Genetics: Published Articles Ahead of Print, published on November 2, 2011 as 10.1534/genetics.111.134437

Linkage disequilibrium under recurrent bottlenecks

E. Schaper[a,*], A. Eriksson[b], M.Rafajlovic[a], S. Sagitov[c], and B. Mehlig[a]
[a]*Department of Physics, University of Gothenburg, SE-41296 Gothenburg, Sweden*
[b]*Department of Zoology, University of Cambridge, Cambridge, UK*
[c]*Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-41296 Gothenburg, Sweden*

In order 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 multi-locus 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 analyse 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.

Keywords: Recurrent bottlenecks, linkage disequilibrium, recombination, gene genealogies, Kingman's coalescent, Xi-coalescent

I. INTRODUCTION

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 time scales, 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 (Liu et al., 2006; Ramachandran et al., 2005), the accompanying expansion of the parasite *Plasmodicum 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 strenghts. 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 time scales in the context of such effective population-size approximations: the time scale of the population-size fluctuations, and the coalescent time scale (which reflects the time to the most recent common ancestor, MRCA). When demographic fluctuations are much slower than the coalescent time scale, 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 (Crow and Kimura, 1970; Wright, 1938) that genetic variation is well described in terms of a population with effective population size $N_{\rm eff}$, given by the harmonic average of the

* Present address: *Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland* and *Computer Science Department, Swiss Federal Institute of Technology Zurich, CH-8092 Zürich, Switzerland*

Copyright 2011.

population size:

$$N_{\text{eff}} = \lim_{T\to\infty}\Big(\frac{1}{T}\sum_{\tau=0}^{T}\frac{1}{N_\tau}\Big)^{-1}. \tag{1}$$

Here N_τ is the population size in generation τ. See Sjödin et al. (2005), Jagers and Sagitov (2004) 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 time scales are of the same order, it has been shown (Eriksson et al., 2010; Kaj and Krone, 2003; Nordborg and Krone, 2003; Sjödin et al., 2005) 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 multi-locus patterns of genetic variation on the same chromosome? Such patterns are influenced by recombination. Genetic recombination introduces a new time scale which 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 multi-locus patterns of genetic variation in biological populations. Measures of linkage disequilibrium quantify the degree of association of genetic variation at pairs of loci on the same chromosome. Common measures of linkage disequilibrium, such as $\hat{r}^2$ (Hill and Robertson, 1968), and its approximation σ_d^2 (McVean, 2002; Ohta and Kimura, 1971), 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).

Fig. 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 Section II. In Fig. 1, grey 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 grey line corresponds to a single realisation of the sequence of bottlenecks. The red lines in Fig. 1 are the averages of the covariances within each panel. Panels **a** and **b** show results for two different sets of parameters of the model. In panel **a**, 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 Eq. (1). For such populations, the coalescent approximation predicts (Griffiths, 1981; Hudson, 1983, 1990)

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

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 Eq. (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 Fig. 1 effective population-size approximations, according to Eq. (2), are shown as dashed lines. In Fig. 1**a** we observe good agreement between Eq. (2) and the average covariance of the times to the MRCA obtained from computer simulations of the model. However, in Fig. 1**b**, Eq. (2) agrees with the simulated covariance only for short genetic distances. For large genetic distances, by contrast, Fig. 1**b** shows that the covariance decreases much more slowly than expected according to Eq. (2). Thus, the results shown in Fig. 1**b** imply long-range association between two loci.

The examples shown in Fig. 1 raise many questions. What are the conditions for the effective population-size approximation to be valid in the multi-locus 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 paper 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 Section II. Our analysis enables us to qualitatively and quantitatively determine the effects of fluctuating population size on the two-locus statistics in terms of the time scales of population-size fluctuations, of coalescence, and of recombination. Using both analytical (Section III), and numerical approaches (Section IV), 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 Fig. 1**a**) and sometimes significant (as in the case shown in Fig. 1**b**). 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 Section V that the long-

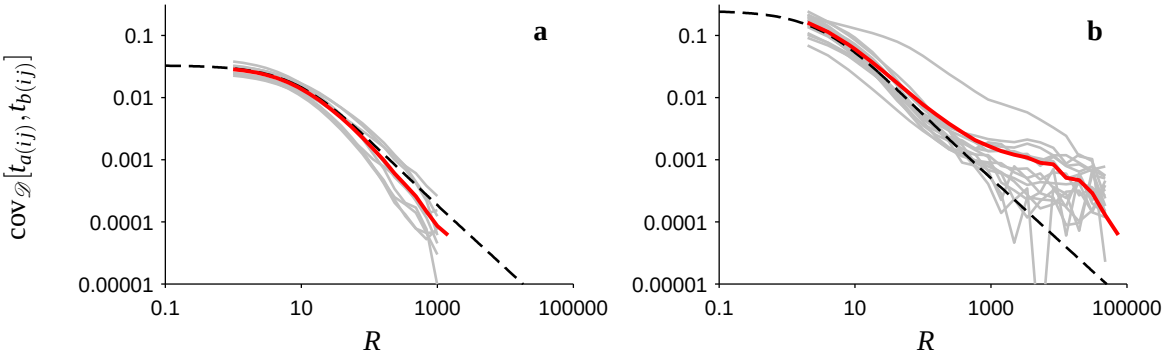


FIG. 1 Covariance of the times to the MRCA at two loci, in a sample of two chromosomes in a population subject to repeated bottlenecks (details in Section II). (**a**) Rapid population-size fluctuations. Wright-Fisher simulations for ten random sequences of bottlenecks with $p = 10^{-3}$, $q = 10^{-3}$, $N_0 = 10^5$, and $N_B = 10^4$ (grey lines). Each grey 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, Eq. (2). (**b**) Same, but for the case of severe reductions of population size during bottlenecks. Wright-Fisher simulations for fifteen randomly generated sequences of bottlenecks, with parameters $p = 10^{-5}$, $q = 2 \cdot 10^{-2}$, $N_0 = 10^6$, and $N_B = 5 \cdot 10^2$. Averages are over 100 gene genealogies for each demographic history.

range association has a surprisingly small effect upon σ_d^2. In Section VI we discuss 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 (Möhle and Sagitov, 2001; Sagitov, 2003; Schweinsberg, 2000), which allows for simultaneous multiple coalescent events (called multiple mergers below).

In summary we describe in this paper 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 analysed 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 Section VI (see Eq. (14)).

II. MODEL

Using a Wright-Fisher model of a population of gametes, we trace the ancestry of two loci on a pair of gametes backwards 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 McPeek 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 ancestry 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 Eriksson et al. (2010); Sjödin et al. (2005), see also Sargsyan and Wakeley (2008).

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

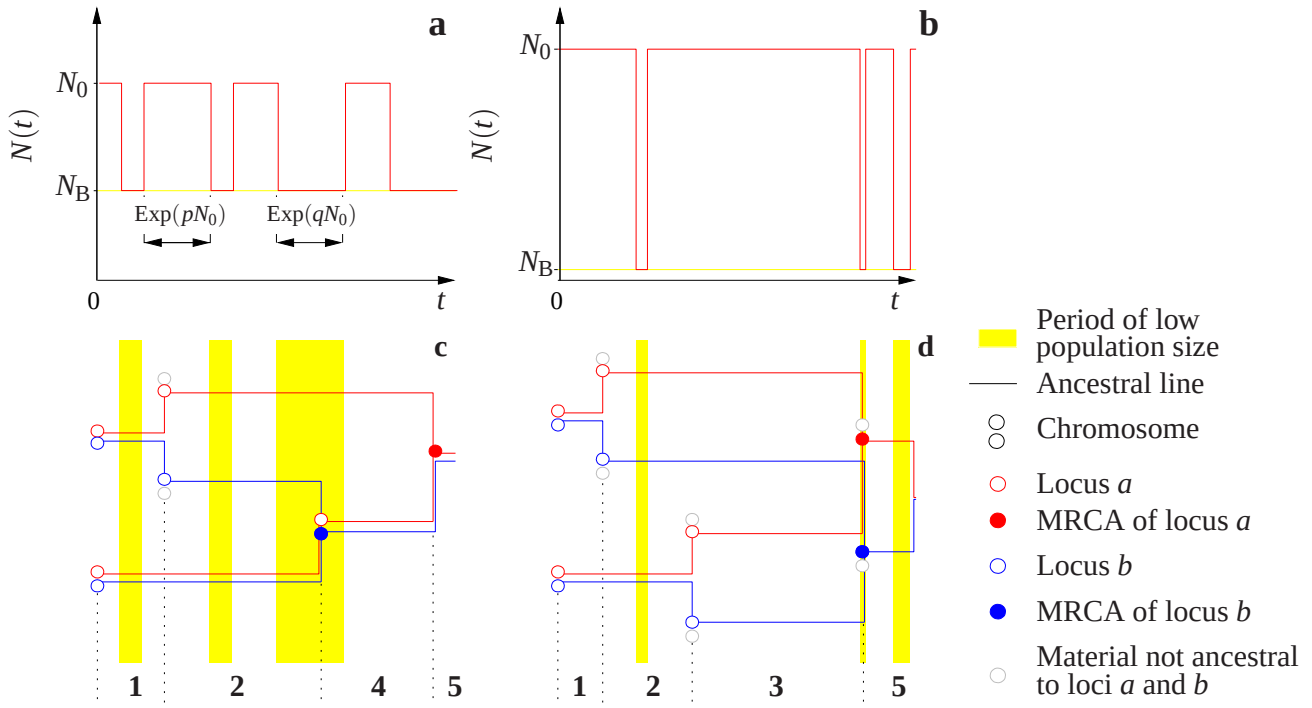


FIG. 2 Panels **a** and **b** show two realisations of the population-size curve, $N(t)$, backwards in time ($t = 0$ denotes the present time). Initially, the population size is N_0. Going backwards in time, the population size randomly jumps between two values, N_0 and N_B, with transition rates pN_0 (from N_0 to N_B) and qN_0 (from N_B to N_0). Panels **c** and **d** show examples of ancestral histories of two loci (blue and red empty 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 $1,\ldots,5$ denote the possible states of the system (they are explained in detail in Fig. 3). Grey circles denote genetic material not ancestral to the sampled loci. Blue and red filled 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 time scale of the gene genealogy. An example is shown in panel **d**: an almost instantaneous transition from state 3 to state 5.

recombination between two loci occurs, the chromosomes are represented by circles instead of lines. MRCA are shown as filled circles. In some cases recombination causes the ancestry of one locus to become associated with a chromosome that lacks direct descendants in the sample (grey 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.

III. COVARIANCE OF THE TIMES TO THE MRCA

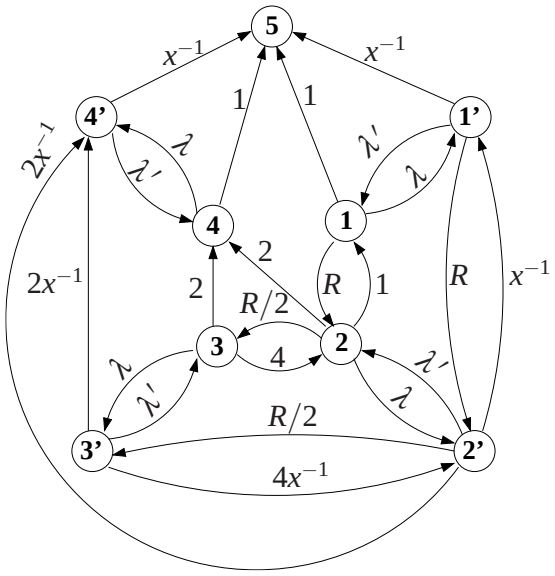
Consider a sample of two chromosomes (denoted by i and j), and two loci (called a and b). As in the introduction, the time to the MRCA at locus a is denoted by $t_{a(ij)}$ (and $t_{b(ij)}$ at locus b). Consider $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ij)}]$, the covariance conditional on a particular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

$$\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ij)}] \rangle = \langle \langle t_{a(ij)} t_{b(ij)} | \mathscr{D} \rangle - \langle t_{a(ij)} | \mathscr{D} \rangle \langle t_{b(ij)} | \mathscr{D} \rangle \rangle = \langle t_{a(ij)} t_{b(ij)} \rangle - \langle \langle t_{a(ij)} | \mathscr{D} \rangle^2 \rangle \,. \tag{3}$$

Here $\langle \ldots | \mathscr{D} \rangle$ denotes the expectation conditional on a particular demographic history $\mathscr{D}$. 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:

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

We remark that biological data correspond to a particular individual realisation $\mathscr{D}$. The additional term in Eq. (4) reflects the variance over different demographic histories of the average time to the MRCA at a single locus, and is thus irrelevant to analysing the covariance of the times to the MRCA in a particular empirical data set.



group	state
$\mathbf{1},\mathbf{1}'$	a_ib_i, a_jb_j
	a_ib_j, a_jb_i
$\mathbf{2},\mathbf{2}'$	$a_i\circ, \circ b_i, a_jb_j$
	$a_ib_i, a_j\circ, \circ b_j$
	$a_i\circ, \circ b_j, a_jb_i$
	$a_ib_j, a_j\circ, \circ b_i$
$\mathbf{3},\mathbf{3}'$	$a_i\circ, \circ b_i, a_j\circ, \circ b_j$
$\mathbf{4},\mathbf{4}'$	$a_i\bullet, a_j\bullet$
	$\bullet b_i, \bullet b_j$
$\mathbf{5}$	$\bullet\bullet$

FIG. 3 Left: a 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 Section II. 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: a table of possible states of the system. The two loci are denoted by a and b, and the corresponding chromosomes by i and j. Empty circles denote genetic material not ancestral to sampled loci, and full circles denote the MRCA.

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

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

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 $\tau = 0, 1, 2, \ldots$ 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 summarised above, we find the following expression for the term $\langle\langle t_{a(ij)}|\mathscr{D}\rangle\langle t_{b(ij)}|\mathscr{D}\rangle\rangle \equiv \langle\langle t_{a(ij)}|\mathscr{D}\rangle^2\rangle$ occurring in Eq. (3):

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

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

In order to evaluate the remaining term in Eq. (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 Fig. 2**c**, **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 Fig. 3, where the vertices represent states (combinations of ancestral lines), and the edges represent transitions between the states (the transition rates w_{mn} from n to m are shown along the edges from n to m). 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 sub-graph of the graph shown in Fig. 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 (Eriksson and Mehlig, 2004)

$$\langle t_{a(ij)} t_{b(ij)} \rangle = \int_0^\infty \mathrm{d}t_1\, t_1^2 \boldsymbol{u}\, \mathrm{e}^{\mathbf{M}t_1} \boldsymbol{v} + \int_0^\infty \mathrm{d}t_1 \int_{t_1}^\infty \mathrm{d}t_2\, t_1 t_2 \boldsymbol{c}\, \mathrm{e}^{\mathbf{K}(t_2-t_1)}\, \mathbf{Q}\, \mathrm{e}^{\mathbf{M}t_1} \boldsymbol{v}\,. \tag{7}$$

The vectors $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{c}$ and the matrices $\mathbf{Q}$ and $\mathbf{K}$ in Eq. (7) are given by:

$$\boldsymbol{v} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^\mathsf{T}, \quad \boldsymbol{u} = \begin{bmatrix} 1 & 0 & 0 & x^{-1} & 0 & 0 \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} 1 & x^{-1} \end{bmatrix},$$
$$\mathbf{Q} = \begin{bmatrix} 0 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2x^{-1} & 2x^{-1} \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} -(\lambda+1) & \lambda_\mathrm{B} x^{-1} \\ \lambda & -(\lambda_\mathrm{B}+1)x^{-1} \end{bmatrix}.$$

Here T 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 Eq. (7) corresponds to a common MRCA of the loci a and b (i. e. to a transition from states 1 or $1'$ to state 5). The second term in Eq. (7) corresponds to different MRCA (transitions from states 2 or 3 to state 4, or from states $2'$ 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 $\boldsymbol{v}$. The elements of the vector $\boldsymbol{u}$ correspond to the transition rates from the states $1, 2, 3, 1', 2', 3'$ to 5. The two rows in the matrix $\mathbf{Q}$ contain the transition rates from $1, 2, 3, 1', 2', 3'$ to 4 and $4'$, respectively. The matrix $\mathbf{K}$ contains the transition rates between states 4 and $4'$, whereas the vector $\boldsymbol{c}$ contains the coalescent rates in each population-size regime. Both $\mathbf{M}$ and $\mathbf{K}$ have negative real eigenvalues. Hence, the integrals in Eq. (7) can be evaluated in terms of matrix inverses (Eriksson and Mehlig, 2004). We find:

$$\langle t_{a(ij)} t_{b(ij)} \rangle = 2\boldsymbol{u}(-\mathbf{M})^{-3}\boldsymbol{v} + 2\boldsymbol{c}(-\mathbf{K})^{-2}\left\{(-\mathbf{K})^{-1}\mathbf{Q} + \mathbf{Q}(-\mathbf{M})^{-1}\right\}(-\mathbf{M})^{-2}\boldsymbol{v}\,. \tag{8}$$

Combining Eqs. (6) and (8) yields

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

The coefficients $C_0, \ldots, C_3$, and $D_0, \ldots, D_4$, 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 analyse 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 Eq. (9) by taking the limit $\lambda \to 0$. In this limit we obtain Eq. (2) with $x_\mathrm{eff} = 1$, which is the expression valid in the case of constant population size.

The second case is described by the conditions $p \gg N_0^{-1}$ and $q \gg N_\mathrm{B}^{-1}$. Formally, this case corresponds to taking the limit $\lambda \to \infty$, and $\lambda_\mathrm{B} \to \infty$ in Eq. (9), in such a way that the ratio $\lambda/\lambda_\mathrm{B}$ is kept constant. In this limit we obtain Eq. (2), with $x_\mathrm{eff} = (x + \lambda_\mathrm{B}/\lambda)/(1 + \lambda_\mathrm{B}/\lambda)$. 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 $N_\mathrm{B} \ll N_0$. 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 (Liu et al., 2006; Ramachandran et al., 2005), the accompanying expansion of the parasite *Plasmodicum 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)). This case can be treated analytically by taking the limit $x \to 0$ in Eq. (9), keeping λ, and λ_B fixed. In this limit we find:

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

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 Eq. (10) reaches a plateau for large values of R:

$$\langle \mathrm{cov}_\mathscr{D}[t_{a(ij)}, t_{b(ij)}] \rangle \approx \frac{2\lambda_\mathrm{B}(1+\lambda_\mathrm{B})\lambda}{(1+\lambda_\mathrm{B}+\lambda)(2+\lambda_\mathrm{B}+\lambda)(9(2+\lambda)+\lambda_\mathrm{B}(27+8\lambda+\lambda_\mathrm{B}(10+\lambda_\mathrm{B}+\lambda)))}\,. \tag{11}$$

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

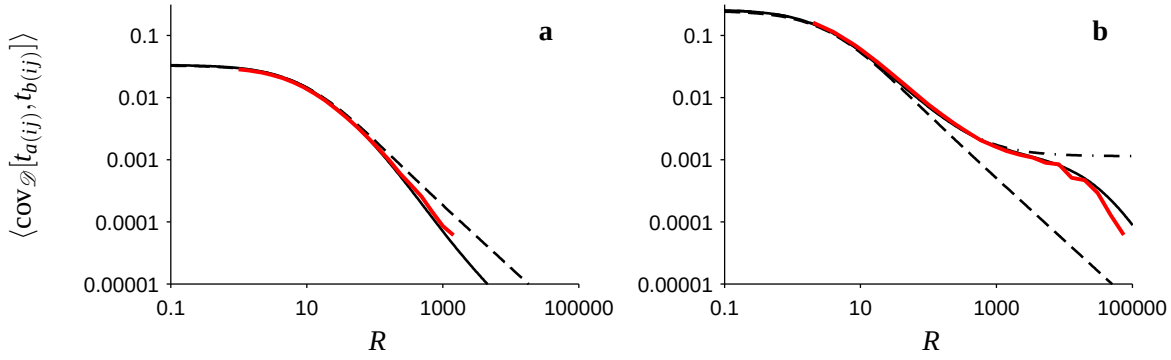


FIG. 4 Covariance of the times to the MRCA of two loci averaged over random population-size histories. (**a**) The red line shows the average covariance corresponding to $p = 10^{-3}$, $q = 10^{-3}$, $N_0 = 10^5$, and $N_B = 10^4$, determined numerically from Wright-Fisher simulations (same as in Fig.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

IV. COMPARISON OF THE COALESCENT CALCULATIONS TO THE WRIGHT-FISHER SIMULATIONS

In order to further illustrate the role of the relevant time scales of genetic drift and recombination in shaping genetic association between two loci, we compare the full coalescent result, Eq. (9), and the different limiting cases considered in Section III, to the average covariance calculated from the Wright-Fisher simulations. These comparisons are shown in Fig. 4.

The parameters used in Fig. 4**a** correspond to rapid population-size fluctuations (the second case described in the previous section). As can be seen in Fig. 4**a**, the agreement between the numerical result (red line) and the approximation Eq. (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 Fig. 4**a**. 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.

Fig. 4**b** shows results for parameters corresponding to the third case analysed 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) which cannot be accurately described by the effective population-size approximation Eq. (2). The average covariance curve obtained by computer simulations (Fig. 4**b**) can be divided into three regions in which the curve behaves qualitatively differently. First, at very small recombination rates, for the parameters chosen in Fig. 4**b**, 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 Fig. 4**b**. 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, Eq. (10) agrees well with the resulting covariance in this region. We show in Appendix B that Eq. (10) can be derived using the Xi-coalescent approximation, which allows for simultaneous multiple coalescences of lines. Third, the agreement between Eq. (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 Eq. (9), visible in Fig. 4**b** at large values of R, are discussed in the concluding section.

V. 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, Eq. (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 based on a sample of n chromosomes (McVean, 2002; Ohta and Kimura, 1971). In this section, we discuss the effect of recurrent bottlenecks upon σ_d^2 in the case $n \gg 1$. In this case, Eq. (9) in McVean (2002)

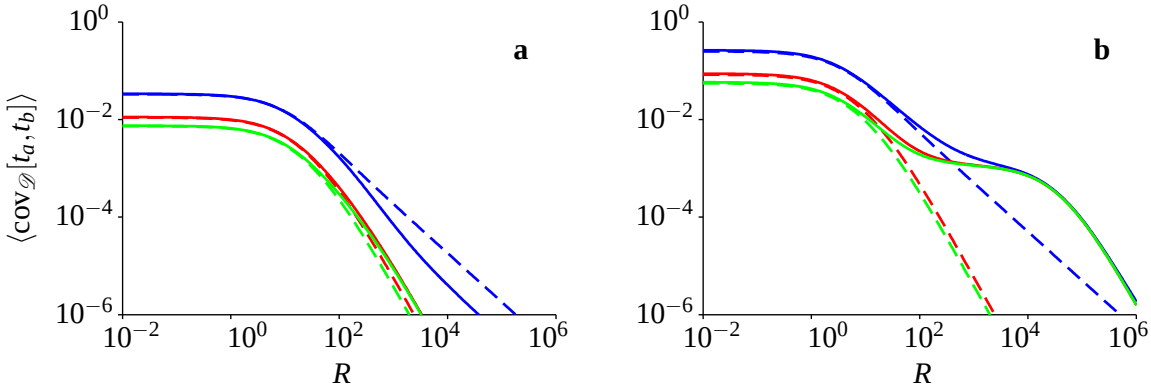


FIG. 5 Average covariances $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(ij)}]\rangle$ (blue lines), $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(ik)}]\rangle$ (red lines), and $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(kl)}]\rangle$ (green lines). In panel **a** and **b**, the values of the parameters p, q, N_0, and N_B, are the same as in Figs. 1**a** and 4**a**. The parameters in panel **b** are the same as in Figs. 1**b**, 4**b**. Exact results are shown as solid lines, whereas results obtained within the effective population-size approximation are shown as dashed lines.

becomes:

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

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). In order to simplify the analysis, we therefore focus on the expected value of the numerator:

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

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

Fig. 5 shows how the covariances $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(ij)}]\rangle$ (blue lines), $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(ik)}]\rangle$ (red lines), and $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(kl)}]\rangle$ (green lines) depend on R. This figure demonstrates that when R is small each covariance is well approximated by the corresponding effective population-size approximation. In the case shown in panel **b**, 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 Eq. (13) these enhancements cancel. We have checked that this conclusion also holds for σ_d^2 given by Eq. (12). In Fig. 6 we show for two random demographic histories how the covariances $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(ij)}]$ (blue lines), $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(ik)}]$ (red lines), and $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)},t_{b(kl)}]$ (green lines) depend on R. Two panels in Fig. 6 show 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 Figs. 4**b**, 5**b**. In both examples in Fig. 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 Eq. (13).

VI. DISCUSSION

The aim of this paper 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 Fig. 2**a**, and assuming Wright-Fisher dynamics, we have derived a coalescent result for the covariance of the times to the MRCA of two loci, Eq. (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 Eq. (1). Third, if the population size is severely reduced during bottlenecks, the results of our computer simulations depend on the relation between the time scales of recombination and of population-size changes. For the parameters chosen in Fig. 4**b**, we find that when recombination

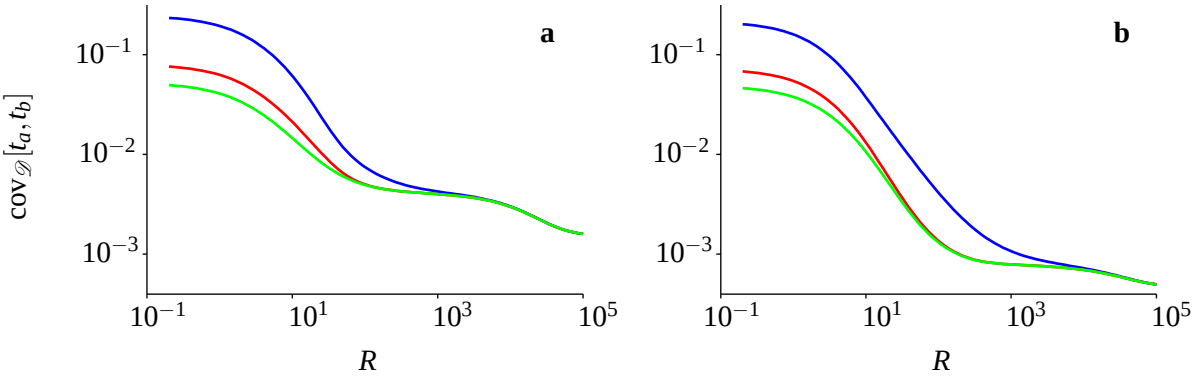


FIG. 6 Covariances $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ij)}]$ (blue lines), $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ik)}]$ (red lines), and $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(kl)}]$ (green lines) for two random realisations of population-size history. In both panels the parameters are: $p = 10^{-4}, q = 0.2, N_0 = 10^5$, and $N_{\mathrm{B}} = 50$, resulting in the same values of x, λ, and λ_{B}, as in Figs. 4**b**, 5**b**.

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 it is also approximately independent of R. In this case we say that pairs of distant loci exhibit long-range association.

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

The coalescent approximations employed in this paper 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 Fig. 1**b**. 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 Eq. (10) in terms of the so-called Xi-coalescent approximation. Xi-coalescents form a broad family of gene-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 Eq. (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, 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 analysed in this paper. 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 $x \to 0$ and $\lambda_{\mathrm{B}} \to 0$. In this limit we find (averaging numerator and denominator in Eq. (12) separately):

$$\sigma_d^2 \approx \frac{10 + R + 2\lambda}{22 + R^2 + 16\lambda + 2\lambda^2 + 13R + 3\lambda R}\,. \tag{14}$$

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 (Marjoram and Wall, 2006; McVean and Cardin, 2005).

We conclude with the observation that σ_d^2, which is a function of the covariances $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ij)}]$, $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ik)}]$, and $\mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ij)}]$, fails to show the plateaus observed in these covariances in Fig. 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 characterising 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.

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

References

Birkner, M., Blath, J., Möhle, M., Steinrücken, M., Tams, J., 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., Kimura, M., 1970. An introduction to population genetics theory. Harper & Row, London, discussion of linkage equilibrium on p. 50.

Durrett, R., Schweinsberg, J., 2004. Approximating selective sweeps. Theor. Popul. Biol. 66, 129 – 138.

Eldon, B., Wakeley, J., 2008. Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178 (3), 1517–1532.

England, P., Osler, G., Woodworth, L., Montgomery, M., Briscoe, D., Frankham, R., 2003. Effects of intense versus diffuse population bottlenecks on microsatellite genetic diversity and evolutionary potential. Conservation Genetics 4 (5), 595–604.

Eriksson, A., Fernstrom, P., Mehlig, B., Sagitov, S., 2008. An accurate model for genetic hitchhiking. Genetics 178 (1), 439–451.

Eriksson, A., Mahjani, B., Mehlig, B., 2009. Sequential markov coalescent algorithms for population models with demographic structure. Theor. Pop. Biol. 76, 84–91.

Eriksson, A., Mehlig, B., 2004. Gene-history correlation and population structure. Physical Biology 1, 220–228.

Eriksson, A., Mehlig, B., Rafajlovic, M., Sagitov, S., 2010. The total branch length of sample genealogies in populations of variable size. Genetics 186, 601–611.

Ewens, W., 1982. The concept of the effective population size. Theor. Popul. Biol. 21, 373–378.

Fisher, R. A., 1930/1999. The Genetical Theory of Natural Selection, variorum Edition. Oxford University Press.

Griffiths, R. C., 1981. Neutral 2-locus multiple allele models with recombination. Theor. Pop. Biol. 19, 169–186.

Hill, W. G., Robertson, A., 1968. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics 38, 226–231.

Hudson, R. R., 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23, 183–201.

Hudson, R. R., 1990. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7, 1–44.

Jagers, P., Sagitov, S., 2004. Convergence to the coalescent in populations of substantially varying size. Journal of Applied Probability 41 (1), 368–378.

Johannesson, K., 2003. Evolution in littorina: ecology matters. Journal of Sea Research 49 (2), 107–117.

Kaj, I., Krone, S., 2003. The coalescent process in a population with stochastically varying size. J. Appl. Prob. 40, 33–48.

Kingman, J. F. C., 1982. The coalescent. Stochastic Processes and their Applications 13, 235–248.

Liu, H., Prugnolle, F., Manica, A., Balloux, F., 2006. A geographically explicit genetic model of worldwide human-settlement history. Am. J. Hum. Genet. 79 (2), 230–237.

Marjoram, P., Wall, J., 2006. Fast coalescent simulations. BMC Genetics 7, 360.

McPeek, M. S., Speed, T. P., 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., Cardin, N., 2005. Approximating the coalescent with recombination. Phil. Trans. Roy. Soc. B 360, 1387–1393.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29, 1547–1562.

Nordborg, M., Krone, S., 2003. Modern Developments in Population Genetics: The Legacy of Gustave Malécot. Oxford University Press, Oxford, pp. 194–232.

Ohta, T., Kimura, M., 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68 (4), 571–580.

Pujolar, J., Vicenzi, S., Zane, L., Jesensek, D., De Leo, G., Crivelli, A., 2011. The effect of recurrent floods on genetic composition of marble trout populations. PLoS ONE 6 (9), e23822.

Ramachandran, S., Deshpande, O., Roseman, C., Rosenberg, N., Feldman, M., Cavalli-Sforza, L., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 102 (44), 15942–15947.

Sabatti, C., Risch, N., 2002. Homozygosity and linkage disequilibrium. Genetics 160 (4), 1707–1719.

Sagitov, S., 2003. Convergence to the coalescent with simultanous multiple mergers. J. Appl. Probab 36, 1116–1125.

Sargsyan, O., Wakeley, J., 2008. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical Population Biology 74 (1), 104 – 114.

Schweinsberg, J., 2000. Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5, 1–55.

Sjödin, P., Kaj, I., Krone, S., Lascoux, M., Nordborg, M., 2005. On the meaning and existence of an effective population size. Genetics 169 (2), 1061–1070.

Tanabe, K., Mita, T., Jombart, T., Eriksson, A., Horibe, S., Palacpac, N., Ranford-Cartwright, L., Sawai, H., Sakihama, N., Ohmae, H., Nakamura, M., Ferreira, M. U., Escalante, A. A., Prugnolle, F., Björkman, A., Färnert, A., Kaneko, A., Horii, T., Manica, A., Kishino, H., Balloux, F., 2010. Plasmodium falciparum accompanied the human expansion out of Africa. Curr Biol 20 (14), 1283–1289.

Wakeley, J., Sargsyan, O., 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., Wiehe, T., 2008. Second-order moments of segregating sites under variable population size. Genetics 180, 341–357.

Appendix A: Coefficients appearing in Eqs. (9) and (10)

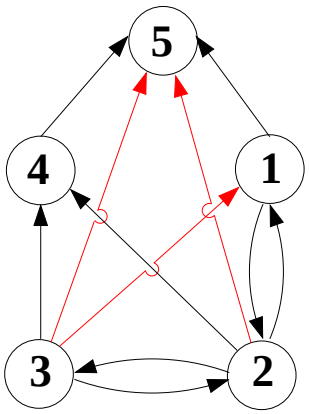
In this Appendix we list the coefficients appearing in Eqs. (9) and (10).
The coefficients in Eq. (9) are:

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

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

The coefficients in Eq. (10) are:

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



$$
\begin{aligned}
w_{12} &= 1+\lambda\lambda_{\mathrm{B}}\big((1+\lambda_{\mathrm{B}})(3+\lambda_{\mathrm{B}})\big)^{-1}\\
w_{13} &= 4\lambda\lambda_{\mathrm{B}}\big((1+\lambda_{\mathrm{B}})(3+\lambda_{\mathrm{B}})(6+\lambda_{\mathrm{B}})\big)^{-1}\\
w_{21} &= 2w_{32} = R\\
w_{23} &= 4(1+\lambda\lambda_{\mathrm{B}}\big((3+\lambda_{\mathrm{B}})(6+\lambda_{\mathrm{B}})\big)^{-1})\\
w_{42} &= 2(1+\lambda\lambda_{\mathrm{B}}\big((1+\lambda_{\mathrm{B}})(3+\lambda_{\mathrm{B}})\big)^{-1})\\
w_{43} &= 2(1+\lambda\lambda_{\mathrm{B}}(7+\lambda_{\mathrm{B}})\big((1+\lambda_{\mathrm{B}})(3+\lambda_{\mathrm{B}})(6+\lambda_{\mathrm{B}})\big)^{-1})\\
w_{51} &= w_{54} = 1+\lambda(1+\lambda_{\mathrm{B}})^{-1}\\
w_{52} &= 3\lambda\big((1+\lambda_{\mathrm{B}})(3+\lambda_{\mathrm{B}})\big)^{-1}\\
w_{53} &= 2\lambda(9+\lambda_{\mathrm{B}})\big((1+\lambda_{\mathrm{B}})(3+\lambda_{\mathrm{B}})(6+\lambda_{\mathrm{B}})\big)^{-1}
\end{aligned}
$$

FIG. 7 Left: graph showing the states and possible transitions corresponding to the limit of severe reductions of population size during bottlenecks. The states $1,\ldots,5$ are explained in Fig. 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 $\lambda = pN_0$, and $\lambda_{\mathrm{B}} = qN_{\mathrm{B}}$.

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 Section III. Formally, we describe this case by the following limit: $N_0 \to \infty$, $N_{\mathrm{B}} \to \infty$, $N_{\mathrm{B}}/N_0 \to 0$, while $\lambda = pN_0$ and $\lambda_{\mathrm{B}} = qN_{\mathrm{B}}$ 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 $(1+\lambda_{\mathrm{B}})^{-1}$. 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 $p/q \to 0$. This not only allows for neglecting the durations of bottlenecks, but it 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 Eq. (10) for the mean covariance $\langle \mathrm{cov}_{\mathscr{D}}[t_{a(ij)}, t_{b(ij)}]\rangle$. Our method for calculating the term $\langle t_{a(ij)} t_{b(ij)}\rangle$ is described in the main text (Section III). But in the Xi-coalescent approximation employed here, the Markov process differs from the one described in Section III. The corresponding graph is shown in Fig. 7. It consists of the same five states $1,\ldots,5$ shown in Fig. 3, but in Fig. 7 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 organised 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 Eq. (10).

1. Formulae 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 $i = 1,\ldots,l$. By construction, the following conditions must be satisfied:

$$l = \sum_{i=1}^{l} i k_i,\ b = \sum_{i=1}^{l} k_i\,. \tag{B1}$$

In our model, the collision rate $\phi_{\{l;k_1,\ldots,k_l\}}$ of l lines colliding into a particular partition $\{l;k_1,\ldots,k_l\}$, such that Eq. (B1) is satisfied, is given by:

$$\phi_{\{l;k_1,\ldots,k_l\}} = 1_{\{b=l-1\}} + \lambda\Xi_{\{l;k_1,\ldots,k_l\}}\,. \tag{B2}$$

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, $\Xi_{\{l;k_1,\ldots,k_l\}}$ can be calculated according to:

$$\Xi_{\{l;k_1,\ldots,k_l\}} = C_{lb}\, p_{\{l;k_1,\ldots,k_l\}}\,, \tag{B3}$$

where

$$C_{lb} = \frac{\lambda_{\mathrm{B}}}{\binom{b}{2}+\lambda_{\mathrm{B}}} \prod_{i=b+1}^{l} \frac{\binom{i}{2}}{\binom{i}{2}+\lambda_{\mathrm{B}}}\,, \tag{B4}$$

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

$$p_{\{l;k_1,\ldots,k_l\}} = \frac{(l-b)!\,b!\,(b-1)!}{l!\,(l-1)!} \prod_{i=1}^{l} (i!)^{k_i}\,. \tag{B5}$$

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

$$\phi^{\mathrm{tot}}_{\{l;k_1,\ldots,k_l\}} = \binom{l}{2} 1_{\{b=l-1\}} + \lambda C_{lb}\, p_{\{l;k_1,\ldots,k_l\}} S_{\{l;k_1,\ldots,k_l\}}\,. \tag{B6}$$

Here

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

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

The graph corresponding to the Markov process in the limit described in this appendix consists of five states $1,\ldots,5$ (see Fig. 7). We now show how the corresponding transition rates between the states $1,\ldots,5$ can be derived from Eq. (B6).

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

$$w_{51} = w_{54} = \phi^{\mathrm{tot}}_{\{2;0,1\}}\,. \tag{B8}$$

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:

$$w_{12} = \frac{1}{3}\phi^{\mathrm{tot}}_{\{3;1,1,0\}}\,, \quad w_{42} = \frac{2}{3}\phi^{\mathrm{tot}}_{\{3;1,1,0\}}\,. \tag{B9}$$

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

$$w_{52} = \phi^{\mathrm{tot}}_{\{3;0,0,1\}}\,. \tag{B10}$$

Now consider transitions from state 3, consisting of four ancestral lines. We analyse 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:

$$w_{23} = \frac{2}{3}\phi^{\mathrm{tot}}_{\{4;2,1,0,0\}}\,. \tag{B11}$$

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 Fig. 2**d**). 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:

$$w_{13} = \frac{2}{3}\phi^{\mathrm{tot}}_{\{4;0,2,0,0\}}\,, \quad w_{43} = \frac{1}{3}\phi^{\mathrm{tot}}_{\{4;2,1,0,0\}} + \phi^{\mathrm{tot}}_{\{4;1,0,1,0\}}\,, \quad w_{53} = \frac{1}{3}\phi^{\mathrm{tot}}_{\{4;0,2,0,0\}} + \phi^{\mathrm{tot}}_{\{4;0,0,0,1\}}\,. \tag{B12}$$

The remaining non-vanishing 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. Explict formulae for the rates w_{mn} in terms of λ and λ_B are given in Fig. 7.

2. Obtaining Eq. (10) under the Xi-coalescent approximation

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

$$\langle t_{a(ij)} t_{b(ij)} \rangle = \int_0^\infty \mathrm{d}t_1 \, t_1^2 \boldsymbol{u} \, \mathrm{e}^{\mathbf{M}t_1} \boldsymbol{v} + \int_0^\infty \mathrm{d}t_1 \int_{t_1}^\infty \mathrm{d}t_2 \, t_1 t_2 \, c \, \mathrm{e}^{K(t_2 - t_1)} \boldsymbol{Q} \, \mathrm{e}^{\mathbf{M}t_1} \boldsymbol{v} \, . \tag{B13}$$

This equation corresponds to Eq. (7) given in Section III. But the dimensions of the matrices and vectors in Eq. (B13) differ from those in Eq. (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}:

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

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 Eq. (B13) are $K = -1$, $c = 1$, $\boldsymbol{v} = [1,0,0]^\mathsf{T}$, and:

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

Combining Eq. (B13), with Eqs. (B14)-(B15) results in Eq. (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.