Genet. Res., Camb. (2002), **79**, *pp.* 129–139. *With 6 figures.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672301005493 Printed in the United Kingdom

Extending the coalescent to multilocus systems: the case of balancing selection

NICK H. BARTON AND ARCADIO NAVARRO*
Institute of Cell, Animal and Population Biology, University of Edinburgh, UK

(*Received 18 May 2001 and in revised form 31 August 2001*)

Summary

Natural populations are structured spatially into local populations and genetically into diverse 'genetic backgrounds' defined by different combinations of selected alleles. If selection maintains genetic backgrounds at constant frequency then neutral diversity is enhanced. By contrast, if background frequencies fluctuate then diversity is reduced. Provided that the population size of each background is large enough, these effects can be described by the structured coalescent process. Almost all the extant results based on the coalescent deal with a single selected locus. Yet we know that very large numbers of genes are under selection and that any substantial effects are likely to be due to the cumulative effects of many loci. Here, we set up a general framework for the extension of the coalescent to multilocus scenarios and we use it to study the simplest model, where strong balancing selection acting on a set of n loci maintains 2^n backgrounds at constant frequencies and at linkage equilibrium. Analytical results show that the expected linked neutral diversity increases exponentially with the number of selected loci and can become extremely large. However, simulation results reveal that the structured coalescent approach breaks down when the number of backgrounds approaches the population size, because of stochastic fluctuations in background frequencies. A new method is needed to extend the structured coalescent to cases with large numbers of backgrounds.

1. Introduction

The increasing wealth of information on DNA sequence polymorphism that has accumulated during the past two decades has stimulated a variety of studies that describe the patterns of neutral DNA diversity that are to be expected under different evolutionary forces, and that use those patterns to detect natural selection. Broadly, the effects of selection on linked neutral variability depend on parameters such as the population size (N), the neutral mutation rate (μ), the recombination rate between the selected and the neutral loci (r) and, of course, the strength of natural selection (s). The general case is difficult, but a considerable amount of progress has been made for strong selection ($Ns \gg 1$) using the structured coalescent, which follows back in time the genealogies of sets of genes that can be in different locations or associated with different alleles (cf. Hudson, 1990; Hey, 1991; Nordborg, 1997). A powerful heuristic argument used in those studies is that the effect of selection on linked neutral variability is analogous to the effect of spatial subdivision. When a population is subdivided into demes of constant size, its effective population size is increased and neutral variability can be larger than in a panmictic population (Nagylaki, 1982). By contrast, if the sizes of demes fluctuates, for example by an extinction-recolonization process, population size decreases and neutral variability is lost (Whitlock & Barton, 1997). If one thinks in terms of the coalescent times, stable subdivision increases them because it takes time to move from one location to another in order to coalesce. Fluctuations make times shorter because lineages only tend to trace back to the limited number of successful demes.

Natural populations are also structured into diverse backgrounds (i.e. haplotypes) defined by different combinations of selected alleles. Just as with spatial subdivision, genetic structure influences neutral di-

* Corresponding author. Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK. Tel: +44 (0) 131 650 5513. Fax: +44 (0) 131 650 6564. e-mail: arcadi@holyrood.ed.ac.uk

versity. Diversity at linked neutral loci is reduced if background frequencies fluctuate more than expected by drift either because selection eliminates deleterious variants (Charlesworth *et al*., 1993; Hudson & Kaplan, 1994; Charlesworth, 1994; Hudson & Kaplan, 1995) or because it forces the fixation of advantageous alleles (Kaplan *et al*., 1989; Stephan *et al*., 1992; Aquadro & Begun, 1993; Aquadro *et al*., 1994; Barton, 1998). By contrast, if balancing selection maintains genetic backgrounds at constant frequencies then neutral diversity is enhanced (Strobeck, 1983; Hudson & Kaplan, 1988; Kaplan *et al*., 1988; Hey, 1991; Nordborg, 1997).

The analogy with spatial subdivision, however, has limitations. Genetic structure is more complex, and harder to study, than its spatial analogue. Genomes are formed by very large numbers of genes (e.g. $\sim 3\times 10^4$ in humans) upon which selection can act, and substantial effects of neutral variability are only expected through the accumulated effects of many loci. Thus, the relevant backgrounds will be defined by combinations of alleles from several (perhaps many) different loci, possibly spanning large regions of the genome. This implies that one needs to consider genealogies that can be transferred between backgrounds by complex recombination events, rather than by simple migration. Owing to these difficulties, almost all the theoretical results obtained up to now deal with a single selected locus. Some multilocus results have been produced for purifying selection, where the exact multilocus coalescent can be approximated by a simpler version, allowing the consideration of classes of backgrounds instead of the backgrounds themselves. Genetic backgrounds are classified according to the number of deleterious mutations that they harbour. Backgrounds within a class are considered equivalent, so one only needs to follow lineages across classes and can ignore the exact background they are associated with (Charlesworth *et al*., 1993; Hudson & Kaplan, 1994; Charlesworth, 1994; Hudson & Kaplan, 1995). Other forms of multilocus selection, such as balancing selection, do not yield so easily to simplification and multilocus results are only starting to emerge. For example, the patterns of neutral sequence variation generated by balancing selection acting upon two diallelic loci that interact epistatically has been recently investigated by Kelly and Wade (2000). Yet, the way in which multilocus balancing selection may be affecting linked neutral variability in a genome-wide scenario is still poorly understood.

Here, we start by setting up a general analytical method that extends the structured coalescent to any situation where selection is strong enough (i.e. Ns is large enough) for background frequencies to be constant. Then we apply the method to a model in which a neutral locus is linked to a set of selected loci where diallelic polymorphisms are maintained at equilibrium by strong balancing selection. We choose this model because of its simplicity. Frequencies can considered to be stable and no mutation in the selected loci needs to be taken into account. We study variability by focusing on a pairwise measure: the average probability of identity in state between two randomly chosen alleles. By doing so, we are studying the distribution of pairwise coalescent times, of which identities are the Laplace transform (Hudson, 1990). Afterwards, we check the validity of the analytical approach by simulations. We aim to answer the question of how much neutral DNA sequence variability can be inflated by balancing selection at many linked loci.

2. Results

(i) *Framework*

In a random-mating diploid population of fixed size N, the probability that two genes are identical by descent (F) from the previous generation is $1/(2N)$. The identity by descent of two alleles in the next generation is:

$$F' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F = F + \frac{1-F}{2N}. \tag{1}$$

This well-known result can be easily extended to the multilocus case following the approach of Maruyama (1972; generalized by Nagylaki, 1982) for a spatially structured population and adapting it to genetical structure. Suppose that genetic variation at many loci defines a set of genetic backgrounds, labelled i. The elements of the backward matrix, Γ_{ij}, give the chance that a gene that is presently in background i was in background j in the previous generation. The probability of i.b.d. in a given generation between a gene in background i and a gene in background j is given by F_{ij}

$$F'_{ij} = \sum_k \sum_l \left(\Gamma_{ik}\Gamma_{jl} \left(F_{kl} + \frac{\delta_{kl}(1-F_{kk})}{2N_k} \right) \right), \tag{2}$$

where $\delta_{kl} = 1$, $\delta_{kl} = 0$ if $k \neq l$ and N_k is the population size of background k. The recursions for the identity in allelic state, f, are:

$$f'_{ij} = z^2 \sum_k \sum_l \left(\Gamma_{ik}\Gamma_{jl} \left(f_{kl} + \frac{\delta_{kl}(1-f_{kk})}{2N_k} \right) \right), \tag{3}$$

where $z = 1-\mu$ (μ is the mutation rate) and f can be regarded as the generating function or the Laplace Transform for the distribution of coalescence times (Hudson, 1990).

These recursions can be simplified by changing the co-ordinates so as to diagonalize $\mathbf{\Gamma}$. Let the eigenvalues

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

of $\mathbf{\Gamma}$ be λ_α; the matrix of left eigenvectors is $\nu_{\alpha i}$, and the matrix of right eigenvectors is $\eta_{i\alpha}$. The leading eigenvalue is $\lambda_1 = 1$, with η_{i1} independent of i, and corresponds to complete mixing. ν_{1i} gives the ultimate contribution that a gene currently in background i will make to future generations.

Following Nagylaki (1982), we can define

$$f_{ij} = \sum_\alpha \sum_\beta \eta_{i\alpha}\eta_{j\beta} f_{\alpha\beta} \quad \text{and} \quad f_{\alpha\beta} = \sum_j \sum_k \nu_{\alpha j}\nu_{\beta k} f_{jk}, \tag{4}$$

where

$$\sum_i \nu_{\alpha i}\eta_{i\beta} = \delta_{\alpha\beta} \quad \text{and} \quad \sum_\alpha \eta_{i\alpha}\nu_{\alpha j} = \delta_{ij}. \tag{5}$$

Thus, in the new coordinate system, Equation (3):

$$f'_{\alpha\beta} = z^2\lambda_\alpha\lambda_\beta \left(f_{\alpha\beta} + \sum_k \frac{\nu_{\alpha k}\nu_{\beta k}(1-f_{kk})}{2N_k} \right). \tag{6}$$

Hence, at equilibrium, assuming that the background frequencies remain constant over time:

$$f_{\alpha\beta} = \sum_k \frac{z^2\lambda_\alpha\lambda_\beta\nu_{\alpha k}\nu_{\beta k}(1-f_{kk})}{2N_k(1-z^2\lambda_\alpha\lambda_\beta)}. \tag{7}$$

This gives a simple formula for the full set of identities, $f_{\alpha\beta}$, in terms of the diagonal elements, f_{kk}.

(ii) *Simple examples: one and two selected loci*

Let us start with the simplest example: a single selected locus. Suppose that two genetic backgrounds are defined by the segregation of two alleles at a single locus (P, Q), at frequencies p and q at a given generation. Selection (or drift) then alters the frequencies to p', q' at the next generation. A linked neutral locus is associated with one or other of these backgrounds; the recombination rate is r.

The forward transition matrix is $\mathbf{M}$; element M_{ij} gives the chance that a gene now in background i will be in background j in the next generation. It is obtained as the product of two matrices; the first accounting for the change in background frequency under selection or drift; the second, for recombination in a pool with allele frequencies p' and q'.

$$\mathbf{M} = \begin{pmatrix} \dfrac{(1-rq')p'}{p} & \dfrac{rq'p'}{p} \\ \dfrac{rq'p'}{q} & \dfrac{q'(1-rp')}{q} \end{pmatrix} = \begin{pmatrix} 1-rq' & rq' \\ rp' & 1-rp' \end{pmatrix}. \tag{8}$$

The elements in the background matrix Γ_{kl} give the chance that a gene that is presently in background k was in background l in the previous generation. It can be obtained by applying Bayes' rule and normalizing. In the single locus case, it is

$$\mathbf{\Gamma} = \begin{pmatrix} 1-rq' & rq' \\ rp' & 1-rp' \end{pmatrix}. \tag{9}$$

Note that the forward and the backward matrices are identical because selection or drift do not move alleles between backgrounds (i.e. they are not transitions).

The eigenvectors and eigenvalues have very simple forms

$$\boldsymbol{\lambda} = (1,\ 1-r) \tag{10}$$

$$\boldsymbol{\nu} = \begin{pmatrix} q' & p' \\ 1 & -1 \end{pmatrix} \tag{11}$$

$$\boldsymbol{\eta} = \begin{pmatrix} 1 & p' \\ 1 & -q' \end{pmatrix}. \tag{12}$$

Consider now the case of two selected loci, each segregating for two alleles P and Q with frequencies p_i and q_i. In this case, there are four possible backgrounds: P_1P_2, P_1Q_2, Q_1P_2 and Q_1Q_2. We define the four background frequencies as b_1, b_2, b_3 and b_4 in one generation and b'_1, b'_2, b'_3 and b'_4 in the next. Assume a linear map $n-1-2$ (where n stands for 'neutral') as depicted in Fig. 1 and no interference.

In this case, the backward transition matrix is (we drop all the prime symbols)

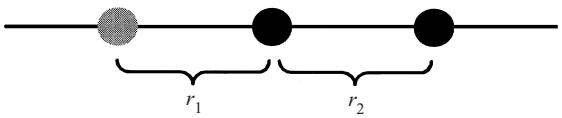

Fig. 1. The neutral locus (gray) lies at the left of the set of selected loci (black).

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

$$\boldsymbol{\Gamma} = \begin{pmatrix} \frac{b_1(b_1+\bar{r}_1(b_3+b_4\bar{r}_2)+b_2(r_1r_2+\bar{r}_1+\bar{r}_2))}{b_2b_3r_2-b_1(-1+b_4r_2)} \\ \frac{b_1(b_4r_2\bar{r}_1+b_2(r_2\bar{r}_1+r_1\bar{r}_2))}{b_1b_4r_2+b_2(1-b_3r_2)} \\ \frac{b_1r_1(b_3+b_4r_2)}{b_1b_4r_2+b_3(1-b_2r_2)} \\ \frac{b_1b_4r_1\bar{r}_2}{b_2b_3r_2+b_4(1-b_1r_2)} \end{pmatrix}$$

$$\begin{pmatrix} \frac{b_2(b_3r_2\bar{r}_1+b_1(r_2\bar{r}_1+r_1\bar{r}_2))}{b_2b_3r_2+b_1(1-b_4r_2)} \\ \frac{b_2(b_2+\bar{r}_1(b_4+b_3\bar{r}_2)+b_1(r_1r_2+\bar{r}_1\bar{r}_2))}{b_1b_4r_2-b_2(-1+b_3r_2)} \\ \frac{b_2b_3r_1\bar{r}_2}{b_1b_4r_2+b_3(1-b_2r_2)} \\ \frac{b_2r_1(b_4+b_3r_2)}{b_2b_3r_2+b_4(1-b_1r_2)} \end{pmatrix} \cdots \tag{13}$$

$$\begin{pmatrix} \frac{b_3r_1(b_1+b_2r_2)}{b_2b_3r_2+b_1(1-b_4r_2)} \\ \frac{b_2b_3r_1\bar{r}_2}{b_1b_4r_2+b_2(1-b_3r_2)} \\ \frac{b_3(b_3+\bar{r}_1(b_1+b_2\bar{r}_2)+b_4(r_1r_2+\bar{r}_1\bar{r}_2))}{b_1b_4r_2-b_3(-1+b_2r_2)} \\ \frac{b_3(b_2r_2\bar{r}_1+b_4(r_2\bar{r}_1+r_1\bar{r}_2))}{b_2b_3r_2+b_4(1-b_1r_2)} \end{pmatrix}$$

$$\begin{pmatrix} \frac{b_1b_4r_1\bar{r}_2}{b_2b_3r_2+b_1(1-b_4r_2)} \\ \frac{b_4r_1(b_2+b_1r_2)}{b_1b_4r_2+b_2(1-b_3r_2)} \\ \frac{b_4(b_1r_2\bar{r}_1+b_3(r_2\bar{r}_1+r_1\bar{r}_2))}{b_1b_4r_2+b_3(1-b_2r_2)} \\ \frac{b_4(b_4+\bar{r}_1(b_2+b_1\bar{r}_2)+b_3(r_1r_2+\bar{r}_1\bar{r}_2))}{b_2b_3r_2-b_4(-1+b_1r_2)} \end{pmatrix},$$

where r_1 and r_2 are according to Fig. 1 and where $\bar{r}_i = 1-r_i$.

Although the approach we use here is entirely general and can be applied to diverse multilocus scenarios, calculations become easier when it is assumed that there is no linkage disequilibrium between selected loci, so the frequency of each background is just the product of allelic frequencies. In this case, the matrix can be greatly simplified.

$$\boldsymbol{\Gamma} = \begin{pmatrix} q_1\bar{r}_1(p_2+q_2\bar{r}_2)+p_1(p_2+q_2(r_1r_2+\bar{r}_1\bar{r}_2)) \\ p_2(q_1r_2\bar{r}_1+p_1(r_2\bar{r}_1+r_1\bar{r}_2)) \\ p_1r_1(p_2+q_2r_2) \\ p_1p_2r_1\bar{r}_2 \end{pmatrix}$$

$$\begin{pmatrix} q_2(q_1r_2\bar{r}_1+p_1(r_2\bar{r}_1+r_1\bar{r}_2)) \\ q_1\bar{r}_1(q_2+p_2\bar{r}_2)+p_1(q_2+p_2(r_1r_2+\bar{r}_1\bar{r}_2)) \\ p_1q_2r_1\bar{r}_2 \\ p_1r_1(q_2+p_2r_2) \end{pmatrix} \cdots \tag{14}$$

$$\begin{pmatrix} q_1r_1(p_2+q_2r_2) \\ p_2q_1r_1\bar{r}_2 \\ p_2(q_1+p_1\bar{r}_1)+q_2(p_1\bar{r}_1\bar{r}_2+q_1(r_1r_2+\bar{r}_1\bar{r}_2)) \\ p_2(p_1r_2\bar{r}_1+q_1(r_2\bar{r}_1+r_1\bar{r}_2)) \end{pmatrix}$$

$$\begin{pmatrix} q_1q_2r_1\bar{r}_2 \\ q_1r_1(q_2+p_2r_2) \\ q_2(p_1r_2\bar{r}_1+q_1(r_2\bar{r}_1+r_1\bar{r}_2)) \\ p_1\bar{r}_1(q_2+p_2\bar{r}_2)+q_1(q_2+p_2(r_1r_2+\bar{r}_1\bar{r}_2)) \end{pmatrix}.$$

Then, its eigenvectors and eigenvalues are

$$\boldsymbol{\lambda} = (1, \bar{r}_1, \bar{r}_1\bar{r}_2+r_1r_2, \bar{r}_1\bar{r}_2) \tag{15}$$

$$\boldsymbol{\nu} = \begin{pmatrix} q_1q_2 & q_1p_2 & p_1q_2 & p_1p_2 \\ -q_2 & -p_2 & q_2 & p_2 \\ -q_1 & q_1 & -p_1 & p_1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \tag{16}$$

$$\boldsymbol{\eta} = \begin{pmatrix} 1 & -p_1 & -p_2 & p_1p_2 \\ 1 & -p_1 & q_2 & -p_1q_2 \\ 1 & q_1 & -p_2 & -p_2q_1 \\ 1 & q_1 & q_2 & q_1q_2 \end{pmatrix}, \tag{17}$$

where r_1 and r_2 are according to Fig. 1 and where $\bar{r}_i = 1-\bar{r}_i$.

The eigenvectors are products of contributions from each locus. The first left eigenvector (first row of ν) corresponds to the steady state, where each background contributes in proportion to its frequency; in the long term, there is the same expected marker frequency in each background (first column of η). The second eigenvalue corresponds to linkage disequilibrium of the neutral marker with just the first locus, which breaks down at rate $(1-r_1)$. The third eigenvalue corresponds to linkage disequilibrium with just the second locus, which breaks down at a rate equal to the chance that there will be no recombination of the neutral marker with the second locus, i.e. $(1-r_1)(1-r_2)+r_1r_2$. The fourth eigenvalue expresses three-way disequilibrium. A different transition matrix is obtained when the neutral marker lies between the two selected loci (map 1-*n*-2). Its eigenvectors, however, are identical to the ones in (16) and (17); only the eigenvalues differ:

$$\boldsymbol{\lambda} = (1, \bar{r}_1, \bar{r}_2, \bar{r}_1\bar{r}_2). \tag{18}$$

Assuming constant background frequencies, one can easily plug eigenvalues and eigenvectors into (7) to obtain the equilibrium identities.

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

(iii) *Many loci*

The previous examples extend to many loci in a straightforward way by means of the multilocus machinery developed by Barton and Turelli (1991) and the principal components approach of Bennet (1954; see Dawson, 2000 for an insightful update and extension). The eigenvectors of the transition matrix are products across loci, which correspond to the steady decay of various higher-order linkage disequilibria. They have a simple form because of the key assumption that the backgrounds are in linkage equilibrium. Therefore, only the disequilibria of the marker with each set of selected loci, which decay geometrically (Bennet, 1954; Barton & Turelli, 1991; Dawson, 2000), need to be tracked. If the backgrounds were in linkage disequilibrium then the eigenvectors would be extremely complicated functions of recombination rates.

(iv) *Defining eigenvalues*

We first define the multilocus eigenvectors and find the eigenvalues in terms of the recombination rates. This requires no assumptions about the map. However, with equally spaced loci on a linear map and no interference, the eigenvalues simplify. We then find the identities to leading order in $\frac{1}{N}$, and give a recursion for the higher-order terms.

It is convenient to describe the backgrounds by a vector $\mathbf{X}$, with values $X_i = 0$ or 1 for alleles P or Q. The eigenvectors correspond to linkage disequilibria between sets of loci U (set (2,4,9), for example, includes loci 2, 4 and 9). We write the eigenvectors as functions of $\mathbf{X}$, $\nu_{\mathrm{U}}(\mathbf{X})$, $\eta_{\mathrm{U}}(\mathbf{X})$. Let $S_i = 2X_i - 1$, so that $S_i = \pm 1$. Let $g_i = (1-X_i) + S_i p_i$ (so that $g_i = q_i$ or p_i); $\bar{g}_i = X_i - S_i p_i$ (so that $\bar{g}_i = p_i$ or q_i). Let $\delta_i(\mathbf{X}, \mathbf{Y}) = 1$ if $X_i = Y_i$, 0 otherwise. We will use the convention that $S_{\mathrm{U}} = \prod_{i \in \mathrm{U}} S_i$, etc. The complete set of loci is L. Thus, the frequency of the gamete containing the set of loci U is g_{U}.

The eigenvectors are

$$\nu_{\mathrm{U}} = S_{\mathrm{U}} g_{\mathrm{L}\setminus\mathrm{U}} \quad \eta_{\mathrm{U}} = S_{\mathrm{U}} \bar{g}_{\mathrm{U}}. \tag{19}$$

Each of the terms in (19) is a function of the genotype $\mathbf{X}$. To show that these eigenvectors are orthogonal, one only has to sum their product over $\mathbf{X}$

$$\sum_{\mathbf{X}} \nu_{\mathrm{U}} \eta_{\mathrm{V}} = \sum_{\mathbf{X}} S_{\mathrm{U}} S_{\mathrm{V}} g_{\mathrm{L}\setminus\mathrm{U}} \bar{g}_{\mathrm{V}} = \delta_{\mathrm{U,V}}. \tag{20}$$

Because each of the elements in (5) is a product across loci, this sum can be simplified by separating it into terms corresponding to the four kinds of loci: those in U and V, in U but not V, in V but not U, and in neither. These contribute $\bar{g}_i; S_i; S_i g_i \bar{g}_i; g_i$, respectively. Because the second and third terms sum to zero, and the first and last terms sum to 1, we have $\delta_{\mathrm{U,V}}$, as required.

(v) *The effect of recombination of the identities*

Now, consider the effect of recombination. Let the chance that the marker remains associated with the set S be r_S. Thus $r_\emptyset = 1$ (where $\emptyset$ stands for the empty set), and for a single selected locus i and recombination rate c, $r_{\{i\}} = 1-c$, (we use c rather than r to avoid confusion). Assuming a linear map and n equally spaced loci, with no interference, the chance of generating k junctions is $c^k(1-c)^{n-k}$. This leads to an expression for r_S. The transition matrix which gives the numbers that move from background $\mathbf{X}$ to $\mathbf{Y}$ is

$$\Gamma(\mathbf{X}, \mathbf{Y}) = \sum_{\mathrm{S}} r_S \delta_S(\mathbf{X}, \mathbf{Y}) g_{\mathrm{L}\setminus\mathrm{S}}(\mathbf{Y}). \tag{21}$$

Multiplying by $\nu_{\mathrm{U}}(\mathbf{X})$, $\nu_{\mathrm{V}}(\mathbf{Y})$ shows that these are indeed eigenvectors of $\mathbf{\Gamma}$, and that the corresponding eigenvalues are just r_{V}.

(vi) *The leading term in $1/N$*

We can find the leading term to $O(\frac{1}{N})$ by setting f_{kk} to zero in (7). This leads to

$$f^*_{\mathrm{U,V}} = \frac{z^2 \lambda_{\mathrm{U}} \lambda_{\mathrm{V}} \delta_{\mathrm{U,V}}}{2N p_{\mathrm{U}} q_{\mathrm{U}} (1 - z^2 r^2_{\mathrm{U}})} + O\left(\frac{1}{N^2}\right), \tag{22}$$

where $f^*_{\mathrm{U,V}}$ is a linear transformation of the identity $f(\mathbf{X}, \mathbf{Y})$ between genes chosen from backgrounds $\mathbf{X}$, $\mathbf{Y}$. It corresponds to the covariance between random fluctuations in linkage disequilibria, just as f for a single locus corresponds to the variance in allelic frequencies. (22) shows that fluctuations in disequilibria involving different sets of loci are independent ($\delta_{\mathrm{U,V}}$), and their variance depends on the probability of remaining associated with that set (r_{U}).

Notice that the average identity between randomly chosen genes is $f_{\emptyset,\emptyset} = z^2/(2N(1-z^2))$, which is just the same as for an unstructured population. This seems to contradict the decrease in identity which one expects for markers linked to a balanced polymorphism. However, it can be shown that this decrease is of order $1/N^2$ and so does not appear in (22).

(vii) *The full solution: transforming the identity within backgrounds*

To get the full solution (valid for smaller N), one needs to solve a matrix equation for the vector $f(\mathbf{X}, \mathbf{X})$, which is the diagonal element of the full matrix of pairwise identities. It is easier to solve by transforming to $f_{\mathrm{U}} = \sum_{\boldsymbol{X}} \nu_{\mathrm{U}}(\mathbf{X}) f(\mathbf{X}, \mathbf{X})$. Note that this is not the same as $f_{\mathrm{U,U}}$, given by (22).

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

Table 1. *The eight classes of terms of the matrices* **H** *and* **H*** *(25). Each matrix is the product of some of these terms according to the presence (1) or absence (0) of loci in the sets* W, U, V *or* Z, U, V

$H_{W,U,V}$				$H^*_{Z,U,V}$			
Presence/Absence			Term	Presence/Absence			Term
W	U	V		Z	U	V	
0	0	0	1	0	0	0	1
0	0	1	0	0	0	1	0
0	1	0	0	0	1	0	0
0	1	1	pq	0	1	1	$1/(pq)$
1	0	0	0	1	0	0	0
1	0	1	1	1	0	1	1
1	1	0	1	1	1	0	1
1	1	1	$(q-p)$	1	1	1	$(q-p)/(pq)$

Applying this transformation to (7), and after some tedious algebra, we have

$$f_{U,V} = \frac{1}{2N}\sum_W \frac{z^2 r_U r_V H^*_{W,U,V}}{(1-z^2 r_U r_V)}(\delta_{W,\emptyset} - f_W) \tag{23}$$

and

$$f_W = \sum_{U,V} H_{W,U,V} f_{U,V}, \tag{24}$$

where

$$H_{W,U,V} = \sum_{\mathbf{X}} \nu_W \eta_U \eta_V, \quad H^*_{Z,U,V} = \sum_{\mathbf{X}} \frac{\eta_Z \nu_U \nu_V}{g_L} \tag{25}$$

Note that $(1-f_{X,X})$ transforms to $(\delta_{Z,\emptyset} - f_Z)$. The matrices **H**, **H*** can be found by separating the sums over **X** into contributions from the eight classes of loci that are defined by belonging or not to the three sets U, V and W. The matrix is the product of contributions (Table 1). The driving term involves $\sum_Z \delta_{Z,\emptyset} H^*_{Z,U,V} = H^*_{\emptyset,U,V} = \delta_{U,V}/(p_U q_U)$. The remaining sum involves $H_{W,U,U} = [(q-p)/pq]_W$ for all U that contain W, and zero otherwise. The driving term (which is just the transform of (22)) is therefore

$$f^*_W = \left(\frac{1}{2N}\right)\left(\frac{q-p}{pq}\right)_W \Lambda_{W,W}, \tag{26}$$

where

$$\Lambda_{W,W} = \sum_V \frac{1}{(1-z^2 r^2_{WV})}. \tag{27}$$

As we can see (23), the full solution is rather complex, and obtaining numerical values for more than, say, five loci becomes excruciatingly slow. However, several simplifications can be made.

(viii) *Simplifying* Λ *for particular models of recombination*

The $\Lambda_{U,V}$ terms simplify for particular models of recombination. Consider $\Lambda_{\emptyset,\emptyset}$. For a linear map with at most one crossover per generation (i.e. total interference), we have $r_U = 1-kc$, where k is the distance spanned by U (if the marker is embedded within it) or the distance from the marker to the furthest locus if the marker is outside U. Suppose there are n_+ loci to the right and n_- to the left; the marker is αc from the right-hand locus and $(1-\alpha)c$ from the left-hand locus.

$$\begin{aligned}\Lambda_{\emptyset,\emptyset} = \frac{1}{1-z^2} &+ \sum_{j=0}^{n_+-1} \frac{2^j}{1-z^2(1-(\alpha+j)c)^2} \\ &+ \sum_{j=0}^{n_--1} \frac{2^j}{1-z^2(1-(1-\alpha+j)c)^2} \\ &+ \sum_{j=1}^{n_++n_--1} \frac{2^{j-1}(n_++n_--j)}{1-z^2(1-jc)^2}\end{aligned} \tag{28}$$

where j is the number of junctions. We have used the fact that there are 2^j sets which have all the loci to the left of the neutral locus, and the leftmost locus $(j+\alpha)c$ away; and there are $2^{j-1}(n_++n_- -j)$ sets which span a distance jc and which include the neutral locus.

(ix) *The average identity*

Because the actual background to which a neutral allele is linked is likely to be unknown, we focus on the average identity between gametes chosen at random from the whole population, $f_{\emptyset,\emptyset}$. From (23), this is

$$f_{\emptyset,\emptyset} = \frac{z^2(1-f_\emptyset)}{2N(1-z^2)}. \tag{29}$$

The coefficient $f_\emptyset$ is the average identity between randomly chosen genes within the same background. It increases with tighter linkage.

(x) *An explicit solution for equal allele frequencies*

Suppose that we have slow change ($\mu, r, N^{-1} \ll 1$). Let $r_U = 1-\rho_U$, $z = 1-\mu$. In this case, (23) becomes

$$f_{U,V} = \sum_Z \frac{H^*_{Z,U,V}}{(4N\mu + 2N\rho_U + 2N\rho_V)}(\delta_{Z,\emptyset} - f_Z). \tag{30}$$

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

Assuming a linear map, the recombination rates ρ_U are just the total recombination rates between the limits of the sets (including the marker). Overall, we have a set of linear equations in the f_Z

$$f_W = \sum_{U,V} \sum_{Z} \frac{H_{W,U,V} H^*_{Z,U,V}}{(4N\mu + 2N\rho_U + 2N\rho_V)} (\delta_{Z,\emptyset} - f_Z). \quad (31)$$

In particular, the average identity between randomly chosen gametes is

$$f_{\emptyset,\emptyset} = \frac{(1 - f_\emptyset)}{4N\mu}, \quad (32)$$

where

$$f_\emptyset = \sum_{U,V} \sum_{Z} \frac{H_{\emptyset,U,V} H^*_{Z,U,V}}{(4N\mu + 2N\rho_U + 2N\rho_U\rho_V)} (\delta_{Z,\emptyset} - f_Z). \quad (33)$$

The array $H_{\emptyset,U,V}$ only has contributions $p_U\, q_U$ from $U = V$, and from $Z \subseteq U$:

$$f_\emptyset = \sum_{U} \frac{1}{(4N\mu + 4N\rho_U)} \left(1 - \sum_{Z \subseteq U} (q-p)_Z f_Z \right). \quad (34)$$

Further simplification is achieved by assuming that some form of symmetrical balancing selection maintains equal and even frequencies at all the selected loci ($p_i = q_i = \frac{1}{2}$). Then, $f_\emptyset$ can be easily expressed as

$$f_\emptyset = 1 - \frac{1}{1 + \sum_{U} \frac{1}{(4N\mu + 4N\rho_U)}}. \quad (35)$$

Recall that U stands for all the possible sets of loci and ρ_U is the total recombination between the limits of each set. These results could be extended to arbitrary linear maps, because all that matters is the chance of a recombination event that dissociates the marker from some set of loci.

Some values predicted using (35) and (32) are plotted in Fig. 2. The predicted identity decreases (i.e. diversity increases) steadily with the number of backgrounds. In fact, as the number of backgrounds becomes large, the identity becomes absurdly small. Examining (35), one can see an obvious reason for the increase in variability predicted by the structured coalescent: the second term in the denominator is a sum across all possible sets of loci that involves $\sim 2^n$ summands and, therefore, can become an enormous quantity. In fact, with a large number of loci, the number of summands may become so big that mutation rates can become almost irrelevant. (32) and (35) predict that, as the number of backgrounds increases, $f_\emptyset \to 1$ and $f_{\emptyset,\emptyset} \to 0$ independently of mutation rates, because the backgrounds will eventually diverge, even for very low mutation. But, of course, as the number of backgrounds approaches population size (N), it is impossible to maintain all of them in the population. For example, if we consider 10 selected

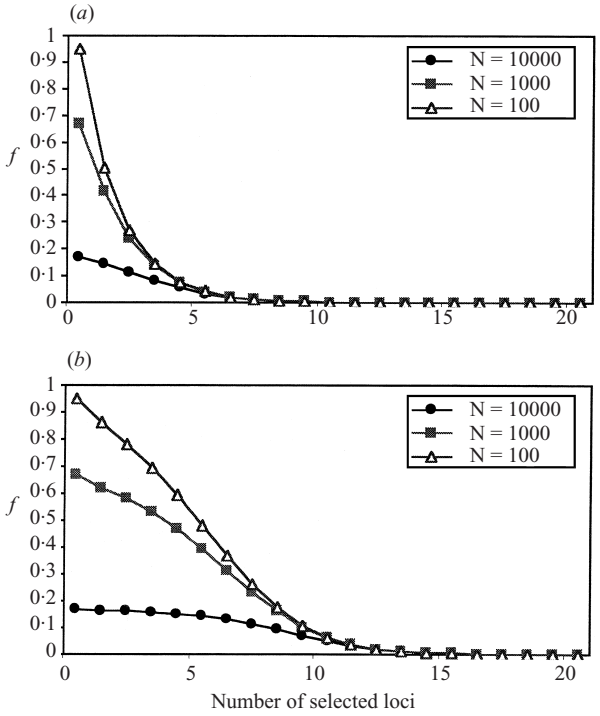

Fig. 2. Theoretically expected identities between two randomly chosen gametes for three different population sizes. All the loci (both selected and neutral) are evenly spaced and the neutral locus lies at an extreme of the map (Fig. 1). Selection maintains even frequencies at all the selected loci. $\mu = 1{\cdot}25 \times 10^{-4}$. (*a*) $r = 10^{-5}$. (*b*) $r = 10^{-3}$.

loci, the number of possible backgrounds is 1024, and a population of, say, size 100 is clearly unable to sustain such a level of subdivision. Although alleles are maintained, some of the backgrounds they define must be absent, rendering the population far less structured than our analytical results assume. The number of backgrounds grows exponentially with the number of loci. For example, 20 loci produce $\sim 10^6$ possible backgrounds, and 24 loci $\sim 1.7 \times 10^7$, so any population is bound to lose selected polymorphism if the number of loci is large. In the following, we use simulations to check that the structured coalescent breaks down for quite small numbers of loci even when N is large.

(xi) *Simulations*

The coalescent approach developed above relies on two related assumptions: that every background is abundant; and that its frequency is both known and constant. These assumptions will, in principle, be valid for strong and constant selection (i.e. Ns large and s constant) but, when many selected loci are considered, the equations predict an unrealistically large hitchhiking effect (Fig. 2), with identity dropping almost to zero. To find out why and under which parameter values the theory breaks down, we simulated a multilocus system. Details of the simulation

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

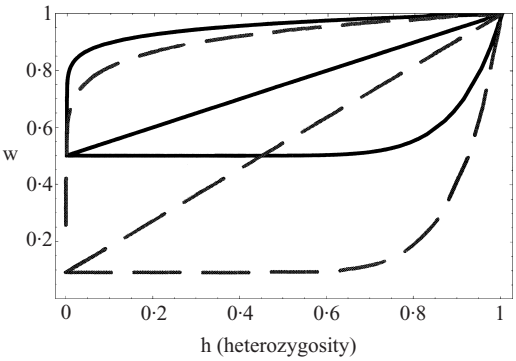

Fig. 3. Some instances of the fitness function. To make values comparable, we divide by $1+\alpha$. Continuous lines $\alpha = 1$; dashed lines $\alpha = 10$. Straight lines $k = 1$ (additive fitness); concave lines $k = 10$ (positive epistasis); convex lines $k = 0{\cdot}1$ (negative epistasis).

system can be found in a companion paper (Navarro & Barton, 2002). Its basics are as follows: the program performs forward simulations of a life cycle with selection and recombination for a population of size N. As in the previous section, we consider a number of selected loci, each segregating for two alleles, and a neutral locus lying alongside them. Selected loci are assumed to be spaced at equal intervals along the genetic map, and the recombination between any two adjacent loci is r. The mutation rate at the neutral locus is μ. All the selected loci have heterozygote advantage and selection is symmetrical at each locus. Simulation values throughout this article were obtained by running the population to drift-selection equilibrium and calculating identities between randomly chosen gametes (f) in the last generation. Every f value plotted in the graphs is the average of 10 runs.

The coalescent approach presented is entirely general and does not assume any specific multilocus fitness regime. Making some assumptions, we have used it to study a polymorphic equilibrium where selection tries to maintain all the possible 2^n backgrounds at constant frequencies and without linkage disequilibrium among selected loci. Such an equilibrium can be reproduced under the n-locus symmetric viability model with negative epistasis (Karlin & Avni, 1981; Christiansen, 1987, 1988, 1990; Barton & Shpak, 2000). This is the model we used in our simulations. For simplicity, we assumed that all loci contribute equally to fitness and that selection acts only on the proportion of heterozygosity of an individual. The fitness of an individual was given by $w\ (\alpha, h, k) = 1+\alpha h^k$, where h is the proportion of heterozygous loci in a given individual ($0 \leqslant h \leqslant 1$) and α is the strength of selection. Epistasis enters the function by means of k, a parameter that allows for different selective regimes (Fig. 3). If $k < 1$, there is negative epistasis, so selection favours decreased variance in heterozygosity and, at equilibrium, all possible backgrounds are present in the population

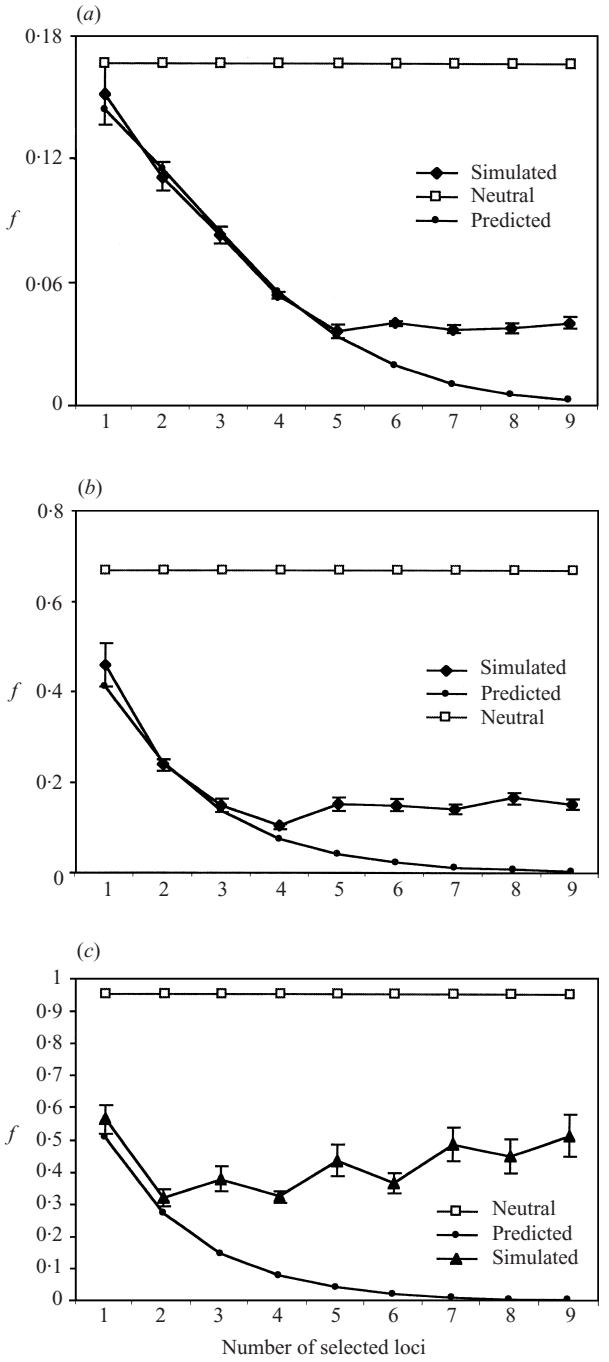

Fig. 4. Identities ($\pm$SE) at a neutral locus with an increasing number of selected loci. All the loci (both selected and neutral) are evenly spaced. The neutral locus lies at an extreme of the map (Fig. 1). $\mu = 1{\cdot}25 \times 10^{-4}$, $r = 10^{-5}$, $\alpha = 1$ and $k = 0{\cdot}1$. (*a*) $N = 10^4$. (*b*) $N = 10^3$. (*c*) $N = 10^2$.

with no linkage disequilibrium (Christiansen, 1987, 1988). If $k > 1$, there is positive epistasis, so selection favours increased variance in heterozygosity. In this case, the equilibrium population has maximum linkage disequilibrium and, when linkage is complete, it is formed by two complementary backgrounds. If $k = 1$, fitness is additive and selection only affects mean heterozygosity (i.e. individual loci), ignoring the

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

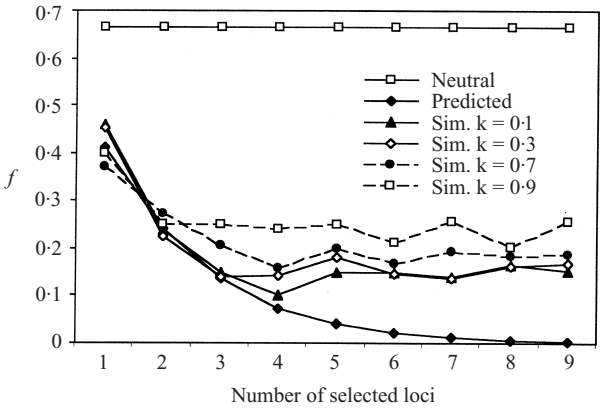

Fig. 5. Identities at a neutral locus with an increasing number of selected loci and different degrees of negative epistasis. Loci distribution and parameters as in Fig. 4 with $N = 10^3$.

variance in heterozygosity and thus ignoring linkage disequilibrium. Detailed results on neutral variability under these different fitness schemes can be found in Navarro & Barton (2002). Here, we only concern ourselves with the case of negative epistasis because it allows straightforward comparisons with the analytical coalescent approach.

Fig. 4 shows predicted and simulated changes in the value of identity as the number of selected loci increases, for different population sizes. For multilocus systems with strong negative epistasis and few selected loci, the analytical predictions hold quite well when compared with the simulations. With few loci, selection is successfully forcing all the possible backgrounds to be present at even and constant frequencies. The population is as subdivided as the theory assumes and neutral variability increases proportionally to the level of subdivision. As expected, when the number of loci in the system becomes too large, the analytical and simulation results start to diverge. Of course, the larger the population size, the more selected loci are needed for theoretically predicted identities to diverge from simulations. Fig. 5 shows another factor affecting the divergence of simulations form analytical results: the strength of epistasis in the system. The closer the fitness scheme is to additivity, the more important is the divergence. As we discuss below, the cause of this divergence is that, when the number of loci is large and/or epistasis is zero or positive, the population is not as subdivided as the theory assumes because fewer genetic backgrounds are maintained.

3. Discussion

There is a substantial body of theory on the effects that balancing selection acting on a single selected diallelic locus is expected to have on linked neutral variability (Strobeck, 1983; Hudson & Kaplan, 1988; Kaplan *et al*., 1988; Hudson 1990; Nordborg 1997). It has been shown that the neutral variants linked to a given selected allele (background) constitute a different 'subpopulation' and that differentiation between subpopulations increases the predicted sequence variability (Hudson & Kaplan, 1988). Recently, Kelly and Wade (2000) extended this single-locus theory to a two-locus scenario and studied neutral sequence variation linked to a pair of balanced polymorphisms held in strong linkage disequilibrium by epistatic selection. Their main result is that neutral variability can be increased across the intervening region for a longer distance than expected with a single selected locus. As we detail elsewhere (Navarro & Barton, 2002), this two-locus selective scenario is a special case of the fitness model studied here (specifically, it is equivalent to symmetric overdominance with positive epistasis, i.e., $k > 1$ in our fitness function).

Here, we have developed a general method to apply coalescent theory to multilocus scenarios. We have used the method to extend the existing theory to consider an arbitrary number of loci under balancing selection. The extension seems to work fairly well for small number of loci but, when the number of loci becomes too big, an absurdly large effect is predicted (Fig. 2). To perform a preliminary exploration of the causes of this behavior, we have simulated the effects on neutral variability of multilocus balancing selection with negative epistasis. We find that such selection has the potential to enhance neutral variability far beyond the predictions for a single selected locus, because it allows many backgrounds to be maintained in the population. There is, however, a limit to that increase. Simulations show that variability stops increasing and seems to stabilize about a certain value, independent of the addition of more selected loci. An obvious explanation is that, even with extremely strong selection, the population size imposes a logical limit to the number of backgrounds that can coexist. However, this can hardly explain the results in Figs 4 and 5. A close examination of these figures shows that simulations diverge from the coalescent predictions even when the number of backgrounds is relatively small compared with population size. (e.g. in Fig. 4a, divergence starts with 64 backgrounds for $N = 10^4$). Stochastic fluctuations in background frequencies are crucial for this divergence.

The increase of variability produced by multilocus balancing selection is opposed by drift. With stable background frequencies, drift acts independently within each subpopulation of size $N/2^n$. This effect is taken into account by our analytical model. When the number of backgrounds is large enough, another source of drift becomes important. The number of backgrounds grows exponentially with the number of loci and, although selection may still be very strong on each individual locus and allele frequencies are close

https://doi.org/10.1017/S0016672301005493 Published online by Cambridge University Press

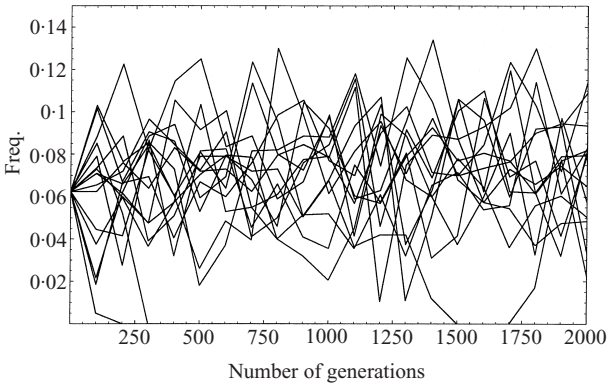

Fig. 6. Stochastic fluctuations of the frequencies of the 16 genetic backgrounds defined by four diallelic loci in a population of size $N = 10^3$. Each line stands for the frequency of a background. Backgrounds had even frequencies at the first generation. Selected loci are evenly spaced $r = 10^{-5}$, $\alpha = 1$, $k = 0{\cdot}1$.

to deterministic expectations, selection quickly becomes very weak on each individual background. In such circumstances, drift is not only caused by random sampling within each background but is also associated with stochastic fluctuations in background frequencies (Fig. 6). These fluctuations are not considered in the analytical model, in which frequencies were assumed to be constant. To put things in another way: the fluctuations violate the assumption of no linkage disequilibrium between selected loci because they generate random associations among them. These random associations render the population less structured than assumed by the coalescent approach, so that the increase in neutral variability is smaller than predicted. Moreover, if fluctuations are very strong, some backgrounds will be eliminated by drift and the neutral variability associated with them will be swept away. Backgrounds can eventually be recovered by recombination (Fig. 6), whereas the neutral variability lost with them can only be replaced by mutation.

The exact multilocus coalescent will be a useful tool to study balancing selection systems provided that the number of relevant backgrounds is not very large and their frequencies are constant. It will be useful to study relatively simple multilocus systems in species with small effective population size, such as humans, but more complex systems could be studies by this approach in species with larger population size, such as *Drosophila*. Of course, some knowledge is needed about which loci are the targets of selection in order to know how many backgrounds can potentially be present in the population. One of the advantages of the coalescent approach is that the exact kind of selection acting upon the system is not important. It does not matter if genetic backgrounds are maintained by, for example, some form of overdominance (as assumed here) or by frequency-dependent selection, as long as their frequencies are kept constant. If this is not the case, some way to account for perturbations in background frequencies must be introduced. In principle, one could consider genealogies conditioning on the random series of background frequencies and then average over these sequences. This has been done to study the earlier phases of selective sweeps, where fluctuations in the frequency of the favoured allele are important (Barton, 1998). Applying the same strategy to multilocus balancing selection is difficult for several reasons. First, the strength of perturbations does depend on the form of selection in the system, which in our case includes epistasis and is therefore far more complex than simple positive selection. Second, a proper analysis must take simultaneously into account fluctuations in all the backgrounds, which means taking into account the whole population all time, rather than only the favoured allele for just a few generations. Finally, in the case of selective sweeps, one can make the simplifying assumption that the favoured allele is destined for fixation, whereas, in the case of multilocus balancing selection, one must account for the loss and recovery of backgrounds by drift and recombination. The case for the application of our extension of the structured coalescent to the study of multilocus balancing selection is made worse by removing some of the simplifying assumptions that underlie our model. First, our analytical model assumes that the selection–drift equilibrium has been reached, but Kelly and Wade (2000) show that the patterns of variability expected during the approach to equilibrium in a two-locus model are quite different than the ones expected at equilibrium. In a multilocus scenario, equilibrium takes much longer to achieve, so the relevance of equilibrium variability predictions is doubtful. Second, although the analytical approach developed here is completely general, our application of it to balancing selection focused on symmetric overdominance acting on several equally spaced, diallelic loci that contributed equally to fitness. The stochastic fluctuations in background frequencies that affect the validity of the coalescent will be even more difficult to account for in more realistic systems in which fitness is not symmetric; selection acts with different strength in different loci and the interactions between loci are not governed by exactly the same kind of epistasis.

An alternative is not to use the exact structured coalescent, but an approximation. This has been successfully done in the case of purifying selection. In that case, the coalescent is structured into different classes of backgrounds with different numbers of deleterious mutations rather than into different individual backgrounds. This approach is based on two approximations: first, that all the backgrounds harbouring a given number of mutations are selectively equivalent; and, second, that the fitness of a gamete

depends only on the mutations it carries, so that one does not need to consider zygotes (Charlesworth *et al*., 1993; Hudson & Kaplan, 1994; Charlesworth, 1994; Hudson & Kaplan, 1995). Unfortunately, neither of these approximations is valid under balancing selection. There is no similar way to classify different backgrounds into equivalent classes because heterozygosity and linkage disequilibrium are crucial. The fitness of a given gamete depends on the zygotes it will form and, therefore, on the other gametes in the population.

It seems clear that, in the case of balancing selection, the study of multilocus genealogies must take into account stochastic fluctuations on each of the possible genetic backgrounds. This can be done by our forward simulations, which we have used extensively to study the effects of balancing selection on several measures of neutral variability (Navarro & Barton, 2002).

We thank B. Charlesworth, D. Charlesworth and F. Depaulis for valuable discussion and criticism. We are also grateful to an anonymous reviewer, who pointed out an imprecision in the original manuscript. This work was supported by the BBSRC.

References

Aquadro, C. F. & Begun, D. J. (1993). *Evidence for and implications of genetic hitch-hiking in the Drosophila genome*. In *Mechanisms of molecular evolution*. (ed. N. Takahata & A. G. Clark), pp. 159–178. Sunderland, MA: Sinauer Press.

Aquadro, C. F. *et al*. (1994). *Selection, recombination and DNA polymorphism in Drosophila*. In *Non-neutral evolution* (ed. B. Golding), pp. 46–56. New York: Chapman & Hall.

Barton, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetical Research* **72**, 123–133.

Barton, N. H. & Shpak, M. (2000). The stability of symmetrical solutions to polygenic models. *Theoretical Population Biology* **57**, 249–263.

Barton, N. H. & Turelli, M. (1991). Natural and sexual selection on many loci. *Genetics* **127**, 229–255.

Barton, N. H. & Wilson, I. (1995). Genealogies and geography. *Philosophical Transactions of the Royal Society* **349**, 49–59.

Bennett, J. H. (1954). On the theory of random mating. *Annals of Eugenics* **18**, 311–317.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* **63**, 213–228.

Charlesworth, B. *et al*. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Christiansen, F. B. (1987). The deviation from linkage equilibrium with multiple loci varying in a stepping-stone cline. *Journal of Genetics* **66**, 45–67.

Christiansen, F. B. (1988). Epistasis in the multiple locus symmetric viability model. *Journal of Mathematical Biology* **26**, 595–618.

Christiansen, F. B. (1990). Simplified models for viability selection at multiple loci. *Theoretical Population Biology* **37**, 39–54.

Dawson, K. J. (2000). The decay of linkage disequilibrium under random union of gametes: how to calculate Bennet's Principal Components. *Theoretical Population Biology* **58**, 1–20.

Hey, J. (1991). A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

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

Hudson, R. B. & Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R. R. & Kaplan, N. L. (1994). Gene trees with background selection. In *Alternatives to the Neutral Model*. (ed. G. B. Golding). Chapman Hall: Pages?

Hudson, R. R. & Kaplan, N. L. (1995). The coalescent process with background selection. *Philosophical Transactions of the Royal Society Series B* **349**, 19–23.

Kaplan, N. L., Darden, T. & Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.

Kaplan, N. L., Hudson, R. R. & Langley, C. H. (1989). The hitch-hiking effect revisited. *Genetics* **123**, 887–899.

Karlin, S. & Avni, H. (1981). Analysis of central equilibria in multilocus systems. *Theoretical Population Biology* **20**, 241–280.

Kelly, J. K. & Wade, M. J. (2000). Molecular evolution near a two-locus balanced polymorphism. *Journal of Theoretical Biology* **204**, 83–101.

Maruyama, T. (1972). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics* **70**, 639–651.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.

Navarro, A. & Barton, N. H. (2002). The effects of multilocus balancing selection on neutral variability. *Genetics*, (in press).

Nordborg, M. (1997). Structured coalescent processes on different timescales. *Genetics* **146**, 1501–1514.

Stephan, W, Wiehe, T. H. & Lenz, M. (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology* **41**, 237–254.

Strobeck, C. (1983). Expected linkage disequilibrium for a neutral locus linked to a chromosome rearrangement. *Genetics* **103**, 545–555.

Whitlock, M. C. & Barton, N. H. (1997). The effective size of a subdivided population. *Genetics* **146**, 427–441.