Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Journal of Theoretical Biology 235 (2005) 207–219

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Background selection and population differentiation

Xin-Sheng Hu*, Fangliang He

Department of Renewable Resources, 751 General Services Building, University of Alberta, Edmonton, AB Canada T6G 2H1

Received 23 July 2004; received in revised form 24 November 2004; accepted 6 January 2005
Available online 8 March 2005

Abstract

A general analytical formula is derived, which predicts the effects of background selection on population differentiation at a neutral locus as a result of its linkage with selected loci of deleterious mutations. The theory is based on the assumptions of random mating, multiplicative fitness, and weak selection in hermaphrodite plants in the island model of population structure. The analytical results show that F_{st} at the neutral locus increases as a result of the effects of background selection, regardless of the dependence or independence among linked background selective loci. The increment in F_{st} is closely related to the magnitude of linkage disequilibria between the neutral locus and selected loci, and can be estimated by the ratio of F_{st} with background selection to F_{st} without background selection minus one. The steady-state linkage disequilibrium between a neutral locus and a selected locus in subpopulations, primarily attained by gene flow, decreases with the recombination rate, and can be enhanced when there are dependence among linked selected loci. Monte Carlo computer simulations with two- and three-locus models show that the analytical formulae perform well under general conditions. Application of the present theory may aid in analyzing the genome-wide mapping of the effect of background selection in terms of F_{st}.

Keywords: Background selection; Population differentiation; Linkage disequilibrium; Gene flow; Selection

1. Introduction

Like the selectively favored mutations that cause hitchhiking effects on linked neutral loci (Maynard Smith and Haigh, 1974), selectively disfavored mutations can also change gene frequencies and reduce genetic diversities at linked neutral loci (“background selection”, Charlesworth et al., 1993). Early studies showed that a substantial reduction in genetic diversity at a neutral locus can result from its linkage to deleterious mutations (e.g., Charlesworth et al., 1993; Hudson and Kaplan, 1995; Nordborg et al., 1996). The genetic basis for maintaining both kinds of effects is the persistence of the linkage disequilibrium (LD) between neutral and selected loci (see the review by Barton, 2000). When LD equals zero, both kinds of effects disappear.

In a natural population without subdivision the LD between two linked neutral loci dissipates with generation as the consequence of recombination, and eventually approaches zero (e.g., Bennett, 1954; Hill and Robertson, 1968; Hill, 1974). In the population with subdivision the dissipation of global LD with generation is enhanced since the inter-subpopulation gene flow can reduce the effective size of the whole population and hence increase the drift speed (Wright, 1943). However, a certain amount of LD in local subpopulations can be attained owing to the effects of inter-subpopulation gene flow that counteracts genetic drift. Stable LD between selected nuclear loci without epistasis can be maintained in subdivided populations (Li and Nei, 1974). When the recombination fraction between selected nuclear loci is of the same order or smaller than the selection coefficient, a substantial amount of LD can be present in a cline (Slatkin, 1975). The LD between selected

*Corresponding author. Tel.: +1 780 492 0715; fax: +1 780 492 4323.
E-mail address: xin-sheng.hu@ualberta.ca (X.-S. Hu).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.01.004

nuclear and cytoplasmic loci that are physically unlinked can even be generated when inter-population gene flow (seed and pollen flow) takes place (e.g., Hu and Li, 2002).

Similarly, the persistence of the LD between one neutral locus and another selected locus is expected in a local population owing to the inter-subpopulation gene flow. As long as a certain amount of LD between neutral and selected loci is preserved, the effect of background selection should be present. The persistence of LD can cause an increase in variance of neutral allele frequencies and hence increase its population differentiation (Barton, 2000).

Previous LD studies are often examined in terms of genes or molecular markers as an ''observation unit'' at the equilibrium between gene flow and genetic drift. At a fine scale, the length of genomes for maintaining a certain amount of LD can be long in terms of the number of base pairs and so is the length within which background selection has a significant effect. It is meaningful to examine the effect of background selection in terms of single nucleotide polymorphisms (SNP) as an observation unit/marker. For example, 1 percent of recombination fraction (1 centiMorgan or cM) is equal to 1 million base pairs on the physical map in human genomes and contains about 1000 SNP (e.g., Wang et al., 1998). Within a few cMs of genetic distance the LD between one selected nucleotide site and another neutral site is likely substantial, and the effect of background selection on population differentiation at the individual neutral sites can be significant. Evidence indicates that high LD may extend over several centiMorgans in cattle and human genomes (e.g., Farnir et al., 2000; Abecasis et al., 2001). Needless to say, LD distribution along genomes varies with populations (e.g., Goddard et al., 2000; Shifman et al., 2003).

SNP are abundant in various organisms, such as in *Arabidopiss thaliana* and the rice genome (see the review by Rafalski, 2002). The genetic diversities of SNP within either the coding or non-coding regions of a gene are affected by their physical distances from the selected sites that may be located within the same gene or in the regions of other genes. For multiple linked genes with unequal numbers of SNP, the spatial pattern of genetic diversity across SNP could exhibit a patchy pattern along chromosomes. These naturally occurring patterns of SNP diversity provide a tool for mapping the effect of background selection in terms of population differentiation.

The purpose of this study is to develop further population genetic theory required for understanding the effect of background selection on population differentiation at a neutral locus. Although the effect of background selection on genetic diversity of a neutral gene has widely been appreciated, the theoretical investigation of such effect on population differentiation is much less fully explored (Charlesworth et al., 1997; Nordborg, 1997; Barton, 2000). In this study, we analytically derive the population differentiation at a neutral site due to its linkage with the sites that are subject to unfavorable mutations. Computer simulations are conducted to validate the analytical results that demonstrate the increase in F_{st} owing to the effect of background selection.

2. Assumptions

Based on the classical island model of population structure (Wright, 1969), here we consider diallelic selected nuclear loci (diploid) that are linked with a neutral locus in a hermaphrodite population of plants. For simplicity the selected loci addressed throughout this study refer to those with selectively disfavored mutation. Weak selection is considered in modeling so that all terms containing the second or higher order of selection coefficient are neglected. Like Nordborg et al. (1996), the selected loci are subject to a balance of mutation–selection–migration, and genetic drift effects are assumed negligible. The dependence among selected loci, caused by gene flow, is considered, relaxing the independence assumption made by Nordborg et al. (1996) and Hudson and Kaplan (1995).

The modeling procedure is based on a sequence of events in the life cycle of hermaphrodite plants: pollen flow, random combination between pollen and ovules (random mating), seed flow, mutation, natural selection, genetic drift, and next adults. This procedure is similar to Hu and Ennos (1999) except that mutation and background selection are included and also similar to Nordborg et al. (1996) except that migration is considered. The gene frequencies in migrants of pollen grains or seeds are equal to the average of gene frequencies over all subpopulations. The gene frequency in ovules before random combination with pollen grains is assumed to be the same as that in the preceding generation.

In the following we first derive the change of gene frequency at a neutral locus as a result of linkage to one and two selected loci, and then give a general expression for the change due to the background selection from an arbitrary number of selected loci. Wright's F_{st} is then employed to describe the population differentiation.

3. Allele frequency

3.1. Two-locus case

Consider a selected locus A that is linked to a neutral locus C in the ith subpopulation. The wild-type allele at the A locus is denoted by A_i, and its mutant allele by a_i;

their frequencies are p_{A_i} and p_{a_i} ($p_{A_i}+p_{a_i}=1$), respectively. Let the mutation rate from the wild-type allele to the mutant allele be u_1 at the A locus. The fitness of genotypes is assumed to be 1, $1-s_{1i}$, and $1-2s_{1i}$ for the genotypes of A_iA_i, A_ia_i, and a_ia_i, respectively. The migration rates of pollen and seeds into each subpopulation are denoted by m_P and m_S, respectively. According to the life cycle mentioned in the assumptions, the change in the allele frequency at the A locus is given by

$$\Delta p_{A_i}=p_{A_i}p_{a_i}s_{1i}-u_1p_{A_i}-\tilde{m}(p_{A_i}-\bar{p}_A), \tag{1}$$

where $\tilde{m}=m_S+m_P/2$, $\bar{p}_A$ is the frequency of the allele A_i in migrants (seeds and pollen grains). This equation can also be implied from Wright's general expression (Wright 1969, p. 474). The first term on the right-hand side of Eq. (1) represents the increment in p_{A_i} due to selection, the second term is the reduction due to mutation, and the third is the change due to immigration. At steady state $\Delta p_{A_i}=0$, the allele frequencies at the A locus can be analytically solved from Eq. (1), $p_{A_i}=((s_{1i}-u_1-\tilde{m})\pm\sqrt{(s_{1i}-u_1-\tilde{m})^2+4s_{1i}\tilde{m}\bar{p}_A})/2s_{1i}$ with the condition of $0\leqslant p_{A_i}\leqslant 1$.

Consider the neutral locus that has alleles C_i and c_i in the ith subpopulation. Let the mutation rate from C_i to c_i be v. There are four types of two-locus gametes: A_iC_i, A_ic_i, a_iC_i, and a_ic_i, with frequencies of $P_{A_iC_i}$, $P_{A_ic_i}$, $P_{a_iC_i}$, and $P_{a_ic_i}$, respectively. Let x_{0i} ($0\leqslant x_{0i}\leqslant 1$) be the probability that the allele C_i is linked with a mutant-free background of gametes with respect to the A locus, $p(C_i|A_i)=x_{0i}$. Let x_{1i} ($0\leqslant x_{1i}\leqslant 1$) be the probability that the allele C_i is linked with the mutant allele a_i, $p(C_i|a_i)=x_{1i}$. The conditional probabilities for x_{0i} and x_{1i} in migrants are denoted as $\bar{x}_0$ and $\bar{x}_1$, respectively. According to the Bayesian theorem, the frequencies of the four types of gametes can be expressed as $P_{A_iC_i}=p_{A_i}x_{0i}$, $P_{A_ic_i}=p_{A_i}(1-x_{0i})$, $P_{a_iC_i}=p_{a_i}x_{1i}$, and $P_{a_ic_i}=p_{a_i}(1-x_{1i})$.

Let r_1 be the recombination fraction between the A and C loci. Following the approach similar to Nordborg et al. (1996, p. 170), the changes in the conditional probabilities of x_{0i} and x_{1i} due to the joint effects of migration, selection, and mutation at the A locus are derived as Eqs. (A.3) and (A.4) in Appendix A. When there is no effect of migration, Eqs. (A.3) and (A.4) reduce to the previous results of Nordborg et al. (1996).

Let $\Delta p'_{C_i}$ be the change in the frequency of the neutral allele C_i due to the joint effects of background selection, mutation, and migration. According to Eqs. (A.3) and (A.4) in Appendix A and the relation of $p_{C_i}=P_{A_iC_i}+P_{a_iC_i}$ the analytical expression for $\Delta p'_{C_i}$ is given by

$$\begin{aligned}\Delta p'_{C_i}&=\Delta P_{A_iC_i}+\Delta P_{a_iC_i}=p_{A_i}\Delta x_{0i}+p_{a_i}\Delta x_{1i}\\&=-\tilde{m}(p_{C_i}-\bar{p}_C)-vp_{C_i}+s_{1i}D_{AC(i)},\end{aligned} \tag{2}$$

where $D_{AC(i)}$ is the LD between the A and C loci in the ith subpopulation. The first term on the right-hand side of Eq. (2) is the change due to migration (seed and pollen flow), the second term is the change due to the mutation of the allele C_i to other alleles, and the third term is the change due to the linkage to the selected A locus. If the linkage disequilibrium is of the order similar to the selection coefficient (s_{1i}), the third term on the right-hand side of Eq. (2) is negligible. Since LD is primarily generated by the inter-subpopulation gene flow, its magnitude can be much greater than the order of selection coefficient when the recombination fraction is very small, say within a few cMs of genome.

From the setting of the conditional probabilities of x_{0i} and x_{1i}, $D_{AC(i)}$ can be expressed by

$$D_{AC(i)}=p_{A_i}p_{a_i}(x_{0i}-x_{1i}). \tag{3}$$

When the neutral allele C_i is equally distributed under the mutant and mutant-free backgrounds of the A locus, i.e. $x_{0i}=x_{1i}$, the effect of background selection equals zero ($D_{AC(i)}=0$).

At the steady state the changes in conditional probability x_{0i} and x_{1i} per generation is equal to zero. Genetic drift does not change the means of the conditional probabilities x_{0i} and x_{1i} and hence the mean of $D_{AC(i)}$ although it alters the distributions of these variables. Instead of using the diffusion model (e.g. Nordborg et al., 1996), the steady state x_{0i} and x_{1i} can be calculated by letting $\Delta x_{0i}=\Delta x_{1i}=0$ according to Eqs. (A.3) and (A.4) in Appendix A, that is

$$\begin{pmatrix} p_{a_i}s_{1i}-u_1-v-\tilde{r}_{1i}p_{a_i}-\tilde{m} & \tilde{r}_{1i}p_{a_i}\\ \tilde{r}_{1i}p_{A_i}+u_1p_{A_i}/p_{a_i} & -p_{A_i}(s_{1i}+\tilde{r}_{1i})-v-\tilde{m}\end{pmatrix}\begin{pmatrix}x_{0i}\\ x_{1i}\end{pmatrix}=\begin{pmatrix}-\tilde{m}\bar{p}_{AC}/p_{A_i}\\ -\tilde{m}\bar{p}_{aC}/p_{a_i}\end{pmatrix}, \tag{4}$$

where $\tilde{r}_{1i}=r_1(1-(1-2p_{a_i})s_{1i})$. From Eq. (4), we obtained

$$x_{0i}=\frac{\tilde{m}(\tilde{r}_{1i}\bar{p}_C+\bar{P}_{AC}(p_{A_i}s_{1i}+v+\tilde{m})/p_{A_i})}{\tilde{r}_{1i}(\tilde{m}+v)}, \tag{5a}$$

$$x_{1i}=\frac{\tilde{m}(\tilde{r}_{1i}\bar{p}_C-(-u_1\bar{P}_{AC}+\bar{P}_{aC}(p_{a_i}s_{1i}-u-v-m))/p_{a_i})}{\tilde{r}_{1i}(\tilde{m}+v)}. \tag{5b}$$

According to Eqs. (1) and (3) and the relation $\bar{p}_C=\bar{p}_{AC}+\bar{p}_{aC}$, the steady-state $D_{AC(i)}$ can be derived as

$$D_{AC(i)}=\frac{\tilde{m}}{\tilde{r}_{1i}}\left(\bar{D}_{AC}-\frac{v}{m+v}(p_{A_i}-\bar{p}_A)\bar{p}_C\right), \tag{6}$$

where $\bar{D}_{AC}$ is the LD in migrants. Eq. (6) explicates that $D_{AC(i)}$ reduces with the increasing recombination fraction, but increases with the increasing migration rate or the selection coefficient.

3.2. Three-locus case

Assume that the neutral locus is linked on either side to a selected locus each with disfavored mutations. The difference between the two- and three-locus cases is that the effects of LD between selected loci generated by gene flow and the double crossover among the three loci are included. Assume that another diallelic selected locus B links to the neutral locus C at the opposite side to the A locus, i.e. the order of ACB. The wild-type allele at the B locus in the ith subpopulation is denoted by B_i, and its mutant allele is denoted by b_i; their frequencies are p_{B_i} and p_{b_i} $(p_{B_i}+p_{b_i}=1)$, respectively. Let the mutation rate from the wild-type allele to the mutant be u_2 at the B locus. The fitness is assumed to be 1, $1-s_{2i}$, and $1-2s_{2i}$ for the genotypes of B_iB_i, B_ib_i, and b_ib_i, respectively.

Using the assumption of multiplicative viability, the fitness for each two-locus genotype can be readily calculated. Following the life cycle the steady-state allele frequencies at the A and B loci at the balance of migration–selection–mutation are shown to have the following relations:

$$p_{A_i}p_{a_i}s_{1i} - u_1p_{A_i} - \tilde{m}(p_{A_i} - \bar{p}_A) - D_{AB(i)}s_{2i} = 0, \tag{7a}$$

$$p_{B_i}p_{b_i}s_{2i} - u_2p_{B_i} - \tilde{m}(p_{B_i} - \bar{p}_B) - D_{AB(i)}s_{1i} = 0, \tag{7b}$$

where $\bar{p}_B$ is the frequency of the allele B_i in migrants (seeds and pollen grains), and $D_{AB(i)}$ is the steady-state LD between the A and B loci.

The steady-state $D_{AB(i)}$ can be obtained from Eq. (B.2) in Appendix B, that is

$$D_{AB(i)} = \frac{\tilde{m}(\bar{D}_{AB} + (p_{A_i} - \bar{p}_A)(p_{B_i} - \bar{p}_B))}{1 - (1 - \tilde{m} - (p_{A_i} - p_{a_i})s_{1i} - (p_{B_i} - p_{b_i})s_{2i} - u_1 - u_2)(1 - r)}, \tag{8}$$

where r is the recombination rate between the A and the B loci. The analytical expressions for the allele frequencies at the A and B loci are hard to obtain using the joint Eqs. (7) and (8). In the specific case where allele frequencies at the A and B loci are coincident, i.e. $s_{1i} = s_{2i} = s_i$, $u_1 = u_2 = u$, $\bar{p}_A = \bar{p}_B = \bar{p}$, and $p_{A_i} = p_{B_i} = p_i$, we obtained a cubic equation

$$d_0p_i^3 + d_1p_i^2 - d_2p_i - d_3 = 0, \tag{9}$$

where

$$d_0 = 4(1-r)s_i^2,$$

$$\begin{aligned} d_1 &= s_i(1 + \tilde{m} - (1 - \tilde{m} - 2u + 2s_i)(1-r) \\ &\quad - 4(s_i - u - m)(1-r)), \end{aligned}$$

$$\begin{aligned} d_2 &= (1 - (1 - \tilde{m} - 2u + 2s_i)(1-r))(s_i - u - \tilde{m}) \\ &\quad + 4s_i\tilde{m}(1-r)\bar{p} + 2s_i\tilde{m}\bar{p} \end{aligned}$$

and

$$d_3 = \tilde{m}((1 - (1 - \tilde{m} - 2u + 2s_i)(1-r))\bar{p} - s_i\bar{D}_{AB} - s_i\bar{p}^2).$$

Solutions to Eq. (9) can be calculated with the *Mathematica* tool.

There are eight types of three-locus gametes: $A_iC_iB_i$, $A_iC_ib_i$, $A_ic_iB_i$, $A_ic_ib_i$, $a_iC_iB_i$, $a_iC_ib_i$, $a_ic_iB_i$, and $a_ic_ib_i$, with frequencies of $P_{A_iC_iB_i}$, $P_{A_iC_ib_i}$, $P_{A_ic_iB_i}$, $P_{A_ic_ib_i}$, $P_{a_iC_iB_i}$, $P_{a_iC_ib_i}$, $P_{a_ic_iB_i}$, and $P_{a_ic_ib_i}$, respectively. Let the probability that the allele C_i is linked with a mutant-free background of gametes with respect to the two selected loci, $p(C_i|A_iB_i)$, be y_{0i}. Similarly, let $p(C_i|a_iB_i) = y_{1i}$, $p(C_i|A_ib_i) = y_{2i}$, and $p(C_i|a_ib_i) = y_{3i}$. All these conditional probabilities are in the range of 0 to 1 $(0 \leqslant y_{ji} \leqslant 1;\ j = 0,1,2,3)$. The frequencies of the three-locus gametes $A_iC_iB_i$ and $A_ic_iB_i$ can be written as $P_{A_iC_iB_i} = P_{A_iB_i}y_{0i}$, and $P_{A_ic_iB_i} = P_{A_iB_i}(1 - y_{0i})$, respectively, where $P_{A_iB_i}$ is the frequency of gamete A_iB_i. The expressions for the remaining six three-locus gametes ($A_iC_ib_i$, $A_ic_ib_i$, $a_iC_iB_i$, $a_iC_ib_i$, $a_ic_iB_i$, and $a_ic_ib_i$) can be written in a similar way. The conditional probabilities for y_{0i}, y_{1i}, y_{2i}, and y_{3i} in migrants are denoted as $\bar{y}_0$, $\bar{y}_1$, $\bar{y}_2$, and $\bar{y}_3$, respectively.

Let r_2 be the recombination fraction between the B and C loci. The changes in the conditional probabilities of y_{0i}, y_{1i}, y_{2i}, and y_{3i} due to the joint effects of migration, selection, and mutation are given in Appendix C. According to Appendix C the analytical expression for $\Delta p'_{C_i}$ is

$$\begin{aligned} \Delta p'_{C_i} &= \Delta P_{A_iC_iB_i} + \Delta P_{a_iC_iB_i} + \Delta P_{A_iC_ib_i} + \Delta P_{a_iC_ib_i} \\ &= P_{A_iB_i}\Delta y_{0i} + P_{a_iB_i}\Delta y_{1i} + P_{A_ib_i}\Delta y_{2i} + P_{a_ib_i}\Delta y_{3i} \\ &= -\tilde{m}(p_{C_i} - \bar{p}_C) - vp_{C_i} + s_{1i}D_{AC(i)} + s_{2i}D_{CB(i)}, \end{aligned} \tag{10}$$

where $D_{CB(i)}$ are the LD between the C and B loci in the ith subpopulation.

According to the conditional probabilities of y_{0i}, y_{1i}, y_{2i}, and y_{3i}, $D_{AC(i)}$ and $D_{CB(i)}$ can be, respectively, given by

$$\begin{aligned} D_{AC(i)} &= p_{A_1}p_{a_i}(p_{B_i}(y_{0i} - y_{1i}) + p_{b_i}(y_{2i} - y_{3i})) \\ &\quad + (p_{a_i}(y_{0i} - y_{2i}) + p_{A_i}(y_{1i} - y_{3i}))D_{AB(i)}, \end{aligned} \tag{11a}$$

$$\begin{aligned} D_{CB(i)} &= p_{B_i}p_{b_i}(p_{A_i}(y_{0i} - y_{2i}) + p_{a_i}(y_{1i} - y_{3i})) \\ &\quad + (p_{b_i}(y_{0i} - y_{2i}) + p_{B_i}(y_{2i} - y_{3i}))D_{AB(i)}. \end{aligned} \tag{11b}$$

The first part on the right-hand side of Eqs. (11a) or (11b) is the amount without the influence of the LD between the A and B loci, and the second part is the increment due to the LD between the A and B loci generated by gene flow. The above equation analytically demonstrates that the presence of the LD among linked selected loci can enhance the effects of background selection. The proportion of $D_{AC(i)}$ and $D_{CB(i)}$ explained by the component of $D_{AB(i)}$ can be assessed by $(p_{a_i}(y_{0i} - y_{2i}) + p_{A_i}(y_{1i} - y_{3i}))D_{AB(i)}/D_{AC(i)}$ and $(p_{b_i}(y_{0i} - y_{2i}) + p_{B_i}(y_{2i} - y_{3i}))D_{AB(i)}/D_{CB(i)}$, respectively.

Letting $\Delta y_{ji} = 0\,(j = 0, 1, 2, 3)$ in Appendix C, we obtain four non-linear equations for calculating the steady-state conditional probabilities (y_{0i}, y_{1i}, y_{2i}, and y_{3i}),

$$\begin{pmatrix} h_{00} & \cdots & h_{05} \\ \cdots & & \\ h_{30} & \cdots & h_{35} \end{pmatrix}_{4\times 6} \left(y_{0i}\ y_{1i}\ y_{2i}\ y_{3i}\ y_{0i}y_{3i}\ y_{1i}y_{2i}\right)^{\mathrm{T}} = \begin{pmatrix} g_0 \\ \cdots \\ g_3 \end{pmatrix}_{4\times 1}, \qquad (12)$$

where

$$\begin{aligned} h_{00} = {} & s_{1i}p_{a_i} + s_{2i}p_{b_i} - u_1 - u_2 - v - \tilde{m} \\ & - (1 + 2s_{1i}p_{a_i} + 2s_{2i}p_{b_i})(P_{a_iB_i}(1 - s_{1i})r_1 \\ & + P_{A_ib_i}(1 - s_{2i})r_2 + P_{a_ib_i}(1 - s_{1i} - s_{2i}) \\ & \times (r_1 + r_2 - r_1r_2)), \ldots, \end{aligned}$$

$$h_{35} = \frac{1 + 2s_{1i}p_{a_i} + 2s_{2i}p_{b_i}}{P_{a_ib_i}}(r_1r_2(1 - s_{1i} - s_{2i})P_{A_ib_i}P_{a_iB_i}),$$

$$g_0 = -\frac{\tilde{m}\bar{P}_{ACB}}{P_{A_iB_i}}, \ldots, g_3 = -\frac{\tilde{m}\bar{P}_{aCb}}{P_{a_ib_i}}.$$

The analytical solution is algebraically complicated, but can be numerically solved with *Mathematica* model solutions.

Our numerical examples demonstrate that both $D_{AB(i)}$ and $D_{AC(i)}$ increase with the migration rate (Fig. 1a). The proportion of $D_{AC(i)}$ explained by the component of $D_{AB(i)}$ can be more than 20% when $\tilde{m} = 0.1$ (Fig. 1b). Both $D_{AB(i)}$ and $D_{AC(i)}$ decrease with the recombination rate (Fig. 2a,b). The proportion of $D_{AC(i)}$ explained by the component of $D_{AB(i)}$ can be substantially increased when the three loci are tightly linked, and more than 20% of $D_{AC(i)}$ can be brought about when $r_1 = r_2 = 0.0005$ (Fig. 2c).

3.3. General case

From the preceding two- and three-locus analyses, extension can be obtained to a more general case where a neutral locus links to an arbitrary number of selected loci among which the interaction may exist. The change in the frequency of the neutral locus C_i in the ith subpopulation, regardless of the magnitude of the linkage disequilibria among selected loci, can be generally expressed as

$$\Delta p'_{C_i} = -\tilde{m}(p_{C_i} - \bar{p}_C) - vp_{C_i} + \Lambda_i, \qquad (13)$$

where $\Lambda_i = \sum_{j=1}^{L} s_{ji}D_{M_jC(i)}$, in which M_j represents the wild-type allele at the jth selected locus. Effects of the LD among multiple selected loci are included in $D_{M_jC(i)}$.

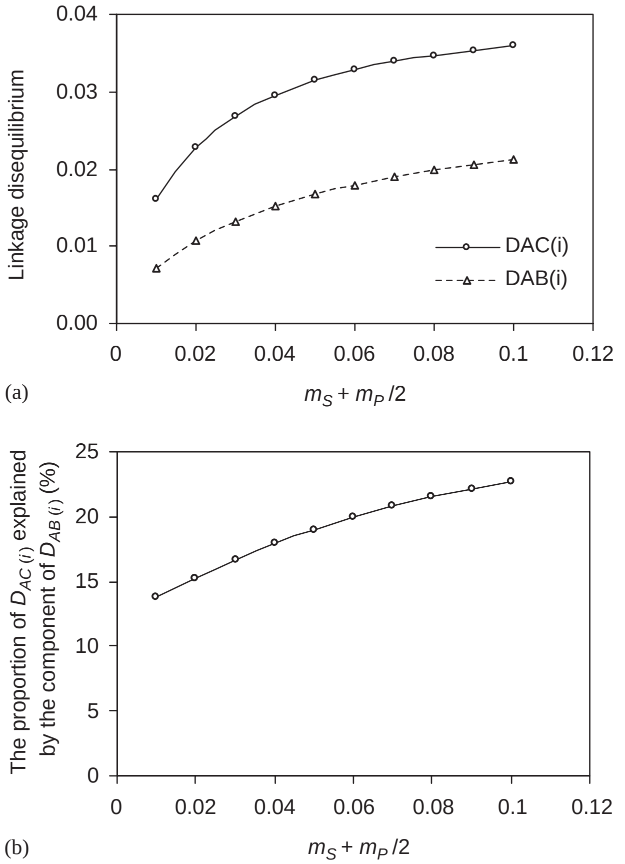

Fig. 1. Effects of the dependence between the selected A and B loci: (a) Changes in $D_{AC(i)}$ and $D_{AB(i)}$ with the migration rate. (b) The proportion of $D_{AC(i)}$ explained by the component of $D_{AB(i)}$. $D_{AB(i)}$ is calculated according to Eqs. (7) and (8) while $D_{AC(i)}$ is calculated according to Eq. (11a). The settings of other parameters are the mutation rate for the A and B loci $u_1 = u_2 = 10^{-5}$ and for the neutral locus $v = 10^{-4}$, the selection coefficients $s_{1i} = s_{2i} = 0.02$, the recombination rates between the A and C loci or the B and C loci $r_1 = r_2 = 0.01$ and between the A and B loci $r = 2r_1$, the LD between the A and B loci in migrants $\bar{D}_{AB} = 0.03$, the migrant allele frequencies $\bar{p}_A = \bar{p}_B = 0.8$, and the conditional probabilities in migrants for the neutral locus C under different backgrounds $\bar{y}_0 = 0.8$, $\bar{y}_1 = \bar{y}_2 = 0.7$, and $\bar{y}_3 = 0.2$.

The cumulative effects Λ_i from multiple selected loci are likely substantial when they closely link to the neutral locus.

4. Population differentiation

4.1. Equal spatial selection

We now examine the effects of background selection on population differentiation using the classical island model (Wright, 1969). Background selection can change both the effective population size and the neutral allele frequency. In the preceding section we have shown the systematical change of allele frequencies at a neutral locus as a result of linkage to selected loci. Previous

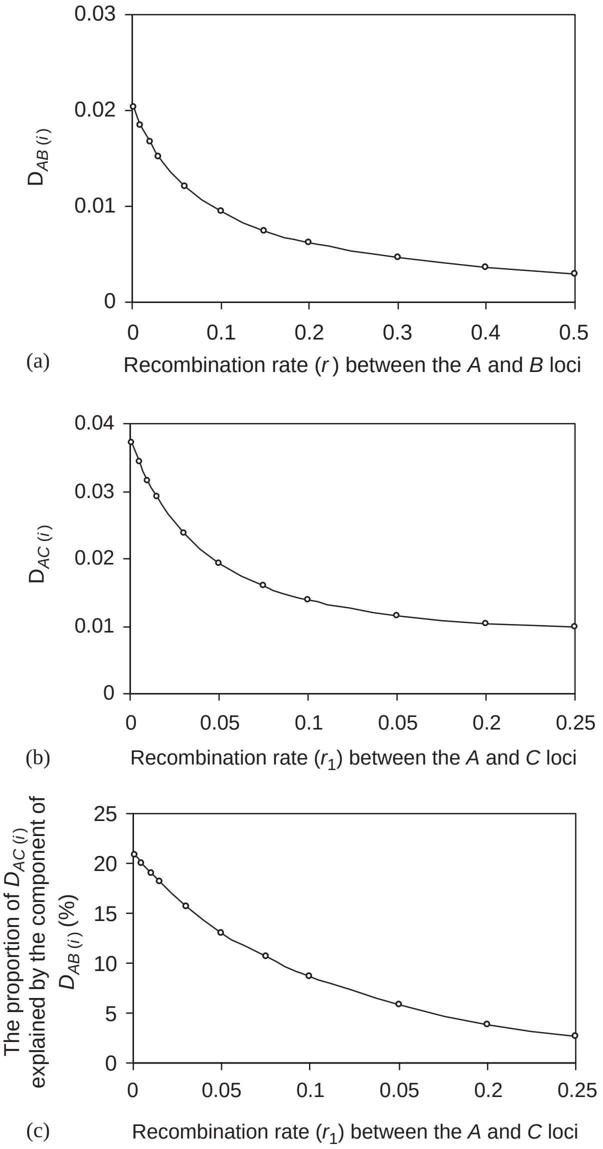

Fig. 2. Effects of the dependence between the selected A and B loci: (a) changes in $D_{AB(i)}$ with the recombination fraction between the A and B loci; (b) changes in $D_{AC(i)}$ with the recombination fraction between the A and C loci; (c) the proportion of $D_{AC(i)}$ explained by the component of $D_{AB(i)}$. $D_{AB(i)}$ is calculated according to Eqs. (7) and (8), while $D_{AC(i)}$ is calculated according to Eq. (11a). The settings of other parameters are the mutation rate for the A and B loci $u_1 = u_2 = 10^{-5}$ and for the neutral locus $v = 10^{-4}$, the selection coefficients $s_{1i} = s_{2i} = 0.02$, the recombination rates between the A and C loci or the B and C loci $r_1 = r_2$ and between the A and B loci $r = 2r_1$, the LD between the A and B loci in migrants $\bar{D}_{AB} = 0.03$, the migration rate $\tilde{m} = 0.05$, the migrant allele frequencies $\bar{p}_A = \bar{p}_B = 0.8$, and the conditional probabilities in migrants for the neutral locus C under different backgrounds $\bar{y}_0 = 0.8$, $\bar{y}_1 = \bar{y}_2 = 0.7$, and $\bar{y}_3 = 0.2$.

studies showed that background selection can reduce the effective subpopulation size (N_e) for the neutral locus (e.g. Nordborg et al., 1996) and hence affect the genetic drift process. Here we include both effects in deriving the expression for population differentiation. According to Eq. (4) of Nordborg et al. (1996), the steady-state effective size for the ith subpopulation under the impacts of background selection, denoted by N'_{ei}, can be approximated by

$$N'_{ei} = N_e e^{-\lambda_i}, \tag{14}$$

where

$$\lambda_i = \sum_{j=1}^{L} \frac{1 - p_{M_j}}{(1 + \tilde{r}_j / s_{ji})^2}.$$

We assume that the effective subpopulation size is not affected by gene flow although the effective size of the whole population is changed (Wright, 1943).

Denote by σ'^2 the variance of allele frequencies among subpopulations after pollen and seed flow and background selection. Suppose that the number of subpopulations (n) is large. According to Eq. (13), σ'^2 can be calculated by

$$\begin{aligned}\sigma'^2 &= \frac{1}{n}\sum_{i=1}^{n}(p'_{C_i} - \bar{p}_C)^2 \\ &\approx E_\Phi(((1-\tilde{m})(p_{C_i} - \bar{p}_C) - v p_{C_i} + \Lambda_i)^2) \\ &= ((1-\tilde{m})^2 - 2v)\sigma^2 + \Pi + \overline{\Lambda^2},\end{aligned} \tag{15}$$

where E_Φ represents the expectation with respect to the allele frequency distribution among subpopulations, $\Pi = 2(1-\tilde{m})E_\Phi((p_{C_i} - \bar{p}_C)\Lambda_i)$, and $\overline{\Lambda^2} = E_\Phi(\Lambda_i^2)$. Note that the expectations of the terms involving coefficients of v^2, mv, and $vs_\bullet$ are neglected in deriving Eq. (15).

When the selection coefficients for the same allele at any selected locus are equal among all subpopulations, i.e. $s_{1j} = \cdots = s_{nj}$, then $\lambda_1 = \cdots = \lambda_n$ and the effective subpopulation size is the same, i.e. $N'_{e1} = \cdots = N'_{en} = N'_e$. The effects of background selection are equal among subpopulations, i.e. $\Lambda_1 = \cdots = \Lambda_n = \Lambda$, and $\overline{\Lambda^2} = \Lambda^2$. The second term on the right-hand side of Eq. (15) varnishes, i.e. $\Pi = 0$. According to Hu and Ennos (1999), the steady-state variance of allele frequencies after genetic drift, can be written as

$$\begin{aligned}\sigma^2 = &\left(1 - \frac{1}{2N'_e}\right)(((1-\tilde{m})^2 - 2v)\sigma^2 + \Lambda^2) \\ &+ \frac{\bar{p}_C(1-\bar{p}_C)}{2N'_e}.\end{aligned} \tag{16}$$

Population differentiation at the neutral locus, denoted by F_{st1} $(= \sigma^2/\bar{p}_C(1-\bar{p}_C))$, can be obtained by substituting Eq. (14) into Eq. (16),

$$F_{st1} = \frac{1}{1 + 4N_e(\tilde{m} + v)e^{-\lambda}}\left(1 + \frac{2N_e e^{-\lambda} - 1}{\bar{p}_C(1-\bar{p}_C)}\Lambda^2\right). \tag{17}$$

Clearly, F_{st1} is greater than the population differentiation under the purely neutral process, denoted by $F_{st.b}$ $(= 1/(1 + 4N\tilde{m}))$ for diploid nuclear genes (Hu and Ennos, 1999).

4.2. Unequal spatial selection

The more general situation is that the selection coefficients of the same allele at any selected locus are unequal among subpopulations, and so is the effective subpopulation size among subpopulations, i.e. $N'_{e1} \neq \cdots \neq N'_{en}$. Eq. (15) remains effective since it is derived before the occurrence of genetic drift. The steady-state increment in the variance of allele frequencies after genetic drift, denoted by $\Delta\sigma_d^2$, is

$$\begin{aligned}\Delta\sigma_d^2 &= E_\Phi\left(\frac{p_{C_i}(1-p_{C_i})}{2N_e e^{-\lambda_i}}\right)\\ &= E_\Phi\left(\frac{p_{C_i}(1-p_{C_i})}{2N_e} + (\lambda_i + \lambda_i^2/2! + \cdots)\frac{p_{C_i}(1-p_{C_i})}{2N_e}\right)\\ &= \frac{1}{2N_e}(\bar{p}_C(1-\bar{p}_C) - \sigma^2) + \rho, \qquad (18)\end{aligned}$$

where $\rho = E_\Phi((\lambda_i + \lambda_i^2/2! + \cdots)\frac{p_{C_i}(1-p_{C_i})}{2N_e})$ is the increment part due to the background selection. Therefore, the steady-state equation for the variance of allele frequencies, equivalent to Eq. (16), is expressed as

$$\begin{aligned}\sigma^2 &= \left(1 - \frac{1}{2N_e}\right)(((1-\tilde{m})^2 - 2v)\sigma^2 + \Pi + \overline{\Lambda^2})\\ &\quad + \frac{\bar{p}_C(1-\bar{p}_C)}{2N_e} + \rho. \qquad (19)\end{aligned}$$

Denote by F_{st2} the population differentiation at the neutral locus. Rearranging Eq. (19) yields

$$F_{st2} = \frac{1}{1 + 4N_e(\tilde{m} + v)}(1 + \varepsilon), \qquad (20)$$

where

$$\varepsilon = \frac{(2N_e - 1)(\Pi + \overline{\Lambda^2}) + 2N_e\rho}{\bar{p}_C(1-\bar{p}_C)}.$$

In the presence of inbreeding in each subpopulation, the variance effective population size reduces to $N_e(1+F_{is})^{-1}$ where F_{is} is the inbreeding coefficient (Caballero and Hill, 1992). Thus, the assumption of random mating can be relaxed by replacing N_e in Eqs. (17) or (20) with $N_e(1+F_{is})^{-1}$. Also the assumption of the large number of subpopulations can be relaxed by replacing $\tilde{m}$ with $(n/(n-1))^2\tilde{m}$ in which n can be an arbitrary number of subpopulations (Hu, 2000).

When the estimates of $\hat{F}_{st2}$ and $\hat{F}_{st.b}$ are available, the increment in F_{st} due to the effect of background selection can be estimated according to the general formula of Eq. (20),

$$\hat{\varepsilon} = \frac{\hat{F}_{st2}}{\hat{F}_{st.b}} - 1. \qquad (21)$$

In order to look at the amount of increment in F_{st} at the neutral locus due to its linkage to selected loci, the above analytical results are applied to the two- and three-locus cases under the model of equal spatial selection. The steady-state F_{st} under the purely neutral process equals 0.2320 for $\tilde{m} = 0.015$, 0.0847 for $\tilde{m} = 0.05$, and 0.0444 for $\tilde{m} = 0.1$, with the settings of other parameters $N_e = 50$, $n = 30$, and $v = 5 \times 10^{-4}$. In the two-locus case, the proportion of increment in F_{st} is about 1.68% for $\tilde{m} = 0.015$, 3.23% for $\tilde{m} = 0.05$, and 4.10% for $\tilde{m} = 0.1$ when the neural locus tightly links to the selected locus A ($r_1 = 0.0001$) (Fig. 3a). The proportion of increment in F_{st} decreases with the recombination rate. In the three-locus case, the proportion of increment in F_{st} is about 4.22% for $\tilde{m} = 0.015$, 7.86% for $\tilde{m} = 0.05$, and 9.16% for $\tilde{m} = 0.1$ when $r_1 =$

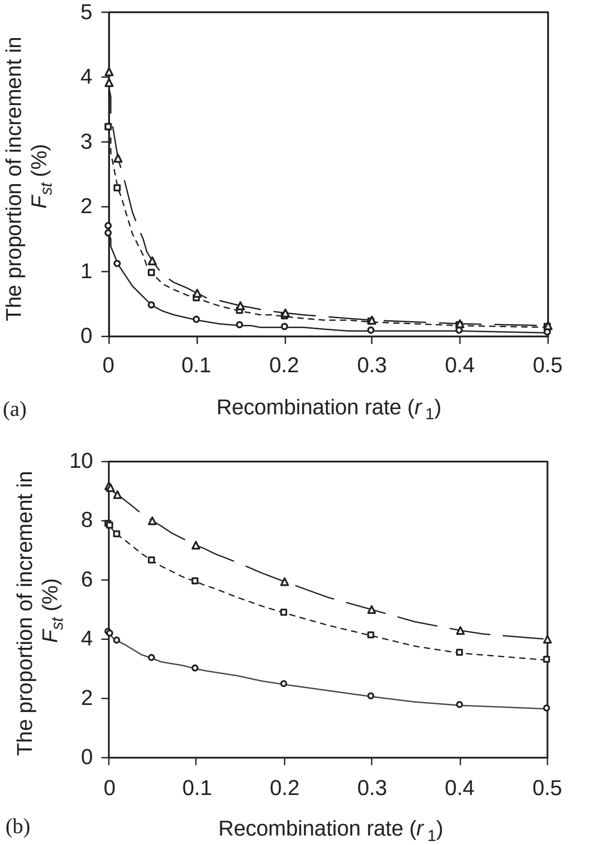

Fig. 3. The proportion of increment in F_{st} due to background selection: (a) two-locus case, with the selection coefficient $s_{1i} = 0.02$ ($i = 1, \ldots, 30$), the mutation rate $u_1 = 10^{-5}$ for the A locus and $v = 5 \times 10^{-4}$ for the C locus, the conditional probabilities in migrants $\bar{x}_0 = 0.8$ and $\bar{x}_1 = 0.2$, the migrant gamete frequencies $\bar{P}_{AC} = 0.76$ and $\bar{P}_{aC} = 0.01$, the migrant allele frequencies $\bar{p}_A = 0.95$ and $\bar{p}_C = 0.8$, and the effective population size $N_e = 50$; (b) three-locus case, with the mutation rate $u_1 = u_2 = 10^{-5}$ and $v = 5 \times 10^{-4}$, the selection coefficients $s_{1i} = s_{2i} = 0.02$ ($i = 1, \ldots, 30$), the recombination rates between the A and C loci or the B and C loci $r_1 = r_2$ and between the A and B loci $r = 2r_1$, the LD between the A and B loci in migrants $\bar{D}_{AB} = 0.03$, the migrant allele frequencies $\bar{p}_A = \bar{p}_B = 0.95$, the conditional probabilities in migrants $\bar{y}_0 = 0.82$, $\bar{y}_1 = \bar{y}_2 = 0.7$, and $\bar{y}_3 = 0.2$, and the effective population size $N_e = 50$. In each figure the line with circles represents the case of $\tilde{m} = 0.015$ ($m_S = m_P = 0.01$), the line with blocks for $\tilde{m} = 0.05$ ($m_S = 0.04$, $m_P = 0.02$), and the line with triangles for $\tilde{m} = 0.10$ ($m_S = 0.05$, $m_P = 0.1$).

0.0001 (Fig. 3b). Although the proportion of increment in F_{st} decreases with the recombination rate, they are greater than those in the two-locus case (Fig. 3a,b). Note that the parameters for the neutral locus C and its linkage to the A (or B) locus in the above two cases are comparable: $\bar{p}_C = 0.8$, $\bar{P}_{AC} = 0.76$, and $\bar{P}_{aC} = 0.01$ in the two-locus case; $\bar{p}_C = 0.7957$, $\bar{P}_{AC} = 0.7769$, and $\bar{P}_{aC} = 0.0187$ in the three-locus case. These numerical results indicate that a certain mount of increment in F_{st} can be attained when the neutral locus tightly links to selected loci or when the migration rate is high.

5. Simulations

5.1. Method

To confirm the analytical results simulation study was conducted according to the sequence of events in the life cycle of hermaphrodite plants. Simulation starts from an initial adult reference population that begins subdivision and produces many subpopulations. The allele frequencies at the selected loci are initially set to be the same as in the reference population. The conditional probabilities for a diallelic neutral gene in the reference population under different backgrounds of selected gametes are also assumed, thus the allele frequencies of the neutral locus can be calculated. The frequencies of two- or three-locus gametes in migrants (seeds and pollen grains) are assumed to be equal to those in the initial reference population. Constant selection coefficients among subpopulations are examined although a more complicated pattern of selection can be modeled.

The detailed simulation procedure is as follows. Given the initial parameter settings, calculate the frequencies of two- or three-locus gametes in pollen and ovules according to the assumption of Wright–Fisher's model. Then calculate the gamete frequencies after pollen flow. According to the assumption of random combination between pollen and ovules, calculate the genotype frequencies in seeds so formed. Seed flow is then considered and the genotype frequencies are calculated after seed flow. Assume that the mutation at the neutral locus is a deterministic process, with a probability of v from the allele C_i to the allele c_i per generation, and then calculate all genotype frequencies after mutation. The model of multiplicative viability is employed to calculate the fitness of each genotype in any subpopulation and the genotype frequencies after selection. A sampling process (genetic drift) is then conducted according to the phenotype frequencies after selection, given an effective subpopulation size (N'_{ei}). Gamete frequencies in ovules and pollen grains are then calculated according to the segregation ratios of gametes from individual cross (10 crosses in the two-locus case and 36 crosses in the three-locus case). The above steps are repeated until the population differentiation reaches steady distribution from generation to generation.

Five thousand independent data sets are created per generation, and each is used to calculate population differentiation. From these replicated datasets, means and standard deviation of F_{st} are calculated. The predicted results are obtained according to Eq. (17), where the LD is calculated from Eq. (6) in the two-locus case and Eq. (11a, b) in the three-locus case. The steady-state values of y_{ji} ($j = 0, 1, 2, 3$) in calculating the expected F_{st} values are calculated according to Eq. (12) with the *Mathematica* tool.

5.2. Results

Let $n = 30$, $v = 5 \times 10^{-4}$, and $N_e = 50$. In the purely neutral process, the theoretical expectation of F_{st} at steady state equals 0.2320 when $\tilde{m} = 0.015$, 0.0847 when $\tilde{m} = 0.05$, and 0.0444 when $\tilde{m} = 0.10$. In the presence of background selection, our simulations clearly show that population differentiation is enhanced, especially when the neutral locus is closely linked to the selected loci or when migration rate is high. In the two-locus case, for example, the average proportion of increment in F_{st} at the 200th generation (steady-state value) reduces from 9.48% for $r_1 = 0.01$–2.97% for $r_1 = 0.1$(Fig. 4a,b), although these estimates are greater than the expected values of 1.12% and 0.25% (Fig. 3a), respectively. The expected values of F_{st} are within the range of one standard deviation of empirical results. The general pattern for the change of F_{st} with generation is that the average F_{st} is initially small when the reference population starts subdivision and all subpopulations are formed by sampling from the same reference population. Then population differentiation gradually increases and approaches a steady distribution after 160 generations (Fig. 4a). Compared with the results in the case of loose linkage ($r_1 = 0.1$), population differentiation displays a greater fluctuation in the case of tight linkage ($r_1 = 0.01$; Fig. 4b).

Population differentiation quickly reaches a steady distribution with the increase of migration rate (Fig. 5a,b). The average proportion of increment in F_{st} generally increases with migration rate. For example, the average proportion of increment in F_{st} at the 200th generation increases from 9.48% for $\tilde{m} = 0.015$, to 10.5% for $\tilde{m} = 0.05$, and to 16.9% for $\tilde{m} = 0.10$ (Fig. 5a), although these estimates are greater than the expected values of 1.12%, 2.28%, and 2.74% (Fig. 3a), respectively. All expected F_{st} values are within the range of one standard deviation of empirical results.

The change of F_{st} with generation in the three-locus case has the pattern similar to that in the two-locus case, which displays a gradual increase with time and eventually approaches a stable distribution (Fig. 6a), but has a greater fluctuation than the latter (Fig. 6b).

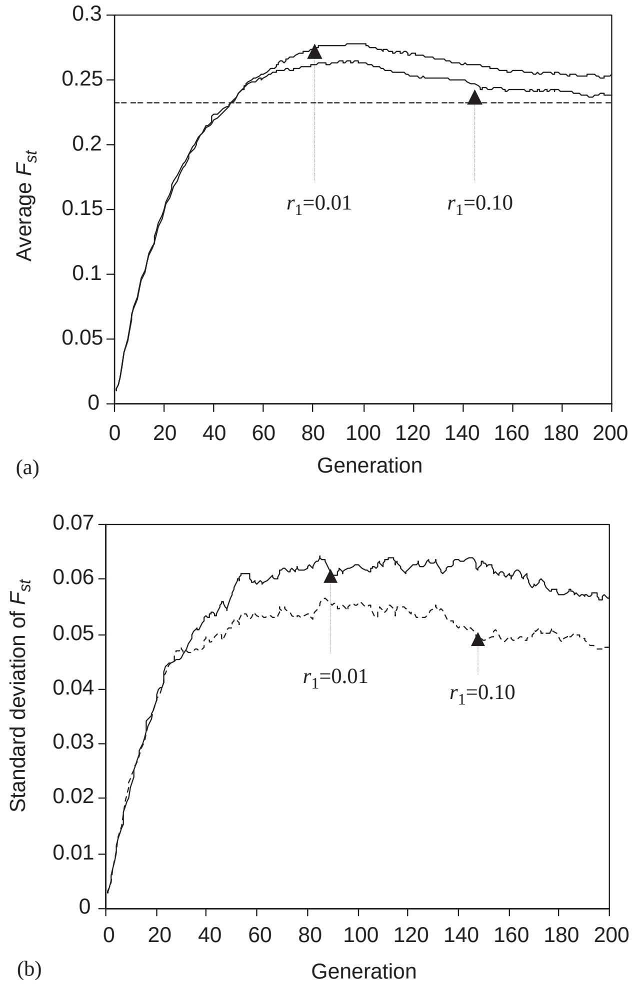

Fig. 4. Effects of recombination rate on background selection in the two-locus case: (a) average F_{st}; (b) standard deviation of F_{st}. Results are obtained from 5000 independent simulations, with $m_S = m_P = 0.01$, $u_1 = 10^{-5}$, $N_e = 50$, $v = 5 \times 10^{-4}$, $\bar{p}_A = 0.95$, $s_{1i} = 0.02$ ($i = 1, \ldots, 30$), and the conditional probabilities in migrants: $\bar{x}_0 = 0.8$ and $\bar{x}_1 = 0.2$. The dashed line in (a) refers to the F_{st} value (= 0.2320) under the purely neutral process.

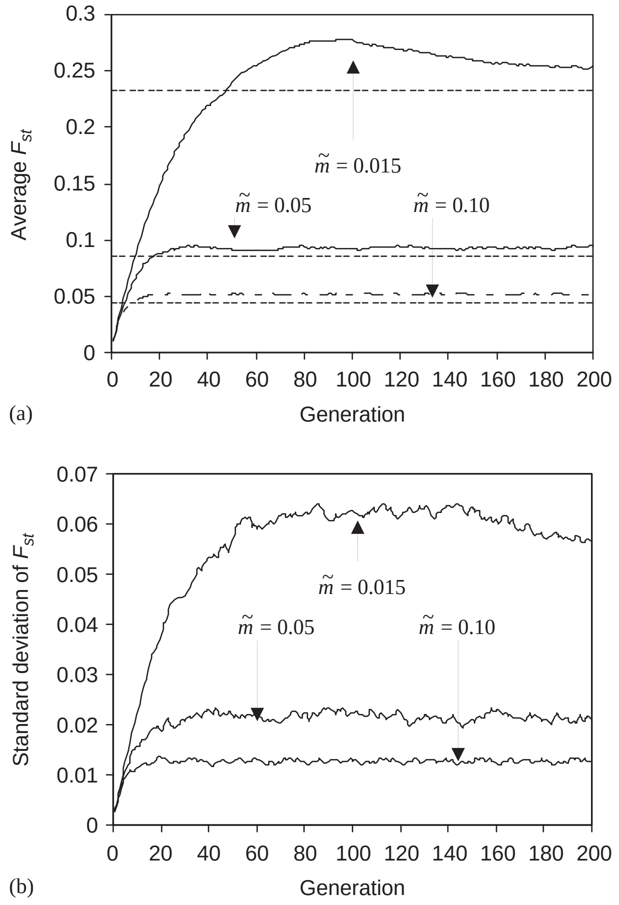

Fig. 5. Effects of migration on background selection in the two-locus case: (a) average F_{st}; (b) standard deviation of F_{st}. Results are obtained from 5000 independent simulations, with $u_1 = 10^{-5}$, $N_e = 50$, $v = 5 \times 10^{-4}$, $\bar{p}_A = 0.95$, $s_{1i} = 0.02$ ($i = 1, \ldots, 30$), $r_1 = 0.01$, and the conditional probabilities in migrants: $\bar{x}_0 = 0.8$, and $\bar{x}_1 = 0.2$. The dashed lines in (a) at the positions of $F_{st} = 0.2320$, 0.0847, and 0.0444 refer to the expected values under the purely neutral process with migration rates of $\tilde{m} = 0.015$ ($m_S = m_P = 0.01$), $\tilde{m} = 0.05$ ($m_S = 0.04$, $m_P = 0.02$), and $\tilde{m} = 0.10$ ($m_S = 0.05$, $m_P = 0.1$), respectively.

Compared with the results in the two-locus case, the average F_{st} in the three-locus case is increased. For example, the average proportion of increment in F_{st} at the 250th generation (steady-state value) is about 13.8% for $r_1 = 0.01$ and 11.6% for $r_1 = 0.1$ (Fig. 6a), although these estimates are greater than the expected values of 3.92% and 2.97% (Fig. 3b), respectively. The general pattern for the effect of migration on increasing the proportion of increment in F_{st} can be observed (Fig. 7a). For example, the average proportions of increment in F_{st} at the 250th generation are 13.8% for $\tilde{m} = 0.015$, 15.7% for $\tilde{m} = 0.05$, and 26.4% for $\tilde{m} = 0.10$ (Fig. 7a,b), although these estimates are greater than the expected values of 3.92%, 7.53%, and 8.88%, respectively (Fig. 3a). Again, all predicted F_{st} values are in the range of one standard deviation of the empirical results. In summary, these simulation results demonstrate that the cumulative effect of background selection on F_{st} at a neutral locus can be substantial if the neutral locus is closely linked to multiple selected loci.

6. Discussion

In this paper we have obtained the analytical expressions for population differentiation at a neutral locus in the island model of population structure. The increase in population differentiation as a result of background selection is analytically demonstrated under very general conditions. Although the individual effects of a single selected locus are likely to be small, the cumulative effect of multiple selected loci could be substantial. Our theoretical results can be applied to a

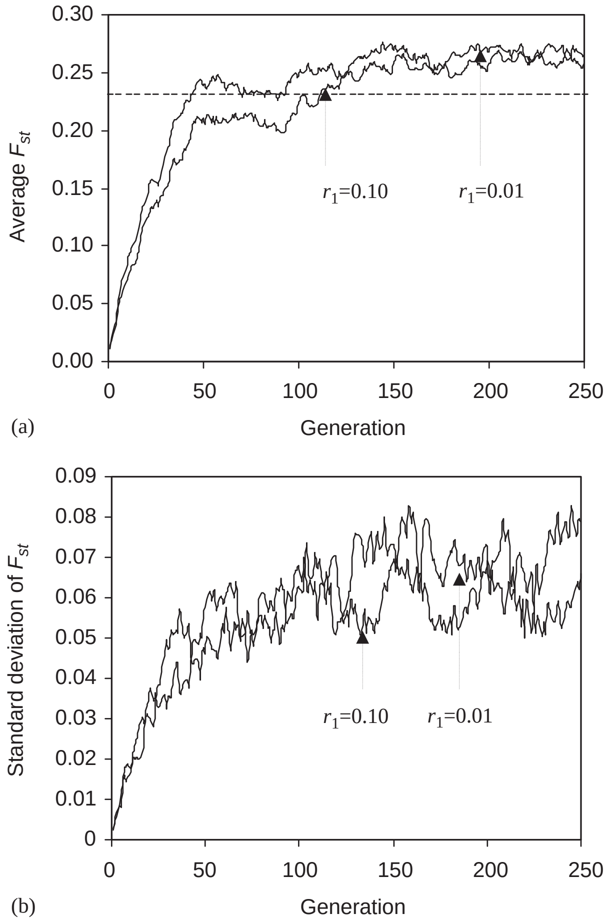

Fig. 6. Effects of recombination rate on background selection in the three-locus case: (a) average F_{st}; (b) standard deviation of F_{st}. Results are obtained from 5000 independent simulations, with $m_S = m_P = 0.01$, $u_1 = u_2 = 10^{-5}$, $N_e = 50$, $v = 5 \times 10^{-4}$, $\bar{p}_A = \bar{p}_B = 0.95$, $s_{1i} = s_{2i} = 0.02$ $(i = 1, \ldots, 30)$, $r_1 = r_2 = r/2$, $\bar{D}_{AB} = 0.03$, and the conditional probabilities in migrants for the neutral locus under different backgrounds $\bar{y}_0 = 0.82$, $\bar{y}_1 = \bar{y}_2 = 0.7$, and $\bar{y}_3 = 0.2$. The dashed line in (a) refers to the expected F_{st} (0.2320) under the purely neutral process.

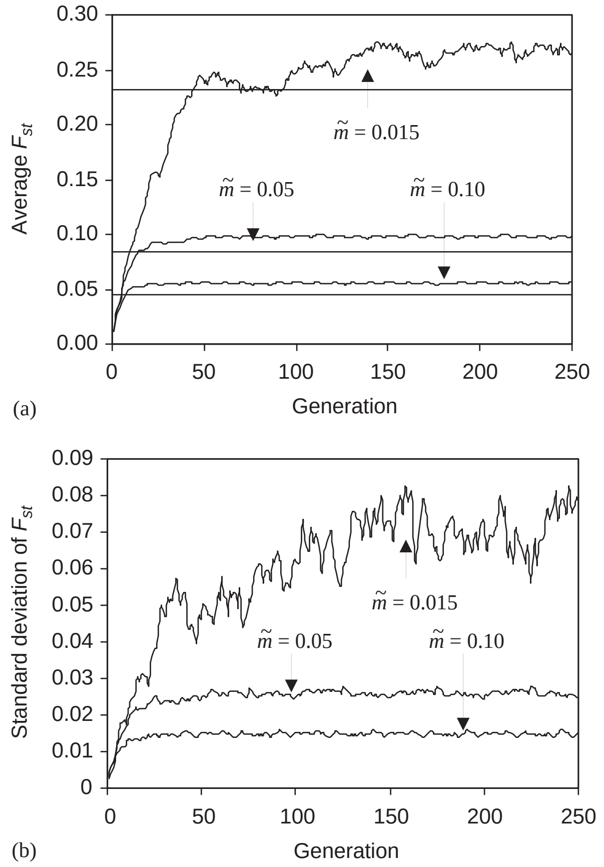

Fig. 7. Effects of migration on background selection in the three-locus case: (a) average F_{st}; (b) standard deviation of F_{st}. Results are obtained from 5000 independent simulations, with $u_1 = u_2 = 10^{-5}$, $N_e = 50$, $v = 5 \times 10^{-4}$, $\bar{p}_A = \bar{p}_B = 0.95$, $s_{1i} = s_{2i} = 0.02$ $(i = 1, \ldots, 30)$, $r_1 = r_2 = r/2 = 0.01$, $\bar{D}_{AB} = 0.03$, and the conditional probabilities in migrants for the neutral locus under different backgrounds $\bar{y}_0 = 0.82$, $\bar{y}_1 = \bar{y}_2 = 0.7$, and $\bar{y}_3 = 0.2$. The dashed lines in (a) at the positions of $F_{st} = 0.2320$, 0.0847, and 0.0444 refer to the expected values under the purely neutral process with migration rates of $\tilde{m} = 0.015$ $(m_S = m_P = 0.01)$, $\tilde{m} = 0.05$ $(m_S = 0.04,\ m_P = 0.02)$, and $\tilde{m} = 0.10$ $(m_S = 0.05,\ m_P = 0.1)$, respectively.

wide situation to map the effects of background selection in terms of population differentiation at a fine genome scale.

Although the present result is qualitatively the same as a previous study (Charlesworth et al., 1997), our approach is fundamentally different in theoretical deduction. By partitioning the total genetic diversity into the components of between and within subpopulations, Charlesworth et al. (1997) showed that the increase of F_{st} due to background selection is mainly caused by the decreased diversity within populations. Our general expression of F_{st} is derived using Wright's approach and seems more rigorous.

The expression of F_{st} given by Charlesworth et al. (1997) is only suitable for a pair of subpopulations. Slatkin and Wiehe (1998) have investigated the hitchhiking effects on population differentiation using the island and stepping-stone models of population structure, and presented the analytical expression of F_{st} suitable for the case of two subpopulations. The present model is an extension of the classical island model (Wright, 1969) to incorporate effects of background selection and necessarily expands the study of Charlesworth et al. (1997) to an arbitrary number of subpopulations simultaneously. Further, when the haploid dispersal of pollen is set to zero, our F_{st} formula can be applied to animal populations where only diploid dispersal occurs.

Until now empirical studies have been concentrated on the observations of high F_{st} in the regions of low recombination as a result of hitchhiking effect, such as in *Drosophia* species (e.g., Stephan and Mitchell, 1992; Begun and Aquadro, 1993). Much less attention has been paid to the effects of background selection on F_{st}. A recent study shows that the average level of nucleotide diversity in regions of low recombination can be used to distinguish background selection from hitchhiking effects (Innan and Stephan, 2003). Because population differentiation of a neutral locus is negatively related to the genetic diversity within subpopulations given a total genetic variation, an interesting question is whether the measure of F_{st} can be used to distinguish background selection from hitchhiking effects. The properties of F_{st} under hitchhiking effects (Slatkin and Wiehe, 1998) and background selection (present study) are very similar, displaying a negative correlation with recombination fraction. The "spatial" pattern of F_{st} along chromosomes is likely very similar between these two types of processes, and this problem presents a challenge for future study.

The present theory provides an implemental technique to examine the magnitude of background selection effects. The conventional approaches for estimating population differentiation at neutral loci remain valid (e.g., Weir, 1996). The problem is how to distinguish which loci are purely neutral and not affected by background selection or hitchhiking effects. With molecular genome sequence data, F_{st} can be calculated for all individual SNP that are distributed on the same chromosome, using the method introduced by Hudson et al. (1992), and hence the pattern of genome-wide F_{st} can be mapped. Those neutral loci with the smallest F_{st} values can be used for approximating $F_{st.b}$, and the effects of background selection at other individual SNP can be estimated according to Eq. (21). The prerequisites for such genomic-wide F_{st} mapping are the tests of neutrality and background selection.

Acknowledgements

We sincerely appreciate Richard A. Ennos and two referees for insightful comments. This work was financially supported from the Department of Renewable Resources, University of Alberta.

Appendix A. Changes in the conditional probability in the two-locus case

Let $P^*_{A_iC_i}$ and $P^*_{a_iC_i}$ be the frequencies of gametes A_iC_i and a_iC_i after selection in the ith population, respectively. The mean fitness in the ith subpopulation, denoted by $\bar{w}_i$, is approximated by $\bar{w}_i = 1 - 2p_{a_i}s_{1i}$. The frequency of each two-locus gamete after selection ($P^*_{A_iC_i}$ and $P^*_{a_iC_i}$) can be calculated using the conventional method,

$$\begin{aligned} P^*_{A_iC_i} = {} & ((1 - p_{a_i}s_{1i})P_{A_iC_i} + \delta_0 r_1)/\bar{w}_i \\ & - (u_1 + v)P_{A_iC_i} - \tilde{m}(P_{A_iC_i} - \bar{P}_{AC}), \end{aligned} \tag{A.1}$$

$$\begin{aligned} P^*_{a_iC_i} = {} & ((1 - (1 + p_{a_i})s_{1i})P_{a_iC_i} - \delta_0 r_1)/\bar{w}_i \\ & + (u_1 - v)P_{A_iC_i} - \tilde{m}(P_{a_iC_i} - \bar{P}_{aC}), \end{aligned} \tag{A.2}$$

where $\delta_0 = P_{A_i}P_{a_i}(x_{1i} - x_{0i})(1 - s_{1i})$. The changes in gamete frequency ($\Delta P_{A_iC_i} = P^*_{A_iC_i} - P_{A_iC_i}$ and $\Delta P_{a_iC_i} = P^*_{a_iC_i} - P_{a_iC_i}$) can be calculated from Eqs. (A.1) and (A.2).

From the relation of $P_{A_iC_i} = p_{A_i}x_{0i}$, we obtained $\Delta P_{A_iC_i} = \Delta p_{A_i}x_{0i} + p_{A_i}\Delta x_{0i}$. Since $\Delta p_{A_i} = 0$ at steady state, the change in the conditional probability of x_{0i} is given by

$$\begin{aligned} \Delta x_{0i} = \Delta P_{A_iC_i}/p_{A_i} = {} & r_1 p_{a_i}(1 - s_{1i})(1 + 2p_{a_i}s_{1i})(x_{1i} - x_{0i}) \\ & + (p_{a_i}s_{1i} - u_1 - v - \tilde{m})x_{0i} + \tilde{m}\bar{p}_{A_iC_i}/p_{A_i}, \end{aligned} \tag{A.3}$$

where $\bar{p}_{A_iC_i}$ is the frequency of the gamete A_iC_i in migrants. Similarly, the change in the conditional probability x_{1i} is given by

$$\begin{aligned} \Delta x_{1i} = \Delta P_{a_iC_i}/p_{a_i} = {} & r_1 p_{A_i}(1 - s_{1i})(1 + 2p_{a_i}s_{1i})(x_{0i} - x_{1i}) \\ & + (p_{a_i}s_{1i} - s_{1i} - v - \tilde{m})x_{1i} + u_1 p_{A_i}x_{0i}/p_{a_i} \\ & + \tilde{m}\bar{p}_{a_iC_i}/p_{a_i}, \end{aligned} \tag{A.4}$$

where $\bar{p}_{a_iC_i}$ is the frequency of the gamete a_iC_i in migrants. Since the genetic drift process does not change the average gamete frequency, Eqs. (A.3) and (A.4) actually represent the per-generation changes in the conditional probabilities of x_{0i} and x_{1i}, respectively.

Appendix B. Recurrent equation for the LD between two selected loci

Let $D_{AB(i)}$ be the LD between the A and B loci in the ith subpopulation in the current generation (adults), and r be the recombination rate between them. The gamete frequencies at the current generation, denoted by P_{jl} ($j = A_i, a_i; l = B_i, b_i$), can be expressed as $P_{jl} = p_j p_l + \delta_{jl}D_{AB(i)}$, where $\delta_{A_iB_i} = \delta_{a_ib_i} = 1$ and $\delta_{A_ib_i} = \delta_{a_iB_i} = -1$. The gamete frequencies in pollen and ovules in the next generation, denoted by P'_{jl}, can be expressed by $P'_{jl} = p_j p_l + (1 - r)\delta_{jl}D_{AB(i)}$. After pollen flow the gamete frequencies in pollen, denoted by P''_{jl}, can be expressed by $P''_{jl} = m_P\bar{P}_{jl} + (1 - m_P)P'_{jl}$, where $\bar{P}_{jl}$ is the gamete frequency in migrants. The gamete frequencies in ovules remain the same as those in the preceding adults.

After random combination between pollen and ovules, the nine genotypic frequencies in seeds so formed can be readily calculated. For example, the frequency for the genotype $A_iA_iB_iB_i$ in seeds, denoted by $P'_{A_iA_iB_iB_i}$, equals $P_{A_iB_i}P''_{A_iB_i}$. The frequency for the genotype $A_iA_iB_iB_i$ after seed flow can be expressed as $P''_{A_iA_iB_iB_i} = m_S\bar{P}_{AABB} + (1 - m_S)P'_{A_iA_iB_iB_i}$, where $\bar{P}_{AABB}$ is the frequency in migrating seeds. The frequencies for other genotypes after seed flow can be expressed in a similar way. Following the procedure of mutation and selection, the genotypic frequencies in adults in the next generation can be derived using the conventional method. The frequencies of four gametes after selection in adults, denoted by P^*_{jl} ($j = A_i, a_i; l = B_i, b_i$), are

$$P^*_{A_iB_i} = \tilde{m}\bar{P}_{AB} + (1 - \tilde{m} + p_{a_i}s_{1i} + p_{b_i}s_{2i} - u_1 - u_2)P'_{A_iB_i}, \tag{B.1a}$$

$$P^*_{A_ib_i} = \tilde{m}\bar{P}_{Ab} + (1 - \tilde{m} + p_{a_i}s_{1i} - p_{B_i}s_{2i} - u_1)P'_{A_ib_i} + u_2P'_{A_iB_i}, \tag{B.1b}$$

$$P^*_{a_iB_i} = \tilde{m}\bar{P}_{aB} + (1 - \tilde{m} - p_{A_i}s_{1i} + p_{b_i}s_{2i} - u_2)P'_{a_iB_i} + u_1P'_{A_iB_i}, \tag{B.1c}$$

$$P^*_{a_ib_i} = \tilde{m}\bar{P}_{ab} + (1 - \tilde{m} - p_{A_i}s_{1i} - p_{B_i}s_{2i})P'_{a_ib_i} + u_1P'_{A_ib_i} + u_2P'_{a_iB_i}. \tag{B.1d}$$

Let $D^*_{AB(i)}$ be the LD in the ith subpopulation in the next adult generation. According to Eqs. (B.1a)–(B.1d), the recurrent equation for LD between the A and B loci is derived as

$$\begin{aligned} D^*_{AB(i)} &= P^*_{A_iB_i}P^*_{a_ib_i} - P^*_{A_ib_i}P^*_{a_iB_i} \\ &= \tilde{m}(\bar{D}_{AB} + (p_{A_i} - \bar{p}_A)(p_{B_i} - \bar{p}_B)) \\ &\quad + (1 - \tilde{m} - (p_{A_i} - p_{a_i})s_{1i} \\ &\quad - (p_{B_i} - p_{b_i})s_{2i} - u_1 - u_2)(1 - r)D_{AB(i)}, \end{aligned} \tag{B.2}$$

where $\bar{D}_{AB}$ is the LD in migrants.

Appendix C. Changes in the conditional probability in the three-locus case

Let $P^*_{A_iC_iB_i}$, $P^*_{A_iC_ib_i}$, $P^*_{a_iC_iB_i}$, and $P^*_{a_iC_ib_i}$ be the frequencies of gametes $A_iC_iB_i$, $A_iC_ib_i$, $a_iC_iB_i$, and $a_iC_ib_i$ after selection in the ith population, respectively. Using the assumption of multiplicative viability model, the mean fitness in the ith subpopulation equals $\bar{W}_i = 1 - 2p_{a_i}s_{1i} - 2p_{b_i}s_{2i}$. Using the same approach as in the two-locus case, the frequencies of the four three-locus gametes are given by

$$\begin{aligned} P^*_{A_iC_iB_i} &= ((1 - s_{1i}p_{a_i} - s_{2i}p_{b_i})P_{A_iC_iB_i} - \delta_1r_1 \\ &\quad - \delta_2r_2 + \delta_0)/\bar{W}_i - (u_1 + u_2 + v)P_{A_iC_iB_i} \\ &\quad - \tilde{m}(P_{A_iC_iB_i} - \bar{P}_{ACB}), \end{aligned} \tag{C.1a}$$

$$\begin{aligned} P^*_{a_iC_iB_i} &= ((1 - s_{1i}(1 + p_{a_i}) - s_{2i}p_{b_i})P_{a_iC_iB_i} \\ &\quad + \delta_1r_1 + \delta_3r_2 - \delta_0)/\bar{W}_i + u_1P_{A_iC_iB_i} - u_2P_{a_iC_iB_i} \\ &\quad - vP_{a_iC_iB_i} - \tilde{m}(P_{a_iC_iB_i} - \bar{P}_{aCB}), \end{aligned} \tag{C.1b}$$

$$\begin{aligned} P^*_{A_iC_ib_i} &= ((1 - s_{1i}p_{a_i} - s_{2i}(1 + p_{b_i}))P_{A_iC_ib_i} \\ &\quad + \delta_4r_1 + \delta_2r_2 - \delta_0)/\bar{W}_i - u_1P_{A_iC_ib_i} + u_2P_{A_iC_iB_i} \\ &\quad - vP_{A_iC_ib_i} - \tilde{m}(P_{A_iC_ib_i} - \bar{P}_{ACb}), \end{aligned} \tag{C.1c}$$

$$\begin{aligned} P^*_{a_iC_ib_i} &= ((1 - s_{1i}(1 + p_{a_i}) - s_{2i}(1 + p_{b_i}))P_{a_iC_ib_i} \\ &\quad - \delta_4r_1 - \delta_3r_2 + \delta_0)/\bar{W}_i + u_1P_{A_iC_ib_i} + u_2P_{a_iC_iB_i} \\ &\quad - vP_{a_iC_ib_i} - \tilde{m}(P_{a_iC_ib_i} - \bar{P}_{aCb}), \end{aligned} \tag{C.1d}$$

where

$$\begin{aligned} \delta_1 &= P_{A_iB_i}P_{a_iB_i}(1 - s_{1i})(y_{0i} - y_{1i}) \\ &\quad + (P_{A_iB_i}P_{a_ib_i}y_{0i} - P_{A_ib_i}P_{a_iB_i}y_{1i})(1 - s_{1i} - s_{2i}), \end{aligned}$$

$$\begin{aligned} \delta_2 &= P_{A_iB_i}P_{A_ib_i}(1 - s_{2i})(y_{0i} - y_{2i}) \\ &\quad + (P_{A_iB_i}P_{a_ib_i}y_{0i} - P_{A_ib_i}P_{a_iB_i}y_{2i})(1 - s_{1i} - s_{2i}), \end{aligned}$$

$$\begin{aligned} \delta_3 &= P_{a_ib_i}P_{a_iB_i}(1 - 2s_{1i} - s_{2i})(y_{3i} - y_{1i}) \\ &\quad + (P_{A_iB_i}P_{a_ib_i}y_{3i} - P_{A_ib_i}P_{a_iB_i}y_{1i})(1 - s_{1i} - s_{2i}), \end{aligned}$$

$$\begin{aligned} \delta_4 &= P_{A_ib_i}P_{a_ib_i}(1 - s_{1i} - 2s_{2i})(y_{3i} - y_{2i}) \\ &\quad + (P_{A_iB_i}P_{a_ib_i}y_{3i} - P_{A_ib_i}P_{a_iB_i}y_{2i})(1 - s_{1i} - s_{2i}), \end{aligned}$$

$$\begin{aligned} \delta_0 &= r_1r_2(1 - s_{1i} - s_{2i})(P_{A_iB_i}P_{a_ib_i}(y_{0i} + y_{3i} - y_{0i}y_{3i}) \\ &\quad - P_{A_ib_i}P_{a_iB_i}(y_{1i} + y_{2i} - y_{1i}y_{2i})). \end{aligned}$$

δ_0 is associated with the effect of double crossover among the three loci.

From the expression of $P_{A_iC_iB_i} = P_{A_iB_i}y_{0i}$, we obtain

$$\Delta P_{A_iC_iB_i} = \Delta P_{A_iB_i}y_{0i} + P_{A_iB_i}\Delta y_{0i}. \tag{C.2}$$

Since $\Delta P_{A_iB_i} = 0$ at steady state, the change in the conditional probability of y_{0i} is

$$\begin{aligned} \Delta y_{0i} &= \Delta P_{A_iC_iB_i}/P_{A_iB_i} \\ &= (s_{1i}p_{a_i} + s_{2i}p_{b_i} - u_1 - u_2 - v - \tilde{m})y_{0i} \\ &\quad + (-\delta_1r_1 - \delta_2r_2 + \delta_0)(1 + 2s_{1i}p_{a_i} + 2s_{2i}p_{b_i})/P_{A_iB_i} \\ &\quad + \tilde{m}\bar{P}_{ACB}/P_{A_iB_i}, \end{aligned} \tag{C.3a}$$

where $\Delta P_{A_iC_iB_i} = P^*_{A_iC_iB_i} - P_{A_iC_iB_i}$ and $\bar{p}_{ACB}$ is the frequency of gamete $A_iC_iB_i$ in migrants.

Similarly, the changes in the conditional probabilities of y_{1i}, y_{2i}, and y_{3i} are derived as

$$\begin{aligned} \Delta y_{1i} &= (-s_{1i}p_{A_i} + s_{2i}p_{b_i} + u_1 - u_2 - v - \tilde{m})y_{1i} \\ &\quad + (\delta_1r_1 + \delta_3r_2 - \delta_0)(1 + 2s_{1i}p_{a_i} + 2s_{2i}p_{b_i})/P_{a_iB_i} \\ &\quad + \tilde{m}\bar{P}_{aCB}/P_{a_iB_i}, \end{aligned} \tag{C.3b}$$

$$\begin{aligned} \Delta y_{2i} &= (s_{1i}p_{a_i} - s_{2i}p_{B_i} - u_1 + u_2 - v - \tilde{m})y_{2i} \\ &\quad + (\delta_4r_1 + \delta_2r_2 - \delta_0)(1 + 2s_{1i}p_{a_i} + 2s_{2i}p_{b_i})/P_{A_ib_i} \\ &\quad + \tilde{m}\bar{P}_{ACb}/P_{A_ib_i}, \end{aligned} \tag{C.3c}$$

$$\begin{aligned}\Delta y_{3i} &= (-s_{1i}p_{A_i} - s_{2i}p_{B_i} + u_1 + u_2 - v - \tilde{m})y_{3i} \\ &\quad + (-\delta_4 r_1 - \delta_3 r_2 + \delta_0)(1 + 2s_{1i}p_{a_i} + 2s_{2i}p_{b_i})/P_{a_ib_i} \\ &\quad + \tilde{m}\bar{p}_{a_iC_ib_i}/p_{a_ib_i}, \qquad \text{(C.3d)}\end{aligned}$$

where $\bar{p}_{ACb}$, $\bar{p}_{aCB}$, and $\bar{p}_{aCb}$ are the frequencies of gametes $A_iC_ib_i$, $a_iC_iB_i$, and $a_iC_ib_i$ in migrants, respectively.

References

Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y.M., Lench, N.J., Carey, A., Cardon, L.R., Moffatt, M.F., Cookson, W.O.C., 2001. Extent and distribution of linkage disequilibrium in three genomic regions. Am. J. Hum. Genet. 68, 191–197.

Barton, N.H., 2000. Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B 355, 1553–1562.

Begun, D.J., Aquadro, C.F., 1993. African and North American populations of *Drosophila melanogaster* are very different at DNA level. Nature 365, 548–550.

Bennett, J.H., 1954. On the theory of random mating. Ann. Eugenic. 18, 311–317.

Caballero, A., Hill, W.G., 1992. Effective size of non-random mating populations. Genetics 130, 909–916.

Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289–1303.

Charlesworth, B., Morgan, M.T., Charlesworth, D., 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. 70, 155–174.

Farnir, F., Coppieters, W., Arranz, J.J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., Georges, M., 2000. Extensive genome-wide linkage disequilibrium in Cattle. Genome Res. 10, 220–227.

Goddard, K.A., Hopkins, P.J., Hall, J.M., Witte, J.S., 2000. Linkage disequilibrium and allele-frequency distribution for 114 single-nucleotide polymorphisms in five populations. Am. J. Hum. Genet. 66, 216–234.

Hill, W.G., 1974. Disequilibrium among several linked neutral genes in finite populations. I. Mean changes in disequilibrium. Theor. Popul. Biol. 5, 366–392.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231.

Hu, X.S., 2000. A preliminary approach to the theory of geographical gene genealogy for plant genomes with three different models of inheritance and its application. Acta Genetica Sinica 27, 440–448.

Hu, X.S., Ennos, R.A., 1999. Impacts of seed and pollen flow on population differentiation for plant genomes with three contrasting modes of inheritance. Genetics 152, 441–450.

Hu, X.S., Li, B.L., 2002. Seed and pollen flow and cline discordance among genes with different modes of inheritance. Heredity 88, 212–217.

Hudson, R.R., Kaplan, N.L., 1995. Deleterious background selection with recombination. Genetics 141, 1605–1617.

Hudson, R.R., Slatkin, M., Aguadé, M., 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583–589.

Innan, H., Stephan, W., 2003. Distinguishing the hitchhiking and background selection models. Genetics 165, 2307–2312.

Li, W.H., Nei, M., 1974. Stable linkage disequilibrium without epistasis in subdivided populations. Theor. Popul. Biol. 6, 173–183.

Maynard Smith, J., Haigh, J., 1974. The hitch-hiking effect of a favorable gene. Genet. Res. 23, 23–35.

Nordborg, M., 1997. Structured coalescent process on different time scales. Genetics 146, 1501–1514.

Nordborg, M., Charlesworth, B., Charlesworth, D., 1996. The effect of recombination on background selection. Genet. Res. 67, 159–174.

Rafalski, J.A., 2002. Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. 162, 329–333.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., Darvasi, A., 2003. Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. 12, 771–776.

Slatkin, M., 1975. Gene flow and selection in a two-locus system. Genetics 81, 787–802.

Slatkin, M., Wiehe, T., 1998. Genetic hitch-hiking in a subdivided population. Genet. Res. 71, 155–160.

Stephan, W., Mitchell, S.J., 1992. Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics 132, 1039–1045.

Wang, D.G., Fan, J.B., Siao, C.J., et al., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280, 1077–1082.

Weir, B.S., 1996. Genetic Data Analysis II. Methods for Discrete Population Genetic Data. Sinauer Associates, Sunderland, MA.

Wright, S., 1943. Isolation by distance. Genetics 28, 114–138.

Wright, S., 1969. Evolutionary and the Genetics of Populations. vol. 2. The Theory of Gene Frequencies. The University of Chicago Press, Chicago.