Fixation Probability in a Two-Locus Model by the Ancestral Recombination–Selection Graph

Sabin Lessard[1] and Amir R. Kermany
Département de Mathématiques et de Statistique, Université de Montréal, Montreal, Quebec H3C 3J7, Canada

ABSTRACT We use the ancestral influence graph (AIG) for a two-locus, two-allele selection model in the limit of a large population size to obtain an analytic approximation for the probability of ultimate fixation of a single mutant allele *A*. We assume that this new mutant is introduced at a given locus into a finite population in which a previous mutant allele *B* is already segregating with a wild type at another linked locus. We deduce that the fixation probability increases as the recombination rate increases if allele *A* is either in positive epistatic interaction with *B* and allele *B* is beneficial or in no epistatic interaction with *B* and then allele *A* itself is beneficial. This holds at least as long as the recombination fraction and the selection intensity are small enough and the population size is large enough. In particular this confirms the Hill–Robertson effect, which predicts that recombination renders more likely the ultimate fixation of beneficial mutants at different loci in a population in the presence of random genetic drift even in the absence of epistasis. More importantly, we show that this is true from weak negative epistasis to positive epistasis, at least under weak selection. In the case of deleterious mutants, the fixation probability decreases as the recombination rate increases. This supports Muller's ratchet mechanism to explain the accumulation of deleterious mutants in a population lacking recombination.

THE Hill–Robertson (HR) effect (Hill and Robertson 1966) is often mentioned as one of the main arguments in favor of the evolution of recombination. In short, it predicts that beneficial mutant alleles arising at different loci in a finite population are more likely to fix in the population as the recombination rate increases even when selection acts independently upon the loci.

Since the early works of Fisher (1930) and Muller (1932), it is generally believed that an evolutionary advantage of recombination is to bring together beneficial mutant alleles arising at different loci. Accordingly the effect of recombination should be to increase the rate of evolution of the population (Crow and Kimura 1965). However, it has been shown that recombination has no effect on this rate in an infinite population if there is initial linkage equilibrium and absence of epistasis so that linkage equilibrium is maintained thereafter in the population (Felsenstein 1965; Maynard Smith 1968).

If recombination can have an effect on the rate of evolution only by breaking down linkage disequilibrium in absolute value, then the effect should be to increase this rate only when linkage disequilibrium in the population is negative (NLD). In the case of a two-locus model, this happens when the frequency of the double mutant is strictly smaller than the product of the frequencies of the mutant alleles. This situation is arguably likely to happen in the view that beneficial mutations are very rare (Crow and Kimura 1969).

On the other hand, NLD could be produced by negative epistasis (NE), with the double mutant being less fit than what it would be under independent effects of the mutant alleles. Then the double mutant would die more often than is expected with mutant alleles acting independently, leaving NLD in the population. In the opposite case of positive epistasis (PE), Eshel and Feldman (1970) showed that the frequency of the double mutant in an infinite population is always larger in an asexual population than in a population with recombination. This suggests that NE rather than PE could be advantageous for the evolution of recombination.

In taking account of a finite population size, Bodmer (1970) considered the expected time until the first formation of a double mutant from two initial single mutants. He concluded that recombination would have a greater advantage in a small population than in a large one. Karlin (1973)

doi: 10.1534/genetics.111.136309
Manuscript received June 10, 2011; accepted for publication November 14, 2011
[1]Corresponding author: Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montreal, Quebec H3C 3J7, Canada. E-mail: lessards@dms.umontreal.ca

showed that this expected time, without selection effects, was indeed a decreasing function of the recombination rate *r*. But he showed also that the expected time until the total fixation of the double mutant was an increasing function of *r*. In other words, increasing recombination might be advantageous in speeding the time until the first formation of the double mutant, but disadvantageous by breaking apart the favored gamete type once formed.

Summing first-order terms for expected changes in gene frequencies in a large finite population under weak selection with additive gene action, which corresponds to an absence of epistasis (AE), Hill and Robertson (1966) deduced that the probability of fixation of an allele *A* initially in NLD with a beneficial allele *B* at another tightly linked locus increases with the recombination fraction between the two loci. The effects of linkage disequilibrium and epistasis on the probability of fixation of gametes and alleles in a finite population under the assumption of weak selection were further studied by diffusion approximations in the limit of a large population size (Ohta 1968).

In the case of initial linkage equilibrium (LE) in a finite population, first-order approximations fail to detect the effect of linkage on the fixation probability. Higher-order effects in the absence of epistasis were first exhibited by simulations (Hill and Robertson 1966).

In a finite population initially in LE, genetic drift creates random instances of linkage disequilibrium. Although random drift can generate both positive and negative disequilibria without any *a priori* bias on average, selection dispels positive linkage disequilibrium (PLD) more efficiently than NLD even in the absence of epistasis, so that the average linkage disequilibrium becomes negative. As shown by simulations and some analytical arguments (Hill and Robertson 1966; see also Barton and Otto 2005), this leads to an average accumulation of NLD. As a consequence, responses to selection at different loci are expected to interfere with each other, even in the absence of gene interaction. This is known as the HR effect. It is by reducing the interference caused by the randomly generated linkage disequilibrium that an increase in the recombination rate raises the rate of fixation of favorable mutants.

The relationship between the HR effect and the Fisher–Muller theory for the evolutionary advantage of recombination was pointed out by Felsenstein (1974). Moreover, it was noted that Muller's ratchet mechanism (Muller 1964) for the accumulation of deleterious mutants in the absence of recombination, which is formally equivalent to the accumulation of advantageous mutants in the presence of recombination, can be explained by the HR effect.

The theory of evolution at a selectively neutral modifier locus that controls the recombination fraction between two major loci that are under selection in an infinite population was developed by Feldman *et al.* (1980). If the major loci are in linkage disequilibrium at a balance between selection against deleterious alleles and mutation toward them, then a mutation increasing recombination succeeds if the linkage disequilibrium is negative, which occurs when epistasis is negative, and the modifier locus is sufficiently tightly linked to the major loci. If the modifier locus is loosely linked, NE has to be weak enough (Otto and Feldman 1997). A similar conclusion has been reached for sweeps of beneficial alleles (Barton 1995a). However, including spatial heterogeneity extends the range of epistasis over which recombination can be favored, from strong NE to PE depending on environmental circumstances (Lenormand and Otto 2000).

On the other hand, Feldman *et al.* (1980) also showed that, if the major loci are at a stable equilibrium in linkage disequilibrium under selection and recombination, then a mutation at the modifier locus increases in frequency when rare if and only if it decreases the recombination fraction. This is part of a general reduction principle for genetic modifiers in an infinite population in a constant environment (Feldman and Liberman 1986).

It has been argued that a modifier allele that increases the recombination rate would be promoted in a finite population due to its role in reducing the negative effect of poor genetic backgrounds on the probability of fixation of favorable mutants, at least in the absence of epistasis. This has been shown by applying a branching process to mutant lines in an infinite population with deterministic changes in the frequencies of the genetic backgrounds (Barton 1995b; Otto and Barton 1997). The same approach has been used to study the probability that both beneficial mutants fix ~~and the~~ analysis of this probability has been refined to deal with the troublesome case where the second mutant is more beneficial than the first (Yu and Etheridge 2010).

Simulations have indicated that this is true across a broad range of epistatic interactions, from weak negative epistasis to positive epistasis, provided that the population size is small enough (Otto and Barton 2001). This suggests that the HR effect overwhelms the influence of epistasis on LD over a wide range of epistasis values.

More recently, a perturbation method to track fluctuations in linkage disequilibrium during the spread of beneficial alleles and to measure the impact on a modifier allele of recombination has been proposed (Barton and Otto 2005). The method consists of considering only the first and second moments of random sampling effects on the deterministic dynamics for the allele frequencies and linkage disequilibrium in an infinite population.

A different perturbation technique to approximate the probability of ultimate fixation of an allele in a multilocus setting assumes small selection effects at different loci in a population of fixed finite size (Lehman and Rousset 2009). This is an extension of a direct Markov chain approach for one-locus models based on expected changes in allele frequencies in one time step or one generation (Rousset 2003; Lessard and Ladret 2007; Lessard and Lahaie 2009). Then the first-order effect of selection can be expressed in terms of expected times that lineages of sampled genes take to merge backward in time, under neutrality. The calculation of these times for one-locus models in the limit of a large population

size makes use of the coalescent (Kingman 1982) and its extension to incorporate multiple mergers in the case of highly skewed reproduction schemes (Pitman 1999; Sagitov 1999; Möhle and Sagitov 2001).

In the case of multilocus selection models with a Wright–Fisher reproduction scheme allowing for recombination in a population of fixed finite size, Lehman and Rousset (2009) considered Taylor expansions of the fixation probability with respect to the intensity of selection. They deduced exact linear recurrence systems of equations for gamete frequencies in sampled individuals backward in time under neutrality to compute the coefficients. Advanced matrix theory was used to interpret these coefficients in terms of mean sojourn times in the backward neutral process. However, a first-order expansion of the fixation probability with respect to the intensity of selection is not sufficient to detect the HR effect in a two-locus model in the absence of epistasis. Actually, a third-order expansion is necessary. In this case, the coefficients of the approximation become difficult to interpret.

Our objective in this article is to consider an ancestral recombination–selection process to deduce an analytic approximation for the probability of ultimate fixation of an allele in a finite but large population under weak selection and tight linkage. The allele is assumed to be a mutant type A introduced at a given locus into the population in which a previous mutant type B is already segregating with a wild type at another linked locus. Exact conditions for a small increase in the recombination rate to increase the probability of ultimate fixation of a single A are addressed.

We focus on a discrete-time two-locus selection model with a Moran reproduction scheme (Moran 1958). We consider the ancestral recombination–selection graph for sampled gametes in the limit of a large population size, which is known as the ancestral influence graph (AIG) (Donnelly and Kurtz 1999). The AIG provides a supragenealogy for a sample of individuals at linked, nonneutral loci in a limiting Fleming–Viot measure-valued diffusion process with selection and recombination. It is a supragenealogy in the sense that the true genealogy of the sample is embedded into it. It combines the ancestral recombination graph (ARG) (Griffiths and Marjoram 1996, 1997) and the ancestral selection graph (ASG) (Krone and Neuhauser 1997; Neuhauser and Krone 1997), extending the coalescent (Kingman 1982) to include both recombination and selection. The ARG and ASG, given the sample composition, have been widely used in likelihood methods to estimate the recombination rate or detect recombination hotspots (*e.g.*, McVean *et al.* 2002; Stephens and Donnelly 2003; Fearnhead *et al.* 2004; Wakeley and Sargsyan 2009) and to locate disease genes from marker loci (*e.g.*, Hudson and Kaplan 1988; Fearnhead 2003; Larribe and Lessard 2008; Larribe and Fearnhead 2011).

We make use of a discrete-time Moran model for mortality selection determined at two loci in a finite haploid population to ascertain the analysis. After recalling the definitions and assumptions, the probability of ultimate fixation of an allele is expressed in terms of sums of expected sample frequencies, which correspond to expected times with given ordered random samples. Then the ancestral graphs obtained by tracing the ~~genealogy~~ of an ordered random sample through coalescence, recombination, or selection events backward in time, whose limit as the population size increases is an AIG, are described. These graphs are used to express the expected times with ordered random samples of given types. It is shown that an approximation of any order of the fixation probability with respect to the population-scaled recombination and selection parameters in the limit of a large population size can be obtained by considering ancestral graphs with enough recombination or selection events. Finally this is applied to directional selection with either beneficial mutants or deleterious mutants, in epistatic interaction or in the absence of interaction, by considering one recombination event and one or two selection events to detect the effect of recombination.

It is expected that the results are valid in the domain of attraction of the Fleming–Viot process with recombination and selection in the same way that a wide class of Cannings exchangeable models including the Moran model and the Wright–Fisher model fall in the domain of attraction of the Kingman coalescent (Möhle and Sagitov 2001).

Definitions and Model

Suppose a population of finite size N distributed over N distinct sites, so that each site is occupied by one and only one individual. Each individual is one of four types, AB, Ab, aB, or ab, with respect to two loci with alleles A, a segregating at locus 1 and B, b at locus 2.

Reproduction is assumed to follow a discrete-time Moran model. At each time step $\tau \geq 0$, two individuals are sampled at random in the population and they produce an offspring. Random sampling of the parents is assumed to take place with replacement so that selfing is permitted and then occurs with probability N^{-1}.

With respect to the two loci, the offspring produced is either an exact copy of one of its parents, with probability $1 - r$, or a recombinant, with probability r. This probability of recombination is inversely proportional to the population size, so that $r = \rho N^{-1}$, where ρ represents a population-scaled recombination fraction. Weak recombination is modeled by keeping ρ constant as $N \to \infty$.

On the other hand, one individual is chosen at random to be replaced by the offspring. Replacement actually occurs with some probability that depends on the type of the individual, called its mortality. It is given by $1 - c_{AB}s$, $1 - c_{Ab}s$, $1 - c_{aB}s$, or $1 - c_{ab}s$ for an individual of type AB, Ab, aB, or ab, respectively (see Figure 1). These can be interpreted as probabilities of dying. If replacement does not occur, then the offspring is eliminated and there is no change in the population during the corresponding time step.

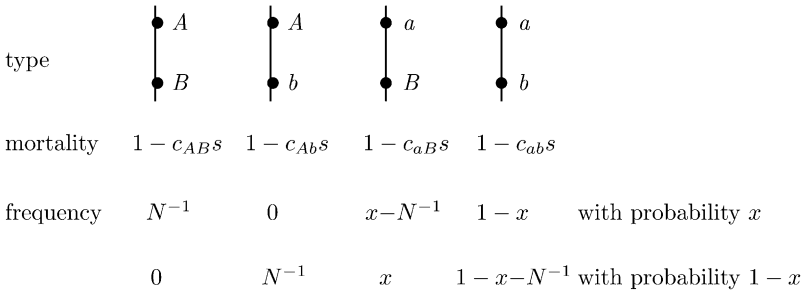

Figure 1 Two mutant alleles, *A* and *B*, segregating with two wild types, *a* and *b*, at two linked loci.

Here, the parameters $0 \le c_{AB}, c_{Ab}, c_{aB}, c_{ab} \le 1$ represent coefficients of selection with respect to an intensity of selection $0 < s < 1$. They can be viewed as viability parameters. Neutrality corresponds to $s = 0$.

The intensity of selection is expressed in the form $s = \sigma N^{-1}$, where N is the population size. The parameter σ stands for a population-scaled intensity of selection. Weak selection is modeled by keeping σ constant as $N \to \infty$.

Alleles *A* and *B* are mutant types, while alleles *a* and *b* are wild types. The mutant alleles *A* and *B* are advantageous when each one reduces the mortality of its carrier compared to what it would be without these alleles. This is the case if the coefficients of selection satisfy the inequalities

$$c_{AB} > \max(c_{Ab}, c_{aB}) \ge \min(c_{Ab}, c_{aB}) > c_{ab}. \tag{1}$$

On the other hand, if we have

$$c_{AB} < \min(c_{Ab}, c_{aB}) \le \max(c_{Ab}, c_{aB}) < c_{ab}, \tag{2}$$

then the mutant alleles *A* and *B* are deleterious.

Allele *B* is a mutant that was introduced some time ago at locus 2 into a population entirely composed of *ab* individuals and its frequency has reached some value $0 < x < 1$. Then a single mutant allele *A* is introduced at random at locus 1 into the population, so that it is linked to *B* with probability *x* and to *b* with the complementary probability $1 - x$. In both cases its frequency is given by the inverse of the population size, that is, N^{-1}. In the former case, the frequency of *aB* is reduced to $x - N^{-1}$ and in the latter the frequency of *ab* is reduced to $1 - x - N^{-1}$.

Linkage disequilibrium (LD) is measured by the difference between the frequency of the double mutant, *AB*, and the product of the frequencies of the mutant alleles, *A* and *B*, which is represented by *D*. Alternatively, *D* is equal to the difference between the product of the frequencies of *AB* and *ab* and the product of the frequencies of *Ab* and *aB*. In the present case, linkage disequilibrium following the introduction of a single *A* is initially positive (PLD) with probability *x* and given by $N^{-1}(1 - x)$, while it is initially negative (NLD) with probability $1 - x$ and given by $-N^{-1}x$. This yields an average LD given by

$$D = N^{-1}(1-x)x + \left(-N^{-1}x\right)(1-x) = 0. \tag{3}$$

Then we are in a situation of an initial average LE.

Epistasis refers to the phenomenon in which the effect of a mutant at one locus, here *B*, is masked or enhanced by a mutant at another locus, here *A*. Population geneticists extended the concept to mean nonindependent or multiplicative effects of mutants.

Epistasis is positive (PE) if interactions between *A* and *B* are such that the double mutant is more fit in comparison to the wild gamete type than what it would be if the mutant alleles have independent effects on fitness. In terms of mortality parameters, this means the inequality

$$\frac{1-c_{AB}s}{1-c_{ab}s} < \frac{1-c_{Ab}s}{1-c_{ab}s} \cdot \frac{1-c_{aB}s}{1-c_{ab}s}. \tag{4}$$

If the inequality is reversed, then epistasis is negative (NE). In the case of an equality, there is no epistasis (AE).

Where advantageous mutations are concerned, PE enhances the fitness increase predicted from individual mutational effects, whereas NE lessens it. It is the opposite for deleterious mutations with respect to fitness decrease.

Note that, in the limit of weak selection when $s = \sigma N^{-1} \to 0$ as $N \to \infty$, epistasis is positive, negative, or null if

$$\varepsilon = c_{AB} - c_{Ab} - c_{aB} + c_{ab} \tag{5}$$

is positive, negative, or null, respectively. Moreover, note that $-2 \le \varepsilon \le 2$ under our general conditions on the coefficients of selection, but $-1 < \varepsilon < 1$ in the case of either advantageous mutations or deleterious mutations.

Expected Change in Allele Frequency

Let $\mathbf{x}(\tau) = (x_{AB}(\tau), x_{Ab}(\tau), x_{aB}(\tau), x_{ab}(\tau))$ be the vector of the individual type frequencies at the current time step $\tau \ge 0$. Then the frequency of *AB* at the next time step will increase by N^{-1} with probability

$$\begin{aligned} \nu_{AB}(\tau) = {} & x_{AB}(\tau)(1 - x_{AB}(\tau)) + N^{-1}\sigma x_{AB}(\tau)(c_{AB}x_{AB}(\tau) - \bar{c}(\tau)) \\ & + N^{-1}\rho(1 - x_{AB}(\tau))(x_A(\tau)x_B(\tau) - x_{AB}(\tau)) \\ & + N^{-2}\sigma\rho(c_{AB}x_{AB}(\tau) - \bar{c}(\tau))(x_A(\tau)x_B(\tau) - x_{AB}(\tau)). \end{aligned} \tag{6}$$

Similarly it will decrease by N^{-1} with probability

$$\mu_{AB}(\tau) = x_{AB}(\tau)(1 - x_{AB}(\tau)) - N^{-1}\sigma c_{AB}x_{AB}(\tau)(1 - x_{AB}(\tau)) - N^{-1}\rho x_{AB}(\tau)(x_A(\tau)x_B(\tau) - x_{AB}(\tau)) + N^{-2}\sigma\rho c_{AB}x_{AB}(\tau)(x_A(\tau)x_B(\tau) - x_{AB}(\tau)). \quad (7)$$

Otherwise it will remain the same with the complementary probability $1 - \nu_{AB}(\tau) - \mu_{AB}(\tau)$. Here we use the notation

$$x_A(\tau) = x_{AB}(\tau) + x_{Ab}(\tau) = 1 - x_a(\tau) \quad (8)$$

for the frequency of allele A and similarly

$$x_B(\tau) = x_{AB}(\tau) + x_{aB}(\tau) = 1 - x_b(\tau) \quad (9)$$

for the frequency of allele B. Moreover,

$$\bar{c}(\tau) = c_{AB}x_{AB}(\tau) + c_{Ab}x_{Ab}(\tau) + c_{aB}x_{aB}(\tau) + c_{ab}x_{ab}(\tau) \quad (10)$$

stands for the mean coefficient of selection at time step τ.

Therefore, the change in the frequency of AB from time step τ to time step $\tau + 1$, given by $\Delta x_{AB}(\tau) = x_{AB}(\tau + 1) - x_{AB}(\tau)$, is found to have

$$E(\Delta x_{AB}(\tau)|\mathbf{x}(\tau)) = N^{-2}\sigma x_{AB}(\tau)(c_{AB} - \bar{c}(\tau)) + N^{-2}\rho(x_A(\tau)x_B(\tau) - x_{AB}(\tau)) - N^{-3}\sigma\rho\bar{c}(\tau)(x_A(\tau)x_B(\tau) - x_{AB}(\tau)) \quad (11)$$

as conditional expectation. Similarly we have

$$E(\Delta x_{Ab}(\tau)|\mathbf{x}(\tau)) = N^{-2}\sigma x_{Ab}(\tau)(c_{Ab} - \bar{c}(\tau)) + N^{-2}\rho(x_A(\tau)x_b(\tau) - x_{Ab}(\tau)) - N^{-3}\sigma\rho\bar{c}(\tau)(x_A(\tau)x_b(\tau) - x_{Ab}(\tau)) \quad (12)$$

for the change in the frequency of Ab. Hence the change in the frequency of allele A, which can be expressed as $\Delta x_A(\tau) = \Delta x_{AB}(\tau) + \Delta x_{Ab}(\tau)$, has conditional expectation

$$E(\Delta x_A(\tau)|\mathbf{x}(\tau)) = N^{-2}\sigma x_A(\tau)(\bar{c}_A(\tau) - \bar{c}(\tau)). \quad (13)$$

Here the quantity

$$\bar{c}_A(\tau) = c_{AB}\frac{x_{AB}(\tau)}{x_A(\tau)} + c_{Ab}\frac{x_{Ab}(\tau)}{x_A(\tau)} \quad (14)$$

represents the marginal coefficient of selection of allele A at time step τ. Straightforward algebraic manipulations lead to the following conclusion.

Proposition 1 *For the discrete-time Moran model with recombination and selection described in Figure 1, the conditional expected change in the frequency of A is given by*

$$N^2\sigma^{-1}E(\Delta x_A(\tau)|\mathbf{x}(\tau)) = x_{AB}(\tau)x_{aB}(\tau)(c_{AB} - c_{aB}) + x_{AB}(\tau)x_{ab}(\tau)(c_{AB} - c_{ab}) + x_{Ab}(\tau)x_{aB}(\tau)(c_{Ab} - c_{aB}) + x_{Ab}(\tau)x_{ab}(\tau)(c_{Ab} - c_{ab}), \quad (15)$$

where N is the population size and $\sigma = sN$ is a population-scaled intensity of selection with coefficients $0 \le c_{AB}, c_{Ab}, c_{aB}, c_{ab} \le 1$ for the individual types AB, Ab, aB, and ab, respectively.

Probability of Fixation of an Allele at One Locus

The random process $\mathbf{x}(\tau) = (x_{AB}(\tau), x_{Ab}(\tau), x_{aB}(\tau), x_{ab}(\tau))$ for $\tau \ge 0$ is a Markov chain on a finite state space S. This is the set of all four-dimensional frequency vectors whose entries are multiples of N^{-1}.

There are four absorbing states represented by

$$\mathbf{e}_{AB} = (1,0,0,0),\ \mathbf{e}_{Ab} = (0,1,0,0,),\ \mathbf{e}_{aB} = (0,0,1,0),\ \mathbf{e}_{ab} = (0,0,0,1).$$

These correspond to the fixation of AB, Ab, aB, and ab, respectively. All other states are transient.

In virtue of the ergodic theorem for Markov chains (see, *e.g.*, Karlin and Taylor 1975; Grimmett and Stirzaker 1982), the probability of transition from state $\mathbf{x}$ to state $\mathbf{y}$ in k time steps, namely

$$P_{\mathbf{xy}}(k) = P(\mathbf{x}(k) = \mathbf{y} \mid \mathbf{x}(0) = \mathbf{x}), \quad (16)$$

converges to some fixation probability in the limit of a large number of time steps, represented by $P_{\mathbf{xy}}(\infty)$. This probability is 0 unless $\mathbf{y}$ is an absorbing state. Therefore,

$$E(x_A(k)|\mathbf{x}(0) = \mathbf{x}) = \sum_{\mathbf{y}}(y_{AB} + y_{Ab})P_{\mathbf{xy}}(k), \quad (17)$$

where the summation is over all $\mathbf{y} = (y_{AB}, y_{Ab}, y_{aB}, y_{ab})$ in S, converges in the same limit to

$$P_{\mathbf{x}\mathbf{e}_{AB}}(\infty) + P_{\mathbf{x}\mathbf{e}_{Ab}}(\infty) = u_A(\mathbf{x}). \quad (18)$$

This is the probability of ultimate fixation of A given an initial population state $\mathbf{x}$.

On the other hand, we have

$$x_A(k) = x_A(0) + \sum_{\tau=0}^{k}\Delta x_A(\tau), \quad (19)$$

and therefore

$$E(x_A(k)|\mathbf{x}(0)) = x_A(0) + \sum_{\tau=0}^{k}E(\Delta x_A(\tau)|\mathbf{x}(0)) \quad (20)$$

by additivity of conditional expectation. As $k \to \infty$, this leads to

$$u_A(\mathbf{x}(0)) = x_A(0) + \sum_{\tau=0}^{\infty}E(\Delta x_A(\tau)|\mathbf{x}(0)). \quad (21)$$

The law of total expectation guarantees that

$$E(\Delta x_A(\tau)|\mathbf{x}(0)) = E(E(\Delta x_A(\tau)|\mathbf{x}(\tau))|\mathbf{x}(0)). \tag{22}$$

Then Proposition 1 for the conditional expected change in the frequency of A leads to the following result.

Proposition 2 *For the discrete-time Moran model with recombination and selection of Proposition 1, the probability of ultimate fixation of A is given by*

$$\begin{aligned} u_A(\underline{x}(0)) = \; & x_A(0) + \frac{\sigma}{2}(c_{AB} - c_{aB})E_{AB,aB}(\underline{x}(0)) \\ & + \frac{\sigma}{2}(c_{AB} - c_{ab})E_{AB,ab}(\underline{x}(0)) \\ & + \frac{\sigma}{2}(c_{Ab} - c_{aB})E_{Ab,aB}(\underline{x}(0)) \\ & + \frac{\sigma}{2}(c_{Ab} - c_{ab})E_{Ab,ab}(\underline{x}(0)), \end{aligned}$$

where

$$E_{z_1,z_2}(\mathbf{x}(0)) = 2N^{-2}\sum_{\tau=0}^{\infty} E(x_{z_1}(\tau)x_{z_2}(\tau)|\mathbf{x}(0)), \tag{23}$$

for $z_1 = AB, Ab$ and $z_2 = aB, ab$.

The quantity $E_{z_1,z_2}(\mathbf{x}(0))$ defined in Proposition 2 represents the expected time in number of $N^2/2$ time steps and over all time steps that two individuals chosen at random *with replacement* in the population at the same time step $\tau \geq 0$ will be of types z_1 and z_2 in this order.

Ancestral Recombination–Selection Graph

An ancestral recombination–selection graph is a Markov chain on ordered samples obtained by tracing backward in time the ancestors, real or virtual, of a given number of individuals chosen at random without replacement in the population at a given time step. It is characterized by a sequence of changes in the ancestry of the sample and times between these events.

As in Krone and Neuhauser (1997), this process is considered in the framework of a Moran model, but in discrete time and with recombination allowed, so that a change in the ancestry can involve simultaneous events of coalescence, recombination, or selection. In the limit of a large population size, however, with time and parameters for recombination and selection appropriately scaled, only one event of coalescence, recombination, or selection can occur at a time with probability one. The limiting process corresponds to the AIG introduced by Donnelly and Kurtz (1999), as described in Fearnhead (2003). An exact description of the ancestral graph incorporating recombination and selection in a discrete-time Moran model could not be found in the literature, although it might exist. Such a description is actually necessary to establish rigorous approximation results for the probability of fixation in the presence of recombination and selection.

Consider the model of the previous section with $1 - c_{AB}s$, $1 - c_{Ab}s$, $1 - c_{aB}s$, and $1 - c_{ab}s$ as mortalities associated to the individual types AB, Ab, aB, and ab, respectively, under the conditions $0 \leq c_{AB}, c_{Ab}, c_{aB}, c_{ab} \leq 1$ and $0 < s < 1$.

The replacement rule for an individual chosen at random can be described as follows. Replacement is inevitable irrespective of the type of the individual with probability $1 - s$, which corresponds to the lowest possible mortality. On the other hand, replacement is type specific with probability s. In this case, replacement occurs with conditional probability

$$\begin{cases} 1 - c_{AB} & \text{if} \quad AB, \\ 1 - c_{Ab} & \text{if} \quad Ab, \\ 1 - c_{aB} & \text{if} \quad aB, \\ 1 - c_{ab} & \text{if} \quad ab. \end{cases}$$

The law of total probability guarantees that the probability of replacement is given by the mortality of the individual. For an individual of type Ab, for instance, replacement will occur with probability $1 - s + s(1 - c_{Ab})$, which is the same as $1 - c_{Ab}s$. With the complementary probability, there is no replacement.

A type-specific replacement is considered to be a selection event. Its probability in one time step is expressed in the form $s = \sigma N^{-1}$. Recall that the probability of a recombination event in one time step is expressed in a similar form, namely $r = \rho N^{-1}$.

The scaling used for the probabilities of selection or recombination events, along with $N^2/2$ time steps as unit of time, will simplify the ancestral process in the limit of a large population size. This timescale is standard for a discrete-time Moran model (see, *e.g.*, Ewens 1990).

Consider a sample of n distinct individuals in the population at a given time step and label them arbitrarily with the integers $i = 1, \ldots, n$. Label arbitrarily the other $N - n$ individuals in the population at the same time step with the integers $i = n + 1, \ldots, N$.

Following the lineages of the sampled individuals in one time step back, there will be pure coalescence of i and j, for $i, j = 1, \ldots, n$ with $i \neq j$, if the offspring produced was an exact copy of j [probability $N^{-1}(1 - \rho N^{-1})$] and the individual replaced irrespective of its type was the individual that occupied the site of i [probability $N^{-1}(1 - \sigma N^{-1})$] or vice versa. We conclude that

$$\frac{2}{N^2}\left(1 - \frac{\sigma}{N}\right)\left(1 - \frac{\rho}{N}\right) = 2N^{-2}\left(1 + O\left(N^{-1}\right)\right) \leq 2N^{-2} \tag{24}$$

is the probability for each pure coalescence event to occur within a sample of size n in one time step back. Then the sample size is reduced by one by merging the lineages of two sampled individuals.

On the other hand, there will be pure recombination of i in one time step back, for $i = 1, \ldots, n$, if the offspring produced was a recombinant of k and l not in the sample and different from each other, that is, for $k, l = n + 1, \ldots, N$ with $k \neq l$ [probability $\rho N^{-1}(N - n)(N - n - 1)N^{-2}$], and the individual replaced irrespective of its type was the individual

that occupied the site of i [probability $N^{-1}(1 - \sigma N^{-1})$]. Therefore, we find that

$$\frac{\rho(N-n)(N-n-1)}{N^4}\left(1-\frac{\sigma}{N}\right) = \rho N^{-2}\left(1+O\left(N^{-1}\right)\right) \leq \rho N^{-2} \tag{25}$$

is the probability for each pure recombination event to occur in a sample of size n in one time step back. In this case the sample size is increased by one by splitting the lineage of one sampled individual into two, each one being actually ancestral to the sampled individual at only one of the loci.

Finally there will be pure selection of i in one time step back, for $i = 1, \ldots, n$, if the offspring produced was an exact copy of k not in the sample, that is, for $k = n + 1, \ldots, N$ [probability $(N - n)N^{-1}(1 - \rho N^{-1})$], and the individual chosen to be replaced according to its type is the one that occupied the site of i (probability σN^{-2}). We conclude that

$$\frac{\sigma(N-n)}{N^3}\left(1-\frac{\rho}{N}\right) = \sigma N^{-2}\left(1+O\left(N^{-1}\right)\right) \leq \sigma N^{-2} \tag{26}$$

is the probability for each pure selection event to occur in a sample of size n in one time step back.

In the case of a pure selection event, the sample size is increased by one by branching the lineage of one sampled individual into two, each one being potentially ancestral to the sampled individual at both loci. The incoming lineage is the lineage of the offspring produced one time step back, while the continuing lineage is the lineage of the individual chosen to be replaced by the offspring. One of these lineages is real and the other virtual, but both lineages must be traced back until ancestors of known types are reached. Then the conditional probability of replacement can be determined.

Note that the probabilities of pure coalescence, recombination, or selection events in one time step back for a sample of fixed size n are all functions of order N^{-2}, denoted by $O(N^{-2})$. On the other hand, the probabilities of multiple events involving simultaneous coalescence, recombination, or selection events that would affect the lineages of the sampled individuals in one time step back are all functions of order $O(N^{-3})$. In all cases the sample size can decrease by at most one, when a pure coalescence event occurs, and increase by at most two, when a selection event and a recombination event occur simultaneously but without any coalescence event occurring.

Given a sample of size n, the total number of pure coalescence events to consider is $n(n - 1)/2$, while this number is n for pure selection events and for pure recombination events. Therefore, the total probability of change in one time step back for the whole sample is given by

$$p_n = 2\lambda_n N^{-2} + O\left(N^{-3}\right), \tag{27}$$

where

$$\lambda_n = \frac{n(n-1+\rho+\sigma)}{2}. \tag{28}$$

This quantity represents the total rate of change in the limit of a large population size with $N^2/2$ time steps as unit of time. Moreover, given a change in one time step back, the conditional probability of each pure coalescence, recombination, or selection event is

$$P(C_n) = \frac{1}{\lambda_n} + O\left(N^{-1}\right), \tag{29}$$

$$P(R_n) = \frac{\rho}{2\lambda_n} + O\left(N^{-1}\right), \tag{30}$$

or

$$P(S_n) = \frac{\sigma}{2\lambda_n} + O\left(N^{-1}\right), \tag{31}$$

respectively, and independently of everything else, while the conditional probability of each multiple event is

$$P(M_n) = O\left(N^{-1}\right). \tag{32}$$

In the limit of a large population size, the conditional probabilities of multiple events vanish.

Let the time back, in number of time steps, for a sample of size n to be affected by any coalescence, recombination, or selection event be represented by τ_n. This sojourn time is a geometric random variable independent of all previous transition events and sojourn times, whose expected value is given by

$$E(\tau_n) = \sum_{k=0}^{\infty} P(\tau_n > k), \tag{33}$$

where

$$P(\tau_n > k) = (1-p_n)^k. \tag{34}$$

The corresponding time back in number of $N^2/2$ time steps, namely

$$T_n = 2\tau_n N^{-2}, \tag{35}$$

converges in distribution to an exponential random variable with parameter λ_n in the limit of a large population size. As a matter of fact,

$$P(T_n > t) = P\left(\tau_n > \left\lfloor \frac{tN^2}{2} \right\rfloor\right), \tag{36}$$

where $\lfloor\ \rfloor$ denotes the integer value, and

$$\lim_{N\to\infty} P\left(\tau_n > \left\lfloor \frac{tN^2}{2} \right\rfloor\right) = \exp(-\lambda_n t), \tag{37}$$

for every $t > 0$. Moreover,

$$E(T_n) = \int_0^\infty P(T_n > t)dt, \tag{38}$$

where

$$P(T_n > t) \leq \left(1 - \lambda_n N^{-2}\right)^{tN^2/2-1} \leq 2\exp\frac{-\lambda_n t}{2}, \tag{39}$$

for N large enough. Therefore,

$$\lim_{N\to\infty} E(T_n) = \int_0^\infty \exp(-\lambda_n t)dt = \lambda_n^{-1} \tag{40}$$

in virtue of the dominated convergence theorem.

Let us summarize.

Proposition 3 *Consider the discrete-time Moran model of Proposition 1 with population-scaled recombination fraction $\rho = rN$ and population-scaled intensity of selection $\sigma = sN$ in the case of coefficients of selection $0 \leq c_{AB}, c_{Ab}, c_{aB}, c_{ab} \leq 1$. In addition, consider an ordered sample without replacement of size n in a population of large size N. Backward in time, each pair of lineages merges as a result of a coalescence event with approximate probability $2N^{-2}$, while each lineage splits into two as a result of a recombination event with approximate probability ρN^{-2} or branches into two as a result of a selection event with approximate probability σN^{-2}, for an approximate total probability of change $2\lambda_n N^{-2} = n(n - 1 + \rho + \sigma)N^{-2}$. In number of $N^2/2$ time steps in the limit of a large population size, the expected time for a change is λ_n^{-1}. Moreover, in the case of a change caused by a selection event, the incoming lineage is real with probability $1 - c_{AB}$, $1 - c_{Ab}$, $1 - c_{aB}$, or $1 - c_{ab}$ if the type of the individual on the continuing lineage is AB, Ab, aB, or ab, respectively.*

Calculation for Fixation Probability

An ordered sample of n individuals is represented by an n-dimensional vector $\mathbf{z} = (z_1, \ldots, z_n)$, where $z_i = AB$, Ab, aB, or ab, for $i = 1, \ldots, n$. The sample configuration is given by the vector $\mathbf{n} = (n_{AB}, n_{Ab}, n_{aB}, n_{ab})$ with $n_{AB} + n_{Ab} + n_{aB} + n_{ab} = n$.

Let $\mathbf{z}(\tau)$ be an ordered sample of n individuals chosen at random *without replacement* at time step $\tau \geq 0$. The probability distribution of this sample will depend on the ancestral recombination-selection graph from time step τ to time step 0, represented by $G(\tau)$, and the type frequencies at time step 0, given by $\mathbf{x}(0)$. What will ~~actually matter is the topology of the graph~~ from time step τ to time step 0. It is represented by sequence events backward in time written in the form

$$G = (G_1, \quad \ldots \quad, G_{m_\tau}), \tag{41}$$

where m_τ is the total number of events. These are events of coalescence, recombination, or selection in one step back, either pure or multiple.

Let n_G be the number of ancestors after the occurrence of the last event of G backward in time. This last event is assumed to take place at time back τ_G. On the other hand, the time with n_G ancestors is represented by τ_{n_G}. For G to be an admissible ~~topology of the graph~~ from time step τ to time step 0, it is necessary that $\tau_G \leq \tau < \tau_G + \tau_{n_G}$. We define

$$G(\tau) = \{G, \tau_G \leq \tau < \tau_G + \tau_{n_G}\}. \tag{42}$$

Note that τ_G and τ_{n_G} are independent random variables. Moreover, τ_{n_G} is a geometric random variable with parameter p_{n_G}, while τ_G is a sum of independent geometric random variables.

The probability of the event $\mathbf{z}(\tau) = \mathbf{z}$, given $\mathbf{x}(0)$, can be expressed in the form

$$P(\mathbf{z}(\tau) = \mathbf{z}|\mathbf{x}(0)) = \sum_{G(\tau)} P(\mathbf{z}(\tau) = \mathbf{z}|G(\tau), \mathbf{x}(0))P(G(\tau)). \tag{43}$$

The conditional probability in the summand of the above equation does not actually depend on time step τ. Therefore, we define

$$P_G(\mathbf{z}|\mathbf{x}(0)) = P(\mathbf{z}(\tau) = \mathbf{z}|G(\tau), \mathbf{x}(0)). \tag{44}$$

On the other hand, we have

$$P(G(\tau)) = P(G)P(\tau_G \leq \tau < \tau_G + \tau_{n_G}), \tag{45}$$

where

$$P(G) = \prod_{k=1}^{m_\tau} P(G_k), \tag{46}$$

with $P(G_k)$ defined by (29)–(31) for pure events of coalescence, recombination, or selection and by (32) for multiple events. Moreover, we have

$$P(\tau_G \leq \tau < \tau_G + \tau_{n_G}) = P(\tau_G + \tau_{n_G} > \tau) - P(\tau_G > \tau). \tag{47}$$

Note that

$$\sum_{\tau=0}^{\infty} P(\tau_H \leq \tau < \tau_G + \tau_{n_G}) = E(\tau_G + \tau_{n_G}) - E(\tau_G) = E(\tau_{n_G}). \tag{48}$$

Summing over $\tau \geq 0$ in (43) yields the following result.

Proposition 4 *Let $\mathbf{z}(\tau)$ be an ordered sample of n individuals chosen at random without replacement at time step $\tau \geq 0$ and $\mathbf{x}(0)$ be the vector of the individual type frequencies at time step 0. Then we have*

$$\sum_{\tau=0}^{\infty} P(\mathbf{z}(\tau) = \mathbf{z}|\mathbf{x}(0)) = \sum_G P_G(\mathbf{z}|\mathbf{x}(0))P(G)E(\tau_{n_G}), \tag{49}$$

where G is a sequence of pure or multiple events of coalescence, recombination, or selection from time step τ to time step 0, n_G

is the number of ancestors at time step 0, *and* τ_{n_G} *is a time back with this number of ancestors*.

Actually the conditional probability of $\mathbf{z}$, given G and $\mathbf{x}(0)$, in Proposition 4 depends on the types of the n_G ordered ancestors at time step 0, represented by $\mathbf{z}(0)$, so that

$$P_G(\mathbf{z}|\mathbf{x}(0)) = \sum_{\mathbf{z}(0)} P_G(\mathbf{z}|\mathbf{z}(0)) P_G(\mathbf{z}(0)|\mathbf{x}(0)). \tag{50}$$

Moreover, we have

$$P_G(\mathbf{z}|\mathbf{z}(0)) = 0, \tag{51}$$

if $\mathbf{z}$ is incompatible with G and $\mathbf{z}(0)$. Otherwise, this conditional probability is 1 times a product of conditional probabilities of replacement, which is different from 1 only in the case of selection events in G. On the other hand,

$$P_G(\mathbf{z}(0)|\mathbf{x}(0)) = N^{-n_G} (N_{AB}(0))_{n_{AB}(0)} (N_{Ab}(0))_{n_{Ab}(0)} \times (N_{aB}(0))_{n_{aB}(0)} (N_{ab}(0))_{n_{ab}(0)}, \tag{52}$$

where $(N)_n = N \times (N-1) \times \ldots \times (N-n+1)$ denotes a falling factorial, while

$$N\mathbf{x}(0) = (N_{AB}(0),\ N_{Ab}(0),\ N_{aB}(0),\ N_{ab}(0)) \tag{53}$$

and

$$\underline{n}(0) = (n_{AB}(0),\ n_{Ab}(0),\ n_{aB}(0),\ n_{ab}(0)) \tag{54}$$

represent the population configuration at time step 0 and the sample configuration of $\mathbf{z}(0)$, respectively. Moreover, this sample satisfies

$$n_{AB}(0) + n_{Ab}(0) + n_{aB}(0) + n_{ab}(0) = ~~nG~~, \tag{55}$$

with the inequalities

$$\begin{aligned} n_{AB}(0) &\le N_{AB}(0), \\ n_{Ab}(0) &\le N_{Ab}(0), \\ n_{aB}(0) &\le N_{aB}(0), \\ n_{ab}(0) &\le Nab(0), \end{aligned}$$

which are necessary conditions for $\mathbf{z}(0)$ to be compatible with G and $\mathbf{x}(0)$.

Consider, for instance, a sequence of events backward in time, $G = ~~R_2S_3C_4~~$ for an ordered sample of size $n = 2$, as illustrated in Figure 2. Here we have a pure recombination event, a pure selection event, and a pure coalescence event, in this order backward in time. In the case of recombination, one lineage splits into two, a left lineage assumed to be ancestral at locus 1 and a right lineage assumed to be ancestral at locus 2. In the case of selection, one lineage branches into two, a continuing lineage and a new incoming lineage, both potentially ancestral. And last, in the case of coalescence, two lineages merge. The probability of the whole sequence of events is

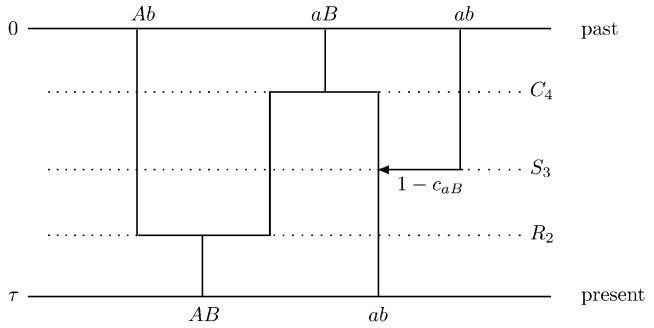

Figure 2 Possible ancestral recombination–selection graph ~~of topology~~ represented by $G = ~~R_2S_3C_4~~$ for two ordered individuals from time step τ to time step 0 to be of types *AB* and *ab* in this order.

$$P(G) = P(R_2)P(S_3)P(C_4). \tag{56}$$

Note that the number of ancestors increases by one following the recombination event and the selection event, but decreases by one following the coalescence event, so that the number of ancestors at the end is $n_G = 3$. The time back to the last event, τ_G, can be expressed in the form

$$\tau_{~~G~~} = \tau_2 + \tau_3 + \tau_4, \tag{57}$$

where τ_2, τ_3, and τ_4 are independent geometric random variables with parameters p_2, p_3, and p_4, respectively. On the other hand, the time back spent with n_G ancestors, τ_{n_G}, is a geometric random variable with parameter p_3.

Finally, given that $G = ~~R_2S_3C_4~~$ and $\tau_G \le \tau < \tau_G + \tau_{~~nG~~}$, the ordered sample $\mathbf{z}(\tau) = (AB, ab)$ occurs with probability $1 - c_{aB}$, if the ancestral state at time step 0 is $\mathbf{z}(0) = (Ab, aB, ab)$. The probability of this initial ancestral state given the initial type frequencies is

$$P(\mathbf{z}(0) = (Ab, aB, ab) | \mathbf{x}(0)) = x_{Ab}(0) x_{aB}(0) x_{ab}(0) (1 + O(N^{-1})). \tag{58}$$

Of course, we have to consider all possible initial ancestral states for this particular G and then all possible G for this particular ordered sample.

Approximation Results

We are now ready to approximate the probability of ultimate fixation of A under the assumptions that the population size is large and the population-scaled recombination and selection parameters are small.

Consider $\mathbf{z} = (z_1, z_2)$, where $z_1 = AB$ or Ab and $z_2 = aB$ or ab. First note that

$$E(x_{z_1}(\tau) x_{z_2}(\tau) | \mathbf{x}(0)) = (1 - N^{-1}) P(\mathbf{z}(\tau) = \mathbf{z} | \mathbf{x}(0)), \tag{59}$$

so that Proposition 4 leads to the expression

$$E_{\mathbf{z}}(\mathbf{x}(0)) = (1 - N^{-1}) \sum_G P_G(\mathbf{z}|\mathbf{x}(0)) P(G) E(T_{n_G}), \tag{60}$$

where $T_{n_G} = 2\tau_{n_G} N^{-2}$, for the quantity defined in Proposition 2.

Let $|G|$ denote the minimum ~~of~~ number of ancestors along a sequence of events G for the ordered sample $\mathbf{z} = (z_1, z_2)$. Note that

$$P_G(\mathbf{z}|\mathbf{x}(0)) = 0, \tag{61}$$

if $|G| = 1$, since alleles A in z_1 and a in z_2 cannot have the same ancestor, while

$$P_G(\mathbf{z}|\mathbf{x}(0)) \le x_A(0), \tag{62}$$

if $|G| \ge 2$, since allele A in z_1 must be present in at least one ancestor. This leads to the inequality

$$E_{\mathbf{z}}(\mathbf{x}(0)) \le x_A(0) E(W_2), \tag{63}$$

where

$$E(W_2) = \sum_{\{G:|G|\ge 2\}} P(G) E(T_{n_G}). \tag{64}$$

Actually, W_2 is the time back in number of $N^2/2$ time steps for the number of ancestors in the ancestral graph starting from a sample of size 2 to reach one for the first time. This occurs when the most recent ultimate ancestor (MRUA) is found. It can be shown that $E(W_2)$ is finite and bounded by a constant that does not depend on N. This is also true for $E(W_n) \ge E(W_{n-1})$, which is defined analogously for a sample of any size $n \ge 3$. (See *Appendix*.)

Now, suppose that the population-scaled recombination and selection parameters, ρ and σ, are small and of the same order of magnitude, so that $\rho = d\sigma << 1$ for some constant $d > 0$. Let n_G^+ designate the sum of all increases in the number of ancestors along a sequence of events G for the ordered sample $\mathbf{z} = (z_1, z_2)$. If $n_G^+ > k$, then

$$G = (G^{(1)}, G^{(2)}), \tag{65}$$

where $G^{(1)}$ is a sequence of events such that $n^+_{G^{(1)}} > k$ and $n_{G^{(1)}} \le k+2$. The number $n_{G^{(1)}}$ is the sample size at the beginning of $G^{(2)}$ just after the first increase in the number of ancestors that brings this number above k. Note that there is a finite number of $G^{(1)}$ satisfying these conditions and that the probability of each one can be neglected compared to σ^k; that is,

$$P(G^{(1)}) = O(\sigma^{k+1}). \tag{66}$$

On the other hand,

$$\sum_{\{G^{(2)}:|G^{(2)}|\ge 2\}} P(G^{(2)}) E(T_{n_{G^{(2)}}}) \le E(W_{k+2}), \tag{67}$$

which is a finite bound. Since $P(G) = P(G^{(1)})P(G^{(2)})$ and $n_G = n_{G^{(2)}}$, we conclude that

$$\sum_{\{G:|G|\ge 2, n_G^+ > k\}} P_G(\mathbf{z}|\mathbf{x}(0)) P(G) E(T_{n_G}) \le x_A(0) O\left(\sigma^{k+1}\right). \tag{68}$$

This gives the order of the error in the following approximation result.

Proposition 5 *Ignoring terms of order* ~~N^{-1}~~$O(\sigma^l \rho^m)$ *for* $l + m \ge k + 1$ *where* $\rho = rN$ *and* $\sigma = sN$ *are the population-scaled parameters for recombination and selection, respectively, the expected times in the probability of ultimate fixation of A given in Proposition 2 are approximated by*

$$E_{\mathbf{z}}(\mathbf{x}(0)) \approx \sum_{\{G:|G|\ge 2, n_G^+ \le k\}} P_G(\mathbf{z}|\mathbf{x}(0)) P(G) E(T_{nG}), \tag{69}$$

for $\mathbf{z} = (z_1, z_2)$, *with* $z_1 = AB, Ab$ *and* $z_2 = aB, ab$, *where all terms in the summation, given by* (40), (46), and (50), *are approximated by their leading terms in the case of a large population size. Here, the summation is over all sequences G of pure coalescence, recombination, or selection events backward in time with at most k pure recombination or selection events and a number of ancestors always larger than two with final value* n_G.

Note that the coefficient of $\sigma^l \rho^m$ for $l + m \le k$ in Proposition 5 is obtained by considering all sequences of events G involving up to l pure selection events and m pure recombination events.

Using MATHEMATICA and (69), a polynomial of degree k with respect to σ and ρ approximating the quantity (23) in Proposition 2 for σ and ρ small enough can be calculated. This approach leads to the main results of this article.

Proposition 6 *Consider the discrete-time Moran model with small population-scaled recombination fraction* $\rho = rN$ *and small population-scaled intensity of selection* $\sigma = sN$ *with coefficients of selection* $0 \le c_{AB}, c_{Ab}, c_{aB}, c_{ab} \le 1$ *such that* $\varepsilon = c_{AB} - c_{Ab} - c_{aB} + c_{ab} \ne 0$. *Given the initial conditions in Figure 1 with* $x_A(0) = N^{-1}$ *and ignoring terms of order* $N^{-1}O(\sigma^l \rho^m)$ *for* $l + m \ge 4$, *the probability of ultimate fixation of A is approximated by*

$$\begin{aligned} u_A(\mathbf{x}(0)) \approx \frac{1}{N} &+ \frac{\sigma}{2N}(c_{Ab} - c_{ab} + \varepsilon x) \\ &+ \frac{\sigma^2}{12N}\Big((c_{Ab} - c_{ab})^2 \\ &\qquad + \varepsilon x(c_{AB} - c_{aB} + c_{Ab} - c_{ab} \\ &\qquad + 2(1-x)(c_{aB} - c_{ab}))\Big) \\ &- \frac{\sigma^3}{24N} x(1-x)(c_{aB} - c_{ab})^2 (c_{Ab} - c_{ab} + \varepsilon x) \\ &+ \frac{\rho\sigma^2 \varepsilon}{432N} x(1-x)\big(3(c_{AB} - c_{Ab}) + 2(c_{aB} - c_{ab})\big). \end{aligned} \tag{70}$$

Proposition 7 *Under the conditions of Proposition 6 but in the case where the coefficients of selection satisfy* $\varepsilon = c_{AB} - c_{Ab} - c_{aB} + c_{ab} = 0$ *and terms of order* $N^{-1}O(\sigma^l \rho^m)$ *for* $l + m \ge 5$ *are ignored, the probability of ultimate fixation of A is approximated by*

$$u_A(x(0)) \approx \frac{1}{N} + \frac{\sigma}{2N}(c_{Ab} - c_{ab}) + \frac{\sigma^2}{12N}(c_{Ab} - c_{ab})^2 - \frac{\sigma^3}{24N}x(1-x)(c_{Ab} - c_{ab})(c_{aB} - c_{ab})^2 - \frac{\sigma^4}{720N}(c_{Ab} - c_{ab})\Big((c_{Ab} - c_{ab})^3 + (c_{aB} - c_{ab})^2 x(1-x)(8(c_{Ab} - c_{ab}) + 7(c_{aB} - c_{ab})(1-2x))\Big) + \frac{19\rho\sigma^3}{432N}x(1-x)(c_{Ab} - c_{ab})(c_{aB} - c_{ab})^2. \quad (71)$$

Discussion

Effect of selection at linked loci on the probability of fixation of a single mutant allele

The conditional expected change in the frequency of an allele A in a two-locus two-allele Moran model from one time step to the next as expressed in (15) can be interpreted [see (13)] as the current frequency of A in the proportion of the population chosen to be replaced (here, N^{-1}) times its average excess in fitness (Fisher 1930). In this average excess, the differences between the coefficients of selection of A-bearing and a-bearing individuals have relative weights given by the products of the frequencies of the individual types. Proposition 2 says that the difference between the probability of ultimate fixation of A and its initial frequency takes the same form with weights given by expected times with pairs of individual types over all time steps. This corresponds to a projected average excess in fitness (Lessard and Lahaie 2009).

The expected times in Proposition 2 depend on the population-scaled parameters, $\sigma = sN$ and $\rho = rN$, for the intensity of selection and the recombination fraction, respectively. As shown in Proposition 5, expansions in $\sigma^l\rho^m$ for $l + m \leq k$ and a population size N large enough are obtained by considering up to k pure recombination or selection events in the ancestral graph of pairs of individuals chosen at random.

This has been applied to get analytical approximations for the probability of ultimate fixation of a single mutant A introduced at random into a population in which a previous mutant B is segregating at another locus and has reached some frequency x. Here we make the assumptions of weak selection ($\sigma << 1$), tight linkage ($\rho << 1$), and large population size ($N >> 1$). In the case of positive or negative epistasis, the first-order effect of recombination on the fixation probability is of order $\rho\sigma^2$. It is detected as soon as one pure recombination event and one pure selection event are considered. This is best understood from the leading recombination terms when the single mutant A is initially linked to B and LD is positive, which occurs with probability x, and when it is the opposite, which occurs with the complementary probability $1 - x$. In the case of coefficients of selection given by $c_{ab} = 0$, $c_{aB} = c_{Ab} = c$, $c_{AB} = 1$ with $0 \leq c \leq 1$, so that A and B are equally advantageous, and epistasis given by $\varepsilon = 1 - 2c$, these leading terms are approximated by

$$L_B(\rho) \approx -\frac{\rho\sigma}{24N}(1-x)(3+\varepsilon) + \frac{\rho\sigma^2}{864N}(1-x)(x(1-\varepsilon)(24+5\varepsilon) + 3\varepsilon(2+\varepsilon) + 27)$$

$$L_b(\rho) \approx \frac{\rho\sigma}{24N}x(3+\varepsilon) - \frac{\rho\sigma^2}{864N}x(x(1-\varepsilon)(24+5\varepsilon) + \varepsilon(1+2\varepsilon) + 27),$$

respectively, which yields

$$L(\rho) = xL_B(\rho) + (1-x)L_b(\rho) \approx \frac{\rho\sigma^2}{864N}x(1-x)\varepsilon(5+\varepsilon),$$

for the weighted average in agreement with Proposition 6. We see that the primary effect of increasing the recombination rate is to increase the fixation probability in the case of initial NLD, but the opposite happens in the case of initial PLD. In the case of initial average LE, the primary effects cancel out, while the weighted average of the secondary effects, obtained by taking into account one pure selection event in addition to one pure recombination event, is of the same sign as epistasis. In the absence of epistasis ($\varepsilon = 1 - 2c = 0$), tertiary effects obtained by taking into account a second pure selection event have to be considered, and their weighted average is approximated by

$$L(\rho) \approx \frac{19\rho\sigma^3}{3456N}x(1-x),$$

according to Proposition 7. This expression is always positive and increases with the recombination rate. This confirms the HR effect when both mutants are advantageous and epistasis is absent (Hill and Robertson 1966).

The following explanation for the HR effect has been given in Barton and Otto (2005, p. @@): "...beneficial alleles that arise in coupling rise rapidly and fix within the population, leading to the disappearance of the positive disequilibrium. The negative disequilibrium persists for a much longer period of time, until one or the other allele becomes fixed. Therefore, with multiplicative selection, the variance in disequilibrium present in the first generation ultimately leads to negative disequilibrium, on average." It is the asymmetric action of selection upon the initial positive and negative disequilibria and further disequilibria generated by random genetic drift that is responsible for the accumulation of negative disequilibrium, on average. This in turn provides an evolutionary advantage to recombination. If this is true in the case of AE, then it should also be true in the case of PE, which enhances the effect of selection. Actually, the average leading recombination term in the fixation probability for A that takes into account two pure selection events in the presence of epistasis is given by

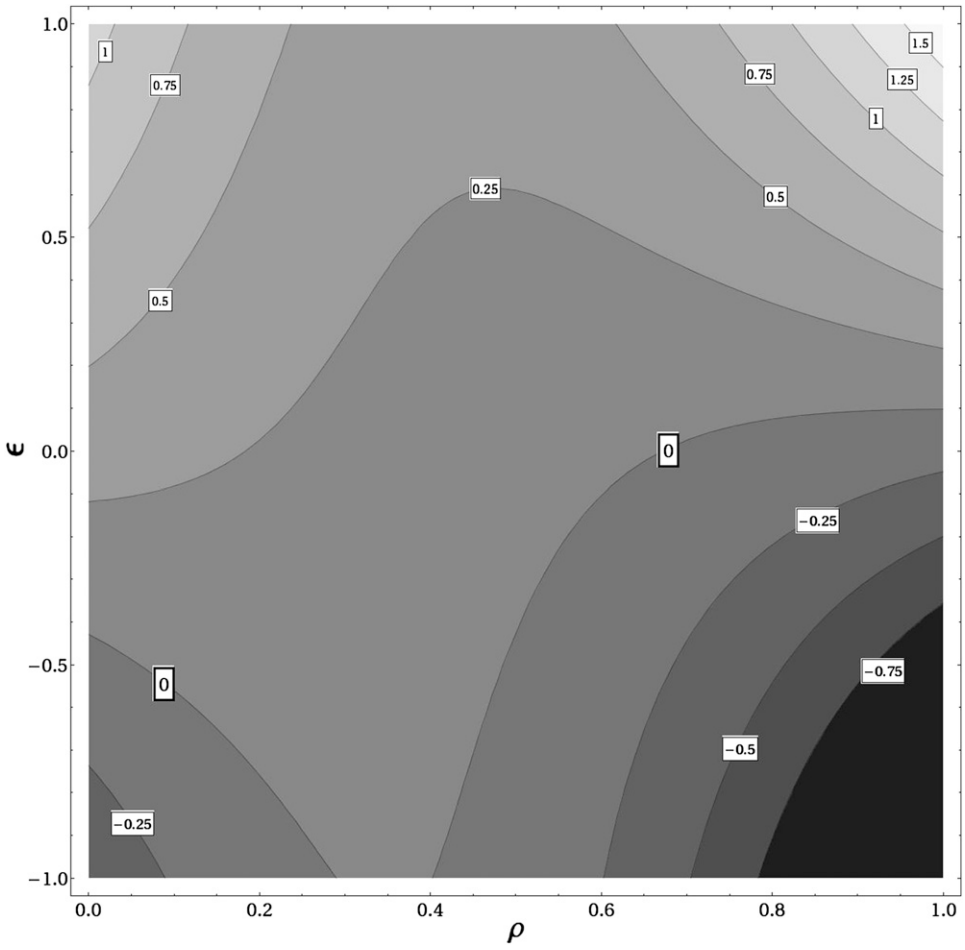

Figure 3 Contour plot of the derivative of the probability of ultimate fixation of *A* with respect to ρ, based on analytical approximations obtained by considering at most four selection or recombination events in the ancestry of samples of size 2. We consider the case where mutations are advantageous and mortalities of *ab*, *aB*, *Ab*, and *AB* are given by 1, $1 - c\sigma N^{-1}$, $1 - c\sigma N^{-1}$, and $1 - \sigma N^{-1}$, respectively. In this case $\varepsilon = 1 - 2c$. Here we assume that $\sigma = 0.5$ and $x = 0.1$. The values of the derivative are expressed in $(100N)^{-1}$ units. Regions with negative values indicate the parameter set for ε and ρ, where increasing recombination reduces the fixation probability.

$$L(\rho) \approx \frac{\rho}{17,280N} x(1-x)(95\sigma + 100\varepsilon + (51 - 44x)\varepsilon\sigma + 20\varepsilon^2 - (45 - 34x)\varepsilon^2\sigma - 5(1 - 2x)\varepsilon^3\sigma).$$

For σ and ε small enough, this expression for the average leading recombination term is positive if $\varepsilon > -0.95\sigma$. More generally, it has been checked that this leading term when at most four pure recombination or selection events are taken into account is positive if ε is larger than some negative value (see Figure 3).

Therefore, our analytical approximations indicate that recombination increases the probability of ultimate fixation of beneficial mutants in the case of positive epistasis or weak negative epistasis, at least under weak enough selection. This is supported by simulations (see Figure 4A for $\varepsilon = 1$). Even in the case of strong negative epistasis, an increase in the recombination fraction may increase the fixation probability in a range of values of the recombination fraction (see Figures 3 and 4B for $\varepsilon = -1$). This effect, though small, should provide some evolutionary advantage to recombination. This suggests that random drift may be an important factor in the evolution of recombination in the presence of selection.

Since negative epistasis is generally considered to be favorable to recombination in an infinite population, our conclusion that positive epistasis or weak negative epistasis, at least under weak selection, is a condition for recombination to help ultimate fixation of beneficial mutants may appear surprising. Some numerical results for approximations based on branching processes (Barton 1995a,b) and simulations for modifiers of the recombination fraction (Otto and Barton 2001) obtained under the assumption of stronger selection ($sN >> 1$) suggested that this could be the case. However, this seems to have been little noticed up to now.

Note also that AE in an exact finite population is modeled by multiplicative gene action. If advantageous mutants act additively and selection is weak, then epistasis is negative and weak. This is the situation that was actually simulated in Hill and Robertson (1966). However, comparisons with results obtained under multiplicative selection showed no significant differences.

Conditions on the selection coefficients for recombination to be favored

In the presence of epistasis and assuming that allele *B* is advantageous, so that the coefficients of selection satisfy $c_{AB} - c_{Ab} > 0$ and $c_{aB} - c_{ab} > 0$, the coefficient of $\rho\sigma^2$ in

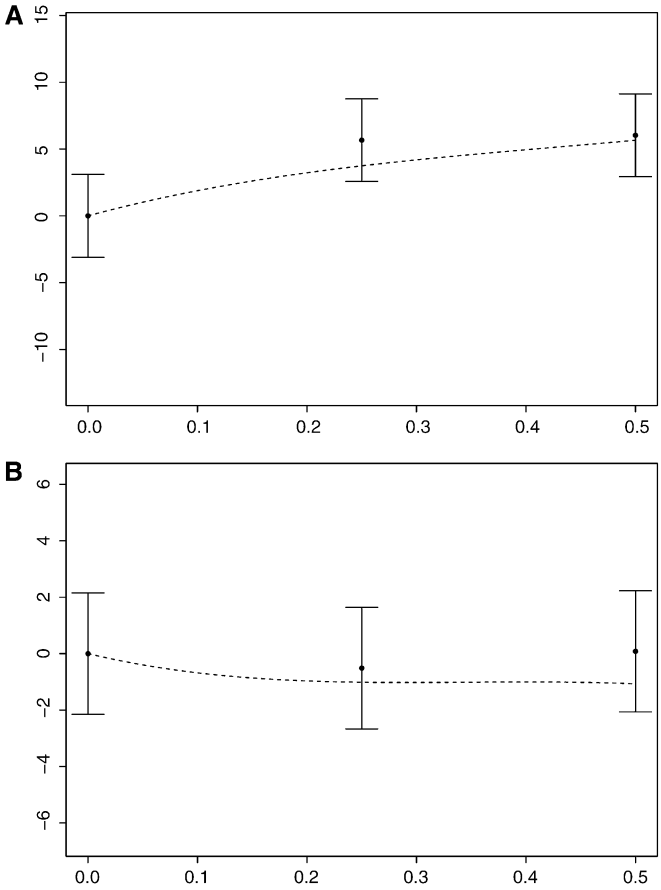

Figure 4 (A and B) Simulation results for the effect of recombination on the probability of ultimate fixation of A assuming that mutations are advantageous with mortalities given by 1, $1 - c\sigma N^{-1}$, $1 - c\sigma N^{-1}$, and $1 - \sigma N^{-1}$ for ab, aB, Ab, and AB, respectively. The values on the y-axis show the mean difference between the number of times (in a block of 5×10^6 runs) that A went to fixation compared to the case of no recombination. Means and 95% confidence intervals are calculated from 4×10^4 blocks of runs for the case of positive epistasis (A) and from 10^5 trials for the case of negative epistasis (B). The dashed line shows the theoretical prediction based on our analytical approximations of the fixation probability. In this simulation $N = 50$; $x = 0.1$; $\sigma = 0.5$; and $\rho = 0.0$, 0.25, and 0.5.

the probability of ultimate fixation of A given in Proposition 6, which is the leading recombination term under weak selection and low recombination, is positive if and only if $\varepsilon = c_{AB} - c_{Ab} - c_{aB} + c_{ab} > 0$; that is, $c_{AB} - c_{Ab} > c_{aB} - c_{ab} > 0$. This means that A enhances the beneficial effect of B. Under these circumstances, the probability of ultimate fixation of A increases as the recombination rate increases. Note, however, that this does not require that allele A itself is beneficial. It is favored to go to fixation as a result of a hitchhiking effect.

In the absence of epistasis, that is when $\varepsilon = 0$, an approximation to the next order is necessary to detect the effect of recombination on the probability of ultimate fixation of A. Then the leading recombination term is given by the term in $\rho\sigma^3$ in Proposition 7, whose coefficient is positive if A is advantageous, so that $c_{Ab} - c_{ab} = c_{AB} - c_{aB} > 0$. Under this condition, an increase of the recombination rate results in an increase of the fixation probability. This is in agreement with the HR effect when both mutants are advantageous (Hill and Robertson 1966). Note, however, that allele B does not have to be beneficial, since the sign of $c_{aB} - c_{ab} = c_{AB} - c_{Ab}$ does not matter.

In both cases, the effect of recombination is more pronounced when the differences in the coefficients of selection are larger and when the frequency of B is closer to $x = \frac{1}{2}$.

The situation with deleterious mutants is comprised in the above discussion, since an allele is deleterious if and only the alternate allele is advantageous. Therefore, the probability of ultimate fixation of an allele A decreases as the recombination rate increases if this allele either is in positive epistatic interaction with a deleterious allele B or is itself deleterious in the case of no epistasis with B. This explains how recombination can reduce the rate at which the ratchet-like mechanism suggested by Muller (1964) slows down the rate of evolution.

Comparisons with previous results

In the absence of selection, the expected allele frequencies do not change and the probability of fixation of a gamete can be obtained from a transformation of a transition matrix (Karlin and McGregor 1968).

In the case where alleles B and b are neutral at locus 2 while alleles A and a are under selection at locus 1, so that $c_{AB} - c_{Ab} = c_{aB} - c_{ab} = 0$ and $c_{AB} - c_{aB} = c_{Ab} - c_{ab} = 1$, there is no epistasis ($\varepsilon = 0$). It follows from (71) that the probability of ultimate fixation of A when its initial frequency is the inverse of the population size, that is N^{-1}, is approximated by

$$u_A(N^{-1}) = N^{-1}\left(1 + \frac{\sigma}{2} + \frac{\sigma^2}{12} - \frac{\sigma^4}{720} + O(\sigma^5)\right). \qquad (72)$$

It is easy to see that this is consistent with a Taylor expansion around $\sigma = 0$ of the formula

$$u_A(N^{-1}) = \frac{1 - e^{-\sigma N^{-1}}}{1 - e^{-\sigma}}, \qquad (73)$$

which is predicted from a diffusion approximation for allele frequencies at a single locus (Kimura 1957, 1962).

When selection acts on both loci, it is possible to use the Kolmogorov forward or backward diffusion equation for two-locus gamete frequencies (Kimura 1955) to approximate the probability of ultimate fixation of a gamete and then that of an allele. The approximations obtained in this way up to the first-order effect of selection (Hill and Robertson 1966; Ohta 1968) are consistent with Propositions 6 and 7.

Lehman and Rousset (2009) expressed the probability of ultimate fixation as a sum of expected changes in frequencies in an exact Wright–Fisher model and then considered

a Taylor expansion of this expression with respect to the intensity of selection. The coefficients in this expansion were obtained from a backward approach in a neutral model and symbolic calculation. An interpretation makes a clever use of matrix theory. Their approximation (A.34) for the reduction in the probability of ultimate fixation of A due to interference in the case of no epistasis is consistent with Proposition 7 with the correspondences $2s_A = (c_{Ab} - c_{ab})\sigma N^{-1}$, $2s_B = (c_{aB} - c_{ab})\sigma N^{-1}$, $2r = \rho N^{-1}$, and $p_B(0) = x$, ignoring terms of order $O(N^{-2})$. Therefore, the Moran and Wright–Fisher models lead to the same results after appropriate rescaling in the limit of a large population size.

One of the main interests in the approach based on the ancestral recombination–selection graph presented in this article is that the term in $\sigma^l\rho^m$ in the fixation probability is known to come up when considering at most $l - 1$ selection events and m recombination events in the genealogy of ordered sampled gametes backward in time to compute expected times in given states. The approach has been applied to the probability of ultimate fixation of a single mutant allele at one locus in initial linkage equilibrium with another segregating locus. It could be applied as well to an allele or a gamete given any initial conditions, *e.g.*, a double mutant given an initial single double mutant or two initial single mutants with mutants being deleterious when alone but beneficial when coupled as in simulations by Michalakis and Slatkin (1996). The results, however, would involve high-order polynomials with respect to the initial frequencies in the case of general initial conditions.

Implementation of the approach

The approximations given in Propositions 6 and 7 for the fixation probability rely on expected times in given ordered-sample states. These expected times are computed by conditioning on events of coalescence, recombination, or selection in the ancestry of the sample. Note that the calculation time can be shortened by considering only the material that is potentially ancestral to the original sample at either of the loci at every state change backward in time. Therefore, if the current number of ancestors at only one locus is n_1 and that of ancestors at both loci is n_2, then the total rate of change is

$$\lambda_n = \frac{n(n-1+\sigma) + n_2\rho}{2}, \tag{74}$$

where $\mathbf{n} = (n_1, n_2)$ with $n = n_1 + n_2$. This is the case since a recombination event on an ancestral sequence will change the state of the ancestral material only if the sequence is ancestral at both loci.

Extensions to other models

The model considered in this article is a particular discrete-time viability model of the Moran type (Moran 1958) expressed in terms of probabilities of mortality that are linear functions with respect to some intensity of selection. This model was used to make as clear as possible the main ideas of the approach and to simplify as far as possible the rigorous justifications of the approximations. In the case of relative mortalities, for instance, the probabilities of replacement would generally be functions of any order with respect to the intensity of selection. Then Propositions 1 and 2 for the expected change in frequency and fixation probability for an allele would give only approximations up to terms of order $O(N^{-2})$. This would introduce technical details in the limit of a large population size. Note, however, that a fertility model with the probability of replacement depending on the type of the offspring produced instead of the type of the individual chosen to be replaced could be analogously treated.

The Wright–Fisher model would introduce other kinds of difficulties. These are related to the possible number of ancestors in the exact ancestral recombination–selection graph and the expected time to reach the most recent ultimate ancestor. The analysis presented in the *Appendix* to justify the approach would have to be refined.

It is relatively easy to understand that the fixation probability in a finite population can be expressed in terms of expected times. It is less obvious to establish the corresponding result in the limiting process for an infinite population. Actually, (69) and (40) show that

$$E_{z_1,z_2}(x(0)) \approx \int_0^\infty E(x_{z_1}(t)x_{z_2}(t)|\mathbf{x}(0))dt, \tag{75}$$

for $z_1 = AB$, Ab and $z_2 = aB$, ab, with time t measured in number of $N^2/2$ time steps, for a large enough population size N and small enough population-scaled parameters ρ and σ for recombination and selection, respectively. This suggests that the fixation probability corresponding to Proposition 2 in the limiting AIG is given by the same expression but with integrals instead of sums.

Like those for the coalescent (Kingman 1982), the results obtained for the Moran model in the limit of a large population size are expected to be valid for a wide range of exchangeable models. This could be established by considering an exact model extending the Cannings neutral model (Cannings 1974) to take into account not only selection (Lessard and Ladret 2007) but also recombination in the limit of a large population size. It could also be done by showing that the fixation probability obtained from a Kolmogorov backward or forward diffusion equation for gamete frequencies (Kimura 1955; Hill and Robertson 1966; Ohta 1968) takes the form given in Proposition 2 with expected values computed according to (75) in the corresponding AIG.

Acknowledgments

We are most grateful to one of the reviewers for the suggestion that there must be some small minimum negative value of epistasis above which recombination should

help fixation of beneficial mutants. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Literature Cited

Barton, N. H., 1995a A general model for the evolution of recombination. Genet. Res. 65: 123–144.

Barton, N. H., 1995b Linkage and the limits to natural selection. Genetics 140: 821–841.

Barton, N. H., and S. P. Otto, 2005 Evolution of recombination due to random drift. Genetics 169: 2353–2370.

Bodmer, W. F., 1970 The evolutionary significance of recombination in prokaryotes. Symp. Soc. Gen. Microbiol. 20: 279–294.

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. Adv. Appl. Probab. 6: 260–290.

Crow, J. F., and M. Kimura, 1965 Evolution in sexual and asexual populations. Am. Nat. 99: 439–450.

Crow, J. F., and M. Kimura, 1969 Evolution in sexual and asexual populations: a reply. Am. Nat. 103: 89–91.

Donnelly, P., and T. G. Kurtz, 1999 Genealogical processes for Fleming-Viot models with selection and recombination. Ann. Appl. Probab. 9: 1091–1148.

Eshel, I., and M. W. Feldman, 1970 On the evolutionary effect of recombination. Theor. Popul. Biol. 1: 88–100.

Ewens, W. J., 1990 Population genetics theory—the past and the future, pp. 177–227 in *Mathematical and Statistical Developments of Evolutionary Theory* (NATO ASI Series C: Mathematical and Physical Sciences, Vol. 299), edited by S. Lessard. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Fearnhead, P., 2003 Ancestral processes for non-neutral models of complex diseases. Theor. Popul. Biol. 63: 115–130.

Fearnhead, P., R. M. Harding, J. A. Schneider, M. Myers, and P. Donnelly, 2004 Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. Genetics 167: 2067–2081.

Feldman, M. W., and U. Liberman, 1986 An evolutionary reduction principle for genetic modifiers. Proc. Natl. Acad. Sci. USA 83: 4824–4827.

Feldman, M. W., F. B. Christiansen, and L. D. Brook, 1980 Evolution of recombination in a constant environment. Proc. Natl. Acad. Sci. USA 77: 4838–4841.

Felsenstein, J., 1965 The effect of linkage on directional selection. Genetics 52: 349–363.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics 78: 737–756.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3: 479–502.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution* (IMA Volumes in Mathematics and Its Applications, Vol. 87), edited by P. Donnelly, and S. Tavaré. Springer-Verlag, Berlin.

Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet. Res. 8: 269–294.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics 120: 831–840.

Karlin, S., 1973 Sex and infinity: a mathematical analysis of the advantages and disadvantages of genetic recombination, pp. 155–194 in *The Mathematical Theory of the Dynamics of Biological Populations*, edited by M. S. Bartlett and R. W. Hiorns. Academic Press, London.

Karlin, S., and J. McGregor, 1968 Rates and probabilities of fixation for two locus random mating finite populations. Genetics 58: 141–159.

Kimura, M., 1955 Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. 20: 33–55.

Kimura, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. 28: 882–901.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics 47: 713–719.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. 13: 235–248.

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. Popul. Biol. 51: 210–237.

Larribe, F., and P. Fearnhead, 2011 Composite likelihood methods in statistical genetics. Stat. Sin. 21: 43–69.

Larribe, F., and S. Lessard, 2008 A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci. Stat. Appl. Genet. Mol. Biol. 7 (1): 27.

Lehman, L., and F. Rousset, 2009 Perturbation expansions of multilocus fixation probabilities for frequency-dependent selection with applications to the Hill-Robertson effect and to the joint evolution of helping and punishment. Theor. Popul. Biol. 76: 35–51.

Lenormand, T., and S. P. Otto, 2000 The evolution of recombination in a heterogeneous environment. Genetics 156: 423–438.

Lessard, S., and V. Ladret, 2007 The probability of fixation of a single mutant in an exchangeable selection model. J. Math. Biol. 54: 721–744.

Lessard, S., and P. Lahaie, 2009 Fixation probability with multiple alleles and projected average allelic effect on selection. Theor. Popul. Biol. 75: 266–277.

Maynard Smith, J., 1968 Evolution in sexual and asexual populations. Am. Nat. 102: 469–673.

McVean, G., P. Awadalla, and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160: 1231–1241.

Michalakis, Y., and M. Slatkin, 1996 Interaction of selection and recombination in the fixation of negative-epistatic genes. Genet. Res. Camb. 67: 257–269.

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

Moran, P. A. P., 1958 Random processes in genetics. Proc. Camb. Philos. Soc. 54: 60–71.

Muller, H. J., 1932 Some genetic aspects of sex. Am. Nat. 66: 118–138.

Muller, H. J., 1964 The relation of recombination to mutational advance. Mutat. Res. 1: 2–9.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics 145: 519–534.

Ohta, T., 1968 Effect of initial linkage disequilibrium and epistasis on fixation probability in a small population, with two-segregating loci. Theor. Appl. Genet. 38: 243–248.

Otto, S. P., and N. H. Barton, 1997 The evolution of recombination: removing the limits to natural selection. Genetics 147: 879–906.

Otto, S. P., and N. H. Barton, 2001 Selection for recombination in small populations. Evolution 55: 1921–1931.

Otto, S. P., and M. W. Feldman, 1997 Deleterious mutations, variable epistatic interactions and the evolution of recombination. Theor. Popul. Biol. 51: 134–147.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902.

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36: 1116–1125.

Rousset, F., 2003 A minimal derivation of convergence stability measures. J. Theor. Biol. 221: 665–668.

Stephens, M., and P. Donnelly, 2003 Ancestral inference in population genetics models with selection. Aust. N. Z. J. Stat. 45: 395–430.

Wakeley, J., and O. Sargsyan, 2009 The conditional ancestral selection graph with strong balancing selection. Theor. Popul. Biol. 75: 355–364.

Yu, F., and A. Etheridge, 2010 The fixation probability of two competing beneficial mutations. Theor. Popul. Biol. 78: 36–45.

Communicating editor: L. M. Wahl

Appendix: Bound for the Expected Time to Reach the Most Recent Ultimate Ancestor

Consider the Markov chain describing the number of ancestors in the ancestral recombination–selection graph backward in time for an ordered sample in a population of size N. For $\sigma = sN$ and $\rho = rN$ small enough, the probabilities of transition $p_{i,j}$, from i to j for $i, j = 1, \ldots, N$, satisfy the inequalities

$$p_{i,i-1} \geq \frac{i(i-1)}{N^2}\left(1-\frac{\sigma}{N}\right)\left(1-\frac{\rho}{N}\right) \geq \frac{i(i-1)}{2N^2}, \tag{A1}$$

$$p_{i,i+1} \leq \frac{i(\sigma+\rho)}{N^2} \leq \frac{i}{12N^2}, \tag{A2}$$

and

$$p_{i,i+2} \leq \frac{i\sigma\rho}{N^3} \leq \frac{i}{12N^2}. \tag{A3}$$

The lower bound comes from the probability of a pure coalescence event, and the upper bounds come from the probability of a recombination event and/or a selection event. Moreover, we have

$$p_{i,i} = 1 - p_{i,i-1} - p_{i,i+1} - p_{i,i+2}, \tag{A4}$$

since $p_{i,j} = 0$ if $j < i - 1$ or $j > i + 2$.

The sojourn time in state $i \geq 2$ in number of time steps is a geometric random variable of parameter $1 - p_{i,i}$, whose expected value is

$$\left(1-p_{i,i}\right)^{-1} \leq \frac{2N^2}{i(i-1)} \leq N^2. \tag{A5}$$

In number of $N^2/2$ time steps, the upper bound is 2. Therefore, the time in number of $N^2/2$ time steps to reach state 1 for the first time from state $i \geq 2$, denoted by W_i, has an expected value satisfying

$$E(W_i) \leq 2E(M_i), \tag{A6}$$

where M_i is the number of state changes before reaching state 1 for the first time from state i.

Given that the chain leaves state $i \geq 2$, the conditional transition probabilities satisfy

$$q_{i,i+1} = \frac{p_{i,i+1}}{1-p_{i,i}} \leq \frac{2}{12(i-1)} \leq \frac{1}{6}, \tag{A7}$$

$$q_{i,i+2} = \frac{p_{i,i+2}}{1-p_{i,i}} \leq \frac{2}{12(i-1)} \leq \frac{1}{6}, \tag{A8}$$

and

$$q_{i,i-1} = \frac{p_{i,i-1}}{1-p_{i,i}} = 1 - q_{i,i+1} - q_{i,i+2} \geq \frac{2}{3}. \tag{A9}$$

Let $\{X_n\}_{n\geq 0}$ be a Markov chain on the integers $i \geq 1$ with $q_{i,j}$ above as transition probabilities for $i \geq 2$ and state 1 absorbing. Such a Markov chain can be constructed from a sequence of independent random variables $\{U_n\}_{n\geq 1}$, each one being uniformly distributed on (0, 1]. Given $X_{n-1} = i \geq 2$, the nth increment for $n \geq 1$ is defined as

$$X_n - X_{n-1} = 2I_{(0,q_{i,i+2}]}(U_n) + I_{(q_{i,i+2},1-q_{i,i-1}]}(U_n) - I_{(1-q_{i,i-1},1]}(U_n), \tag{A10}$$

where $I_{(a,b]}(u) = 1$ if $a < u \leq b$ and 0 otherwise. Analogously, given $Y_{n-1} = i \geq 2$, define

$$Y_n - Y_{n-1} = 2I_{(0,1/6]}(U_n) + I_{(1/6,1/3]}(U_n) - I_{(1/3,1]}(U_n). \tag{A11}$$

In both cases, the increment is 0 if the current state is 1. Moreover, given $X_0 = Y_0 = i \geq 2$, we have

$$X_n = X_0 + \sum_{k=1}^{n}(X_n - X_{n-1}) \leq Y_0 + \sum_{k=1}^{n}(Y_n - Y_{n-1}) = Y_n, \tag{A12}$$

for $n \geq 1$. In particular,

$$P(X_n > 1 | X_0 = i) \leq P(Y_n > 1 | Y_{0=i}), \tag{A13}$$

from which

$$E(M_i) = \sum_{n=0}^{\infty} P(X_n > 1 \mid X_0 = i) \leq \sum_{n=0}^{\infty} P(Y_n > 1 \mid Y_0 = i) = E(N_i). \tag{A14}$$

Here, N_i represents the number of transitions for the Markov chain $\{Y_n\}_{n\geq 0}$ to reach state 1 for the first time from state $i \geq 2$.

The Markov chain $\{Y_n\}_{n\geq 0}$ has transition probabilities $\frac{1}{6}$, $\frac{1}{6}$, and $\frac{2}{3}$ from state i to states $i + 2$, $i + 1$, and $i - 1$, respectively, for $i \geq 2$. State 1 is an absorbing state. Reaching state 1 from state $i \geq 2$ in $\{Y_n\}_{n\geq 0}$ corresponds to extinction in a Galton–Watson process $\{Z_n\}_{n\geq 0}$ starting with $Z_0 = i - 1$ individuals, each individual leaving independently of all others 2, 1, or 0 offspring with probability $\frac{1}{6}$, $\frac{1}{6}$, or $\frac{2}{3}$, respectively. Extinction occurs with probability 1, since the expected number of offspring per individual is $\gamma = \frac{1}{2} < 1$. Moreover,

$$E(N_i) = E\left(\sum_{n=1}^{\infty} Z_n\right) = \sum_{n=1}^{\infty} E(Z_n) = \sum_{n=1}^{\infty} \gamma^n E(Z_0) = \frac{(i-1)\gamma}{1-\gamma}. \tag{A15}$$

We conclude that ~~that~~ $E(W_i) \leq 2E(N_i) < \infty$.

Genetics February (2012)
Author query sheet Lessard (GEN_136309)

Please confirm your final decision on Open Access. yes ☐ no ☐
For information on Author's Choice Open Access, see the Instructions for Authors at
http://www.genetics.org/site/misc/ifora.xhtml#PUBLICATION%20CHARGES

1 Please verify styling of Greek and math symbols in text and equations throughout article. Check carefully for correct use of boldface, italics, operators, qualifiers, spacing, superscripts, and subscripts.

2 The in-text citation "Karlin and Taylor 1975" is not in Literature Cited. Please correct the citation, add the reference to the list, or delete the citation.

3 The in-text citation "Grimmett and Stirzaker 1982" is not in Literature Cited. Please correct the citation, add the reference to the list, or delete the citation.

4 "The time back to the last event, τ_G," as meant? (Original had "of" instead of "to".)

5 Please provide page number for quotation from "Barton and Otto (2005)".

6 Barton 1995a,b" as meant, to match Literature Cited?

7 Please confirm page range for "Maynard Smith 1968".