{K} in Equation 7 are given by

$$\begin{aligned} \mathbf{v} &= [1 \ 0 \ 0 \ 0 \ 0 \ 0]^T, \quad \mathbf{u} = [1 \ 0 \ 0 \ x^{-1} \ 0 \ 0], \\ \mathbf{c} &= [1 \ x^{-1}], \\ \mathbf{Q} &= \begin{bmatrix} 0 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2x^{-1} & 2x^{-1} \end{bmatrix}, \\ \mathbf{K} &= \begin{bmatrix} -(\lambda + 1) & \lambda_B x^{-1} \\ \lambda & -(\lambda_B + 1)x^{-1} \end{bmatrix}. \end{aligned}$$

Here \top denotes the transpose, while $t_1 = \min(t_{a(ij)}, t_{b(ij)})$ and $t_2 = \max(t_{a(ij)}, t_{b(ij)})$ are the times to the first and second coalescent events, respectively. The first term in Equation 7 corresponds to a common MRCA of the loci a and b (i.e., to a transition from state 1 or 1' to state 5). The second term in Equation 7 corresponds to a different MRCA (transition from state 2 or 3 to state 4 or from state 2' or 3' to state 4', followed by a transition to state 5). At the present time the system is assumed to be in state 1, represented by the vector \mathbf{v} . The elements of the vector \mathbf{u} correspond to the transition rates from the states 1, 2, 3, 1', 2', and 3' to 5. The two rows in the matrix \mathbf{Q} contain the transition rates from 1, 2, 3 and 1', 2', 3' to 4 and 4', respectively. The matrix \mathbf{K} contains the transition rates between states 4 and 4', whereas the vector \mathbf{c} contains the coalescent rates in each population-size regime. Both \mathbf{M} and \mathbf{K} have negative real eigenvalues. Hence, the integrals in Equation 7 can be evaluated in terms of matrix inverses (Eriksson and Mehlig 2004). We find



group	state
1, 1'	$a_i b_i, a_j b_j$ $a_i b_j, a_j b_i$
2, 2'	$a_i \circ, \circ b_i, a_j b_j$ $a_i b_i, a_j \circ, \circ b_j$ $a_i \circ, \circ b_j, a_j b_i$ $a_i b_j, a_j \circ, \circ b_i$
3, 3'	$a_i \circ, \circ b_i, a_j \circ, \circ b_j$
4, 4'	$a_i \bullet, a_j \bullet$ $\bullet b_i, \bullet b_j$
5	$\bullet \bullet$

Figure 3 (Left) Graph showing the states and transition rates determining the ancestral history of two loci in a sample of two chromosomes, under the population model introduced in the *Model* section. States where the population is in a bottleneck are marked with a prime. The final state is denoted by 5 (in this state it does not matter whether the population is in a bottleneck or not). Arrows indicate transitions between states. The corresponding transition rates from state n to m , w_{mn} , are displayed next to the lines. For notational convenience we have used the abbreviations $x = N_B/N_0$, $\lambda = pN_0$, and $\lambda' = qN_0$. While x and λ are referred to in the main text, λ' is not. (Right) Table of possible states of the system. The two loci are denoted by a and b , and the corresponding chromosomes by i and j . Open circles denote genetic material not ancestral to sampled loci, and solid circles denote the MRCA.

$$\begin{aligned} \langle t_{a(ij)}t_{b(ij)} \rangle &= 2\mathbf{u}(-\mathbf{M})^{-3} \mathbf{v} \\ &+ 2\mathbf{c}(-\mathbf{K})^{-2} \left\{ (-\mathbf{K})^{-1} \mathbf{Q} + \mathbf{Q}(-\mathbf{M})^{-1} \right\} (-\mathbf{M})^{-2} \mathbf{v}. \end{aligned} \quad (8)$$

Combining Equations 6 and 8 yields

$$\langle \text{cov}_{\mathcal{D}} [t_{a(ij)}, t_{b(ij)}] \rangle = \frac{R^3 C_3 + R^2 C_2 + R C_1 + C_0}{R^4 D_4 + R^3 D_3 + R^2 D_2 + R D_1 + D_0}. \quad (9)$$

The coefficients C_0, \dots, C_3 , and D_0, \dots, D_4 , are functions of the parameters x , λ , and λ_B . They are given in *Appendix A*.

We now discuss three special cases of this result. First, we consider the case when the time to the first bottleneck is much longer than the expected time to the MRCA in the large population-size regime. Second, we discuss the case when the population-size fluctuations are much faster than all other processes. Third, we analyze the case of severe reductions of population size during bottlenecks.

In the first case one has $p \ll N_0^{-1}$. This implies that the time to the first bottleneck is much larger than the time to the MRCA, and as a consequence all bottlenecks are irrelevant to the mean covariance of the times to the MRCA. Formally, this case is obtained from Equation 9 by taking the limit $\lambda \rightarrow 0$. In this limit we obtain Equation 2 with $x_{\text{eff}} = 1$, which is the expression that is valid in the case of constant population size.

The second case is described by the conditions $p \gg N_0^{-1}$ and $q \gg N_B^{-1}$. Formally, this case corresponds to taking the limit $\lambda \rightarrow \infty$, and $\lambda_B \rightarrow \infty$ in Equation 9, in such a way that the ratio λ/λ_B is kept constant. In this limit we obtain

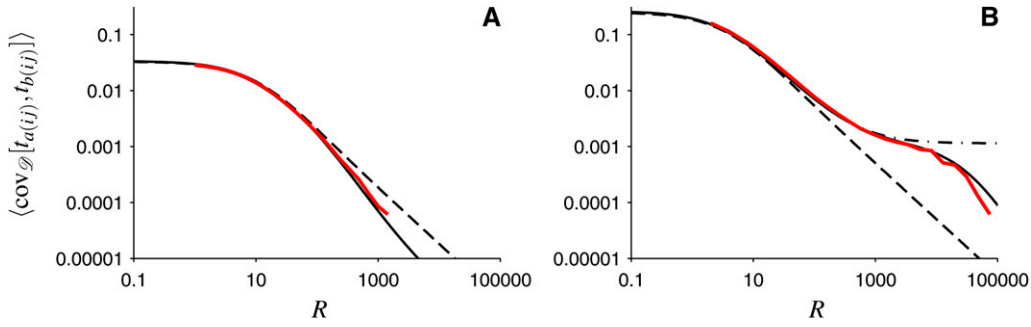


Figure 4 Covariance of the times to the MRCA of two loci averaged over random population-size histories. (A) The red line shows the average covariance corresponding to $p = 10^{-3}$, $q = 10^{-3}$, $N_0 = 10^5$, and $N_B = 10^4$, determined numerically from Wright–Fisher simulations (same as in Figure 1A). The solid line shows our exact result (Equation 9), and the dashed line shows the effective population-size approximation (Equation 2). The numerical result deviates from the effective population-size approximation when the recombination timescale is the smallest ($R > 100$). (B) The same as in A, but for the case of severe reductions of population size during bottlenecks: $p = 10^{-5}$, $q = 2 \times 10^{-2}$, $N_0 = 10^6$, and $N_B = 5 \times 10^2$. The dashed-dotted line denotes the result of Equation 10, corresponding to the Xi-coalescent approximation.

Equation 2, with $x_{\text{eff}} = (x + \lambda_B/\lambda)/(1 + \lambda_B/\lambda)$. This demonstrates that in this limit the resulting two-locus gene genealogies are described by the effective population-size approximation.

Third is the case of severe reductions of population size during bottlenecks, described by the condition $N_B \ll N_0$. Such demographic histories can occur during range expansions, where small groups of animals repeatedly colonize new areas [as mentioned in the Introduction, examples are the human out-of-Africa scenario (Ramachandran *et al.* 2005; Liu *et al.* 2006), the accompanying expansion of the parasite *P. falciparum* causing severe malaria (Tanabe *et al.* 2010), and the recolonization by the marine snail *L. saxatilis* of Sweden’s west coast archipelago (Johannesson 2003)]. This case can be treated analytically by taking the limit $x \rightarrow 0$ in Equation 9, keeping λ and λ_B fixed. In this limit we find

$$\left\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \right\rangle \approx \frac{R^2 A_2 + R A_1 + A_0}{R^2 B_2 + R B_1 + B_0}. \quad (10)$$

Here the coefficients $A_0, A_1, A_2, B_0, B_1,$ and B_2 are functions of the parameters λ and λ_B . They are given in Appendix A. Note that Equation 10 reaches a plateau for large values of R :

$$\left\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \right\rangle \approx \frac{2\lambda_B(1 + \lambda_B)\lambda}{(1 + \lambda_B + \lambda)(2 + \lambda_B + \lambda)(9(2 + \lambda) + \lambda_B(27 + 8\lambda + \lambda_B(10 + \lambda_B + \lambda)))}. \quad (11)$$

This expression implies long-range association between the two loci.

Comparison of the Coalescent Calculations to the Wright–Fisher Simulations

To further illustrate the role of the relevant timescales of genetic drift and recombination in shaping genetic association between two loci, we compare the full coalescent result, Equation 9, and the different limiting cases considered in *Covariance of the Times to the MRCA*, to the average covariance calculated from the Wright–Fisher simulations. These comparisons are shown in Figure 4.

The parameters used in Figure 4A correspond to rapid population-size fluctuations (the second case described in the previous section). As can be seen in Figure 4A, the agreement between the numerical result (red line) and the approximation (Equation 2) (dashed line) is good for a wide range of recombination rates. A small disagreement appears at large values of R , more precisely at $R \approx 100$ for the parameters chosen in Figure 4A. This discrepancy is expected, as at such large recombination rates the population-size fluctuations are no longer rapid compared to the process of recombination. But the corresponding deviations from the effective population-size approximation are very small. In summary, in the case of rapid population-size fluctuations, the results of the Wright–Fisher simulations are well approximated by the effective population-size approximation.

Figure 4B shows results for parameters corresponding to the third case analyzed in the previous section, the case of severe reductions of population size during bottlenecks. As we noted already in the Introduction, this case exhibits long-range association (between the two loci in question) that cannot be accurately described by the effective population-size approximation (Equation 2). The average covariance curve obtained by computer simulations (Figure 4B) can be divided into three regions in which the curve behaves qualitatively differently. First, at very small recombination rates, for the parameters chosen in Figure 4B, the covariance can be approximated using the effective population-size approximation. This is expected since in this region the population-size fluctuations are fast compared to the process of recombination. Second, the effective population-size approximation breaks down when $R \approx pN_0 = 10$ for the parameters chosen in Figure 4B. At recombination rates larger than this value, such that recombination occurs frequently between bottlenecks but rarely within, we find long-range association between the two loci. Hence, in this region the effective population-size approximation fails to describe the covariance of the times to the MRCA. However, Equation 10 agrees well with the resulting covariance in this region. We show in Appendix B that Equation 10 can be derived using the Xi-coalescent approximation, which allows for

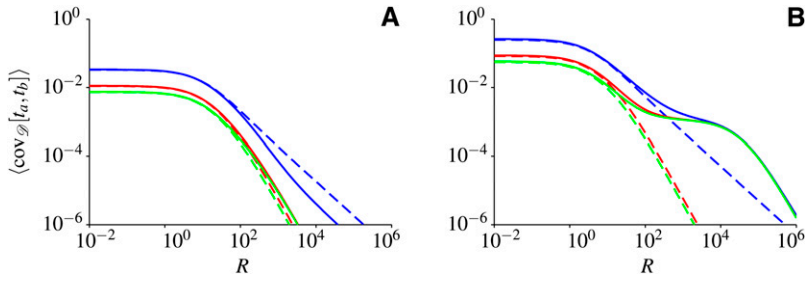


Figure 5 Average covariances $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \rangle$ (blue lines), $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ik)}] \rangle$ (red lines), and $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}] \rangle$ (green lines). (A and B) Values of the parameters p , q , N_0 , and N_B are the same as in Figures 1A and 4A. The parameters in B are the same as in Figures 1B and 4B. Exact results are shown as solid lines, whereas results obtained within the effective population-size approximation are shown as dashed lines.

simultaneous multiple coalescences of lines. Third, the agreement between Equation 10 and the full coalescent result breaks down when recombination events in bottlenecks can no longer be ignored (*i.e.*, when $R \approx qN_0$, where qN_0 is the rate of leaving a bottleneck). For still larger recombination rates, only the full coalescent result agrees with the Wright–Fisher simulations. The slight deviations between the Wright–Fisher simulations and Equation 9, visible in Figure 4B at large values of R , are discussed in the concluding section.

The Effect of Recurrent Bottlenecks upon σ_d^2

In the previous sections we showed how sustained population-size fluctuations in the form of recurrent bottlenecks can give rise to long-range association between loci, measured by the covariance of the times to the MRCA, Equation 3. An important question is how such population-size fluctuations affect common measures of linkage disequilibrium such as, for example, \hat{r}^2 , and its closely related measure, σ_d^2 , on the basis of a sample of n chromosomes (Ohta and Kimura 1971; McVean 2002). In this section, we discuss the effect of recurrent bottlenecks upon σ_d^2 in the case $n \gg 1$. In this case, Equation 9 in McVean (2002) becomes

$$\sigma_d^2 = \left\langle \frac{\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] - 2\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ik)}] + \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}]}{\langle t_{a(ij)} | \mathcal{D} \rangle^2 + \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}]} \right\rangle. \quad (12)$$

As before, a and b denote two loci, and i, j, k , and l denote four different chromosomes in the sample. The main properties of this measure are determined by how the numerator depends on the recombination rate (McVean 2002). To simplify the analysis, we therefore focus on the expected value of the numerator:

$$\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \rangle - 2\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ik)}] \rangle + \langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}] \rangle. \quad (13)$$

The mean of the conditional covariance $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \rangle$ is given by Equation 9. The covariances $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ik)}] \rangle$ and $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}] \rangle$ can be calculated in the same way that Equation 9 was obtained, but starting from different initial conditions (McVean 2002; Eriksson and Mehlis 2004). In our Markov representation, this corresponds to taking $\mathbf{v} = [0 \ 1 \ 0 \ 0 \ 0]^T$ and $\mathbf{v} = [0 \ 0 \ 1 \ 0 \ 0]^T$, respectively, in Equation 7.

Figure 5 shows how the covariances $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \rangle$ (blue lines), $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ik)}] \rangle$ (red lines), and $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}] \rangle$ (green lines) depend on R . Figure 5 demonstrates that when R is small, each covariance is well approximated by the corresponding effective population-size approximation. In the case shown in Figure 5B, all three covariances exhibit a plateau at the same level, in approximately the same range of R values. This demonstrates that all three covariances show long-range association between loci. But, in the linear combination (Equation 13) these enhancements cancel. We have checked that this conclusion also holds for σ_d^2 given by Equation 12. In Figure 6 we show for two random demographic histories how the covariances $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}]$ (blue lines), $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ik)}]$ (red lines), and $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}]$ (green lines) depend on R . Figure 6, A and B, shows results of two different demographic histories, but in both cases the parameters p , q , N_0 , and N_B are chosen to correspond to the case shown in Figures 4B and 5B. In both examples in Figure 6, a plateau appears in all three covariances of the times to the MRCA, as expected. These plateaus cancel in the same way as the three contributions to Equation 13.

Discussion

The aim of this article was to provide an understanding of how sustained population-size fluctuations influence the degree of association between two loci. Our conclusions are based on both analytical and numerical calculations, which we find to agree well. Using the population-size model depicted in Figure 2A, and assuming Wright–Fisher dynamics, we have derived a coalescent result for the covariance of the times to the MRCA of two loci, Equation 9. We have discussed three particular cases of our result. First, if the expected times to the MRCA are much smaller than the expected time to the most recent bottleneck, our model reduces to a constant population-size model with effective population size equal to the population size at the present time. Second, if the population size fluctuates much faster than the remaining two processes (coalescence and recombination), the effective population-size approximation works well with an effective population size given by Equation 1. Third, if the population size is severely reduced during bottlenecks, the results of our computer simulations depend on the relation between the timescales of recombination and of population-size changes. For the parameters chosen

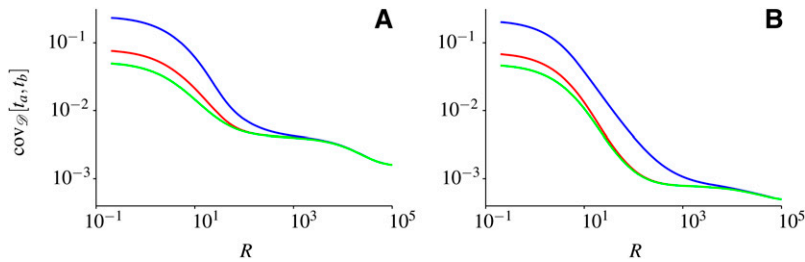


Figure 6 Covariances $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}]$ (blue lines), $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}]$ (red lines), and $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}]$ (green lines) for two random realizations of population-size history. In both A and B the parameters are $\rho = 10^{-4}$, $q = 0.2$, $N_0 = 10^5$, and $N_B = 50$, resulting in the same values of x , λ , and λ_B , as in Figures 4B and 5B.

in Figure 4B, we find that when recombination is the slowest process, that is, at very small values of recombination rates, the effective population-size approximation works well. By contrast, when recombination is frequent between bottlenecks but rare within bottlenecks, the effective population-size approximation fails to describe the covariance of the times to the MRCA. In this region the covariance is not only enhanced with respect to the result of the effective population-size approximation, but also approximately independent of R . In this case we say that pairs of distant loci exhibit long-range association.

These conclusions rely on analyzing covariances averaged over different demographic histories. This raises the question of how typical such averages are. In other words, how large are the fluctuations around the average? We observe that in the case shown in Figure 1B the fluctuations around the mean covariance are much higher than in the case shown in Figure 1A. However, in both cases shown, the averages represent the qualitative behavior of individual realizations.

The coalescent approximations employed in this article assume large population sizes. While we generally find very good agreement between the coalescent approximations and the Wright–Fisher simulations, we observe some deviations, in particular for large recombination rates in the case shown in Figure 1B. As expected, when we increase the parameter N_0 in our computer simulations, the deviations become smaller (results not shown).

In the remainder of this section, we discuss our result (Equation 10) in terms of the so-called Xi-coalescent approximation. Xi coalescents form a broad family of genealogical models allowing for simultaneous multiple mergers. The Kingman coalescent is a special case, allowing only for pairwise mergers. See Schweinsberg (2000) and Möhle and Sagitov (2001) for detailed descriptions of the family of Xi coalescents. We show in *Appendix B* how a Markov process with simultaneous multiple mergers is obtained in the case of severe reductions of population size during bottlenecks and compute the corresponding transition rates. This not only gives an alternative way of deriving Equation 10, it also provides insight into why a plateau forms. It turns out that the plateau arises as a direct consequence of simultaneous multiple mergers. This implies that long-range associations between two loci are also expected in other situations where simultaneous multiple mergers are important: Durrett and Schweinsberg (2004) have studied populations subject to selective sweeps, and Eldon and Wakeley

(2008) have demonstrated the importance of multiple mergers in shaping σ_d^2 in populations with skewed offspring distributions.

We note that there is a particular case where we can find a precise correspondence between the model of Eldon and Wakeley (2008) and the model analyzed in this article. This is the case of extreme reproductive success where one parent alone gives rise to the next generation. This case is obtained by setting $\omega = 1$ in the notation of Eldon and Wakeley (2008). In our model this corresponds to a population subject to one severe and infinitely strong bottleneck, that is, to the limit $x \rightarrow 0$ and $\lambda_B \rightarrow 0$. In this limit we find (averaging numerator and denominator in Equation 12 separately)

$$\sigma_d^2 \approx \frac{10 + R + 2\lambda}{22 + R^2 + 16\lambda + 2\lambda^2 + 13R + 3\lambda R}. \quad (14)$$

This result is equivalent to the equation for Y_2 on p. 1522 in Eldon and Wakeley (2008), identifying $\sigma_d^2 = Y_2$, $\lambda = \phi$, $R = 2\eta$, and setting $\omega = 1$, $\alpha = 1$, and $\beta = 1$ in the notation of Eldon and Wakeley (2008). We remark that for a population subject to a bottleneck, σ_d^2 was calculated explicitly by Eriksson and Mehlig (2004). Eriksson *et al.* (2009) discuss corresponding results within the sequential Markov coalescent approximation (McVean and Cardin 2005; Marjoram and Wall 2006).

We conclude with the observation that σ_d^2 , which is a function of the covariances $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}]$, $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}]$, and $\text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(kl)}]$, fails to show the plateaus observed in these covariances in Figure 6. This was observed already in Eriksson and Mehlig (2004) for the case of a single, recent bottleneck. Because of the close link between σ_d^2 and \hat{r}^2 , a common measure of linkage disequilibrium (McVean 2002), this casts doubt on the suitability of such measures for characterizing the degree of association between two loci [another example is the measure HR^2 (Sabatti and Risch 2002)], in populations that may have been subject to recent population bottlenecks and range expansions. A more accurate approach, especially for detecting long-range association between two loci, may be to estimate the covariance of the times to the MRCA directly. For example, simulations show that the covariance of the number of mutations in small windows (*e.g.*, a few hundred nucleotides long) can be used to estimate the covariance of the times to the MRCA (Eriksson and Mehlig 2004). However, it remains to

investigate which observables are most suitable for detecting long-range dependencies in the underlying gene genealogies for more general demographic histories.

Acknowledgments

Support by Swedish Research Council grants, by the Göran Gustafsson stiftelse, and by the Centre for Theoretical Biology at the University of Gothenburg are gratefully acknowledged. A.E. was supported by a Philip Leverhume Award and a Biotechnology and Biological Sciences Research Council grant (BB/H005854/1).

Literature Cited

- Birkner, M., J. Blath, M. Möhle, M. Steinrücken, and J. Tams, 2009 A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *ALEA* 6: 25–61.
- Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*, Harper & Row, New York.
- Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. *Theor. Popul. Biol.* 66: 129–138.
- Eldon, B., and J. Wakeley, 2008 Linkage disequilibrium under skewed offspring distribution among individuals in a population. *Genetics* 178(3): 1517–1532.
- England, P., G. Osler, L. Woodworth, M. Montgomery, D. Briscoe *et al.*, 2003 Effects of intense vs. diffuse population bottlenecks on microsatellite genetic diversity and evolutionary potential. *Conserv. Genet.* 4(5): 595–604.
- Eriksson, A., and B. Mehlig, 2004 Gene-history correlation and population structure. *Phys. Biol.* 1: 220–228.
- Eriksson, A., P. Fernstrom, B. Mehlig, and S. Sagitov, 2008 An accurate model for genetic hitchhiking. *Genetics* 178(1): 439–451.
- Eriksson, A., B. Mahjani, and B. Mehlig, 2009 Sequential Markov coalescent algorithms for population models with demographic structure. *Theor. Popul. Biol.* 76: 84–91.
- Eriksson, A., B. Mehlig, M. Rafajlovic, and S. Sagitov, 2010 The total branch length of sample genealogies in populations of variable size. *Genetics* 186: 601–611.
- Ewens, W., 1982 On the concept of the effective population size. *Theor. Popul. Biol.* 21: 373–378.
- Fisher, R. A., 1999 *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford University Press, Oxford.
- Griffiths, R. C., 1981 Neutral 2-locus multiple allele models with recombination. *Theor. Popul. Biol.* 19: 169–186.
- Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38: 226–231.
- Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* 23: 183–201.
- Hudson, R. R., 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7: 1–44.
- Jagers, P., and S. Sagitov, 2004 Convergence to the coalescent in populations of substantially varying size. *J. Appl. Probab.* 41(2): 368–378.
- Johannesson, K., 2003 Evolution in littorina: ecology matters. *J. Sea Res.* 49(2): 107–117.
- Kaj, I., and S. Krone, 2003 The coalescent process in a population with stochastically varying size. *J. Appl. Probab.* 40: 33–48.
- Kingman, J. F. C., 1982 The coalescent. *Stoch. Proc. Appl.* 13: 235–248.
- Liu, H., F. Prugnolle, A. Manica, and F. Balloux, 2006 A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* 79(2): 230–237.
- Marjoram, P., and J. Wall, 2006 Fast coalescent simulation. *BMC Genet.* 7: 16.
- McPeck, M. S., and T. P. Speed, 1995 Modelling interference in genetic recombination. *Genetics* 139: 1031–1044.
- McVean, G., 2002 A genealogical interpretation of linkage disequilibrium. *Genetics* 162: 987–991.
- McVean, G., and N. Cardin, 2005 Approximating the coalescent with recombination. *Philos. Trans. R. Soc. B* 360: 1387–1393.
- Möhle, M., and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* 29: 1547–1562.
- Nordborg, M., and S. Krone, 2002 *Modern Developments in Population Genetics: The Legacy of Gustave Malécot*, pp. 194–232. Oxford University Press, Oxford.
- Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 68: 571–580.
- Pujolar, J., S. Vincenzi, L. Zane, D. Jesensek, G. De Leo *et al.*, 2011 The effect of recurrent floods on genetic composition of marble trout populations. *PLoS ONE* 6(9): e23822.
- Ramachandran, S., O. Deshpande, C. Roseman, N. Rosenberg, M. Feldman *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. USA* 102(44): 15942–15947.
- Sabatti, C., and N. Risch, 2002 Homozygosity and linkage disequilibrium. *Genetics* 160(4): 1707–1719.
- Sagitov, S., 2003 Convergence to the coalescent with simultaneous multiple mergers. *J. Appl. Probab.* 40: 839–854.
- Sargsyan, O., and J. Wakeley, 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popul. Biol.* 74(1): 104–114.
- Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* 5: 1–50.
- Sjödin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg, 2005 On the meaning and existence of an effective population size. *Genetics* 169: 1061–1070.
- Tanabe, K., T. Mita, T. Jombart, A. Eriksson, S. Horibe *et al.*, 2010 *Plasmodium falciparum* accompanied the human expansion out of Africa. *Curr. Biol.* 20(14): 1283–1289.
- Wakeley, J., and O. Sargsyan, 2009 Extensions of the coalescent effective population size. *Genetics* 181: 341–345.
- Wright, S., 1931 Evolution in Mendelian populations. *Genetics* 16: 97–159.
- Wright, S., 1938 Size of a population and breeding structure in relation to evolution. *Science* 87: 430–431.
- Zivkovic, D., and T. Wiehe, 2008 Second-order moments of segregating sites under variable population size. *Genetics* 180: 341–357.

Communicating editor: H. G. Spencer

Appendix A: Coefficients Appearing in Equations 9 and 10

In this *Appendix* we list the coefficients appearing in Equations 9 and 10.

The coefficients in Equation 9 are

$$\begin{aligned}
C_0 &= 36x^5(\lambda_B + \lambda + 3)(\lambda_B + \lambda + 6)(\lambda_B(\lambda_B(\lambda_B + 2x\lambda + \lambda + 4) \\
&\quad + \lambda(x((x + 2)\lambda + x + 6) + 1) + 5) + x\lambda(x(\lambda + 1)(\lambda + 3) + 2) + 2), \\
C_1 &= 2x^5(\lambda_B(\lambda_B(\lambda_B(\lambda_B(\lambda_B + 3x(\lambda + 9) + 2\lambda + 13) + (3x(x + 2) + 1)\lambda^2 \\
&\quad + (x(55x + 76) + 29)\lambda + 252x + 59) + x^3\lambda(\lambda + 1)(\lambda + 27) \\
&\quad + 2x^2\lambda(\lambda(3\lambda + 58) + 213) + x(\lambda(\lambda(3\lambda + 80) + 370) + 801) + 2\lambda(8\lambda + 55) + 119) \\
&\quad + x^3\lambda(\lambda + 1)(\lambda + 21)(2\lambda + 9) + x^2\lambda(\lambda(\lambda(3\lambda + 76) + 452) + 903) \\
&\quad + x(\lambda(\lambda(31\lambda + 226) + 647) + 1044) + (3\lambda + 4)(5\lambda + 27)) \\
&\quad + x(\lambda(x^2(\lambda + 1)(\lambda + 3)(\lambda + 6)(\lambda + 15) + x(5\lambda + 18)(\lambda(3\lambda + 16) + 33) \\
&\quad + 44\lambda + 270) + 468) + 18(\lambda + 2)), \\
C_2 &= x^5(\lambda_B(\lambda_B(x\lambda_B(3\lambda_B + 2x(4\lambda + 9) + 4(\lambda + 7)) + x(x^2\lambda(7\lambda + 39) \\
&\quad + 10x(\lambda(\lambda + 7) + 9) + \lambda(\lambda + 25) + 89) + 4\lambda) \\
&\quad + x(2x^3\lambda(\lambda + 1)(\lambda + 9) + x^2\lambda(\lambda(8\lambda + 65) + 129) \\
&\quad + 2^x(\lambda + 2)^2(\lambda + 18) + \lambda(9\lambda + 55) + 116) + 4\lambda) \\
&\quad + x(x(\lambda(2x^2(\lambda + 1)(\lambda + 3)(\lambda + 6) + x(\lambda(\lambda(\lambda + 22) + 83) + 102) \\
&\quad + 4\lambda^2 + 42\lambda + 96) + 72) + 26(\lambda + 2))), \\
C_3 &= x^7(\lambda_B + \lambda + 2)(\lambda_B(\lambda_B + 2x\lambda + 3) + x\lambda(x\lambda + x + 2) + 2), \\
D_0 &= 36x^5(\lambda_B + \lambda + 1)^2(\lambda_B + \lambda + 2)(\lambda_B + \lambda + 3)(\lambda_B + \lambda + 6), \\
D_1 &= 2x^5(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)(\lambda_B(\lambda_B(13\lambda_B + 13(x + 2)\lambda + 27x + 130) \\
&\quad + 13(2x + 1)\lambda^2 + 157(x + 1)\lambda + 9(19x + 39)) \\
&\quad + 13x(\lambda + 1)(\lambda + 3)(\lambda + 6) + 9(\lambda + 2)(3\lambda + 13)), \\
D_2 &= x^5(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)(\lambda_B(\lambda_B(2\lambda_B + 4x\lambda + 39x + 2\lambda + 20) \\
&\quad + x(2x(\lambda(\lambda + 8) + 9) + 4\lambda^2 + 86\lambda + 247) + 16\lambda + 54) \\
&\quad + 2x^2(\lambda + 1)(\lambda + 3)(\lambda + 6) + 13x(\lambda + 2)(3\lambda + 13) + 18(\lambda + 2)), \\
D_3 &= x^6(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)^2(3\lambda_B + x(3\lambda + 13) + 13), \\
D_4 &= x^7(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)^2.
\end{aligned} \tag{A1}$$

$$\begin{aligned}
D_0 &= 36x^5(\lambda_B + \lambda + 1)^2(\lambda_B + \lambda + 2)(\lambda_B + \lambda + 3)(\lambda_B + \lambda + 6), \\
D_1 &= 2x^5(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)(\lambda_B(\lambda_B(13\lambda_B + 13(x + 2)\lambda + 27x + 130) \\
&\quad + 13(2x + 1)\lambda^2 + 157(x + 1)\lambda + 9(19x + 39)) \\
&\quad + 13x(\lambda + 1)(\lambda + 3)(\lambda + 6) + 9(\lambda + 2)(3\lambda + 13)), \\
D_2 &= x^5(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)(\lambda_B(\lambda_B(2\lambda_B + 4x\lambda + 39x + 2\lambda + 20) \\
&\quad + x(2x(\lambda(\lambda + 8) + 9) + 4\lambda^2 + 86\lambda + 247) + 16\lambda + 54) \\
&\quad + 2x^2(\lambda + 1)(\lambda + 3)(\lambda + 6) + 13x(\lambda + 2)(3\lambda + 13) + 18(\lambda + 2)), \\
D_3 &= x^6(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)^2(3\lambda_B + x(3\lambda + 13) + 13), \\
D_4 &= x^7(\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)^2.
\end{aligned} \tag{A2}$$

The coefficients in Equation 10 are

$$\begin{aligned}
A_0 &= 18(\lambda_B + 1)(\lambda_B + \lambda + 3)(\lambda_B + \lambda + 6)(\lambda_B(\lambda_B + \lambda + 3) + 2), \\
A_1 &= (\lambda_B + 1)(\lambda_B(\lambda_B(\lambda_B^2 + 2(\lambda + 6)\lambda_B + \lambda(\lambda + 27) + 47) + \lambda(15\lambda + 83) + 72) + 18(\lambda + 2)), \\
A_2 &= 2\lambda_B(\lambda_B + 1)\lambda, \\
B_0 &= 18(\lambda_B + \lambda + 1)^2(\lambda_B + \lambda + 2)(\lambda_B + \lambda + 3)(\lambda_B + \lambda + 6), \\
B_1 &= (\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)(\lambda_B(13\lambda_B(\lambda_B + 2(\lambda + 5)) + \lambda(13\lambda + 157) + 351) + 9(\lambda + 2)(3\lambda + 13)), \\
B_2 &= (\lambda_B + \lambda + 1)(\lambda_B + \lambda + 2)(\lambda_B(\lambda_B(\lambda_B + \lambda + 10) + 8\lambda + 27) + 9(\lambda + 2)).
\end{aligned} \tag{A3}$$

Appendix B: Severe Reductions of Population Size During Bottlenecks: Connection to the Xi coalescent

In this *Appendix*, we turn our attention to the case of severe reductions of population size during bottlenecks, the third case discussed in *Covariance of the Times to the MRCA*. Formally, we describe this case by the following limit: $N_0 \rightarrow \infty$, $N_B \rightarrow \infty$, $N_B/N_0 \rightarrow 0$, while $\lambda = pN_0$ and $\lambda_B = qN_B$ are kept constant. This limit implies the following. First, it is possible that a bottleneck may host more than one coalescence. How frequently this occurs is determined by the rates λ , and λ_B . We remark that the probability for two lines to coalesce between two successive bottlenecks is $(1 + \lambda)^{-1}$ and the probability for two lines to coalesce within a single bottleneck is $(1 + \lambda_B)^{-1}$. Second, this limit implies that the duration of a single bottleneck is negligible compared to the time between two successive bottlenecks, because in this limit one has $p/q \rightarrow 0$.

This not only allows for neglecting the durations of bottlenecks, but also implies that possible coalescences within a single bottleneck can be considered as a single simultaneous multiple merger. This allows for employing the Xi-coalescent approximation in this case (as was also suggested by Birkner *et al.* 2009). The corresponding coalescent rates for our Xi-coalescent approximation show, in terms of the parameters λ and λ_B , how important multiple mergers are for shaping the resulting gene genealogies in this case.

In what follows we demonstrate how the Xi-coalescent approximation yields Equation 10 for the mean covariance $\langle \text{cov}_D[t_{a(ij)}, t_{b(ij)}] \rangle$. Our method for calculating the term $\langle t_{a(ij)} t_{b(ij)} \rangle$ is described in the main text (*Covariance of the Times to the MRCA*). But in the Xi-coalescent approximation employed here, the Markov process differs from the one described in *Covariance of the Times to the MRCA*. The corresponding graph is shown in Figure B1. It consists of the same five states 1, . . . , 5 shown in Figure 3, but in Figure B1 the states in bottlenecks are omitted, since the time spent in a bottleneck is short compared to the time between two successive bottlenecks.

The remainder of this *Appendix* is organized as follows. First, we show how the transition rates w_{mn} are derived using the Xi-coalescent approximation. Second, using the Xi-coalescent transition rates we demonstrate how to derive Equation 10.

Formulas for w_{mn} Under the Xi-Coalescent Approximation

Assume that l ancestral lines enter a bottleneck, and that b lines leave this bottleneck. In our Xi-coalescent approximation, we assume that the $l - b$ coalescent events in the bottleneck happen instantaneously.

In what follows we derive the coalescent rates for our Xi-coalescent approximation. Let l lines be partitioned into b families ($b \leq l$), such that k_i families are of sizes $i = 1, \dots, l$. By construction, the following conditions must be satisfied:

$$l = \sum_{i=1}^l i k_i, \quad b = \sum_{i=1}^l k_i. \quad (\text{B1})$$

In our model, the collision rate $\phi_{\{l; k_1, \dots, k_l\}}$ of l lines colliding into a particular partition $\{l; k_1, \dots, k_l\}$, such that Equation B1 is satisfied, is given by

$$\phi_{\{l; k_1, \dots, k_l\}} = 1_{\{b=l-1\}} + \lambda \Xi_{\{l; k_1, \dots, k_l\}}. \quad (\text{B2})$$

Here the first term stands for the Kingman coalescent outside bottlenecks, and the second term corresponds to the contribution from (multiple) coalescences during bottlenecks (multiple mergers are obtained in the case $b < l-1$). Bottlenecks occur at a rate λ . Given the probability C_{lb} that during a bottleneck l lines collide into b lines, $\Xi_{\{l; k_1, \dots, k_l\}}$ can be calculated according to

$$\Xi_{\{l; k_1, \dots, k_l\}} = C_{lb} P_{\{l; k_1, \dots, k_l\}}, \quad (\text{B3})$$

where

$$C_{lb} = \frac{\lambda_B}{\binom{b}{2} + \lambda_B} \prod_{i=b+1}^l \frac{\binom{i}{2} + \lambda_B}{\binom{i}{2} + \lambda_B}, \quad (\text{B4})$$

and $P_{\{l; k_1, \dots, k_l\}}$ is the probability of observing a particular partition $\{l; k_1, \dots, k_l\}$ of l lines. As shown by Kingman (1982), it is given by

$$P_{\{l; k_1, \dots, k_l\}} = \frac{(l-b)! b! (b-1)!}{l! (l-1)!} \prod_{i=1}^l (i!)^{k_i}. \quad (\text{B5})$$

The rate $\phi_{\{l; k_1, \dots, k_l\}}$ appearing in Equation B2 is conditional on a particular partition. Thus the total collision rate of l lines into any of partitions of type $\{l; k_1, \dots, k_l\}$ is given by

$$\phi_{\{l; k_1, \dots, k_l\}}^{\text{tot}} = \binom{l}{2} 1_{\{b=l-1\}} + \lambda C_{lb} P_{\{l; k_1, \dots, k_l\}} S_{\{l; k_1, \dots, k_l\}}. \quad (\text{B6})$$

Here

$$S_{\{l; k_1, \dots, k_l\}} = \frac{l!}{\prod_{i=1}^l (i!)^{k_i} k_i!} \quad (\text{B7})$$

denotes the number of possible ways of collisions of l lines into a partition $\{l; k_1, \dots, k_l\}$, such that restrictions in Equation B1 hold.

The graph corresponding to the Markov process in the limit described in this *Appendix* consists of five states $1, \dots, 5$ (see Figure B1). We now show how the corresponding transition rates between the states $1, \dots, 5$ can be derived from Equation B6.

We observe that a collision of type $\{2; 0, 1\}$ describes a transition from either state 1 or 4 to 5. It follows that

$$w_{51} = w_{54} = \phi_{\{2;0,1\}}^{\text{tot}}. \quad (\text{B8})$$

State 2 consists of three lines. A collision of one particular pair of lines, among the three lines, results in a transition from 2 to 1, while a collision of either of the two remaining pairs of lines results in a transition from 2 to 4. Because a collision of a pair of lines among three lines is of type $\{3; 1, 1, 0\}$, we obtain the following transition rates:

$$w_{12} = \frac{1}{3} \phi_{\{3;1,1,0\}}^{\text{tot}}, \quad w_{42} = \frac{2}{3} \phi_{\{3;1,1,0\}}^{\text{tot}}. \quad (\text{B9})$$

A collision of all three lines of state 2 leads to a transition from state 2 to 5 at the rate

$$w_{52} = \phi_{\{3;0,0,1\}}^{\text{tot}}. \quad (\text{B10})$$

Now consider transitions from state 3, consisting of four ancestral lines. We analyze first a collision of a single pair of lines, that is, a collision of type $\{4; 2, 1, 0, 0\}$. There are in total six different ways to pair the four lines entering the bottleneck: four choices describe a transition from state 3 to 2, and the remaining two lead to a transition from 3 to 4. Thus, we have

$$w_{23} = \frac{2}{3} \phi_{\{4;2,1,0,0\}}^{\text{tot}}. \quad (\text{B11})$$

Further, there are three possibilities for simultaneous collisions of two pairs of lines. Two possibilities result in a transition from 3 to 1, and one leads to a transition from 3 to 5 (see Figure 2D). It is also possible to obtain a collision of three lines, in which case the transition from 3 to 4 is obtained. Further, a collision of all four lines results in a transition from 3 to 5. Thus, we obtain the following transition rates:

$$w_{13} = \frac{2}{3} \phi_{\{4;0,2,0,0\}}^{\text{tot}}, \quad w_{43} = \frac{1}{3} \phi_{\{4;2,1,0,0\}}^{\text{tot}} + \phi_{\{4;1,0,1,0\}}^{\text{tot}}, \quad w_{53} = \frac{1}{3} \phi_{\{4;0,2,0,0\}}^{\text{tot}} + \phi_{\{4;0,0,0,1\}}^{\text{tot}}. \quad (\text{B12})$$

The remaining nonvanishing rates, $w_{21} = R$, and $w_{32} = R/2$, describe recombination transitions from state 1 to 2 and from 2 to 3. This shows how the transition rates w_{mn} are expressed in terms of the collision rates of the Xi coalescent. Explicit formulas for the rates w_{mn} in terms of λ and λ_B are given in Figure B1.

Obtaining Equation 10 Under the Xi-Coalescent Approximation

Given the rates w_{mn} , the mean covariance $\langle \text{cov}_{\mathcal{D}}[t_{a(ij)}, t_{b(ij)}] \rangle$ is computed from

$$\langle t_{a(ij)} t_{b(ij)} \rangle = \int_0^\infty dt_1 t_1^2 \mathbf{u} e^{\mathbf{M} t_1} \mathbf{v} + \int_0^\infty dt_1 \int_{t_1}^\infty dt_2 t_1 t_2 c e^{K(t_2 - t_1)} \mathbf{Q} e^{\mathbf{M} t_1} \mathbf{v}. \quad (\text{B13})$$

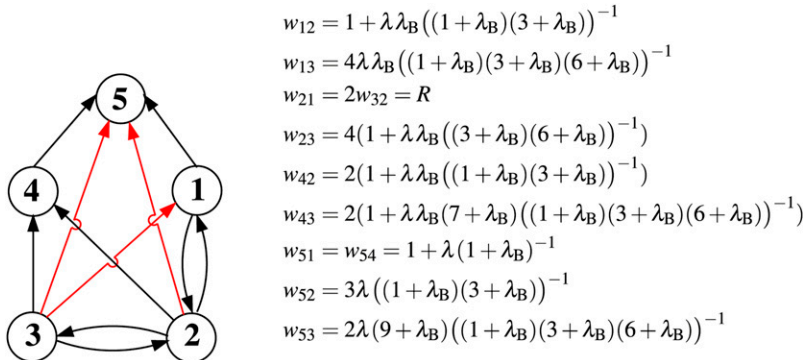


Figure B1 (Left) Graph showing the states and possible transitions corresponding to the limit of severe reductions of population size during bottlenecks. The states $1, \dots, 5$ are explained in Figure 3. The three red arrows in this graph correspond to simultaneous multiple mergers. (Right) Transition rates, w_{mn} , from state n to m , in terms of the parameters $\lambda = \rho N_0$ and $\lambda_B = q N_B$.

This equation corresponds to Equation 7 given in *Covariance of the Times to the MRCA*. But the dimensions of the matrices and vectors in Equation B13 differ from those in Equation 7 because states in the bottleneck are omitted. Here \mathbf{M} is a 3×3 matrix. Its off-diagonal elements correspond to the rates w_{mn} :

$$\begin{aligned} M_{12} &= 1 + \lambda\lambda_B((1 + \lambda_B)(3 + \lambda_B))^{-1}, \\ M_{13} &= 4\lambda\lambda_B((1 + \lambda_B)(3 + \lambda_B)(6 + \lambda_B))^{-1}, \\ M_{21} &= 2M_{32} = R, \\ M_{23} &= 4 + 4\lambda\lambda_B((3 + \lambda_B)(6 + \lambda_B))^{-1}. \end{aligned} \tag{B14}$$

The diagonal elements of \mathbf{M} are given by $M_{nn} = -\sum_{m \neq n} M_{mn}$ for $n = 1, 2, 3$. The remaining quantities appearing in Equation B13 are $K = -1$, $c = 1$, $\mathbf{v} = [1, 0, 0]^T$, and

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} 1 + \lambda(1 + \lambda_B)^{-1} \\ 3\lambda((1 + \lambda_B)(3 + \lambda_B))^{-1} \\ 2\lambda(9 + \lambda_B)((1 + \lambda_B)(3 + \lambda_B)(6 + \lambda_B))^{-1} \end{bmatrix}^T, \\ \mathbf{Q} &= \begin{bmatrix} 0 \\ 2(1 + \lambda\lambda_B((1 + \lambda_B)(3 + \lambda_B))^{-1}) \\ 2(1 + \lambda\lambda_B(7 + \lambda_B)((1 + \lambda_B)(3 + \lambda_B)(6 + \lambda_B))^{-1}) \end{bmatrix}^T. \end{aligned} \tag{B15}$$

Combining Equation B13 with Equations B14 and B15 results in Equation 10. This shows that the covariance of the times to the MRCA in the case of severe reductions of population size can be derived within the Xi-coalescent approximation.