The Hitchhiking Effect on Linkage Disequilibrium between Linked Neutral Loci

Wolfgang Stephan,[*] **Yun S. Song,**[†] **and Charles H. Langley**[‡]

[*] *Section of Evolutionary Biology, Biocenter, University of Munich,*
82152 Planegg-Martinsried, Germany

[†] *Department of Computer Sciences, University of California, Davis, California 95616*

[‡] *Section of Evolution and Ecology, University of California, Davis, California 95616*

Running head:
Hitchhiking and linkage disequilibrium

Keywords:
Positive directional selection, genetic hitchhiking, linkage disequilibrium

Corresponding author:
Wolfgang Stephan
Section of Evolutionary Biology, Biocenter, University of Munich,
Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany
Phone: +49 89 2180 74102
Fax: +49 89 2180 74104
Email: stephan@zi.biologie.uni-muenchen.de

ABSTRACT

We analyzed a three-locus model of genetic hitchhiking with one locus experiencing positive directional selection and two partially linked neutral loci. Following the original hitchhiking approach by Maynard Smith and Haigh, our analysis is purely deterministic. In the first half of the selected phase after a favored mutation has entered the population, hitchhiking may lead to a strong increase of linkage disequilibrium (LD) between the two neutral sites if both are less than $0.1s$ away from the selected site (where s is the selection coefficient). In the second half of the selected phase, the main effect of hitchhiking is to destroy LD. This occurs very quickly (before the end of the selected phase) when the selected site is between both neutral loci. This pattern cannot be attributed to the well-known variation-reducing effect of hitchhiking but is a consequence of secondary hitchhiking effects on the recombinants created in the selected phase. When the selected site is outside the neutral loci (which are, say, less than $0.1s$ apart), however, a fast decay of LD is only observed if the selected site is in the immediate neighborhood of one of the neutral sites (i.e., if the recombination rate r between the selected site and one of the neutral sites satisfies $r \ll 0.1s$). If the selected site is far away from the neutral sites (say, $r > 0.3s$), the decay rate of LD approaches that of neutrality. Averaging over a uniform distribution of initial gamete frequencies shows that the expected LD at the end of the hitchhiking phase is driven toward zero, while the variance is increased when the selected site is well outside the two neutral sites. When the direction of LD is polarized with respect to the more common allele at each neutral site, hitchhiking creates more positive than negative linkage disequilibrium. Thus, hitchhiking may have a distinctively patterned LD-reducing effect, in particular near the target of selection.

Genetic drift is recognized as a fundamental stochastic force shaping the polymorphism within populations and divergence between species of both neutral and selected variants at a locus (Fisher 1930; Wright 1931; Kimura 1983). Remarkably similar patterns as those caused by genetic drift are predicted by theoretical models incorporating stochastically varying selection (Gillespie 1994). In 1974 Maynard Smith and Haigh analyzed a simple and obvious extension of the genetic drift models, incorporating the stochastic coupling due to linkage with another locus undergoing strong directional selection. Addressing the apparent uniformity of allozyme polymorphism across species, Maynard Smith and Haigh (1974) focused on the "hitchhiking effect" on allelic frequencies and heterozygosity.

Since this seminal work the hitchhiking effect associated with positive directional selection has been extended to address observations in the emerging field of molecular population genetics (Aguadé *et al.* 1989; Stephan and Langley 1989; Begun and Aquadro 1992). These studies revealed low levels of DNA sequence polymorphism in *Drosophila* in genomic regions of low crossing-over, and led to theoretical analyses of the hitchhiking effect on nucleotide diversity (Kaplan *et al.* 1989) and on the frequency spectrum of polymorphisms (Braverman *et al.* 1995). Whether formulated in terms of gamete frequencies or in the coalescent framework, the genetic models dealt with a pair of loci: a selected and a linked neutral locus with a defined rate of recombination between them. Independent of whether single or recurrent hitchhiking events were analyzed (as a stochastic process or a deterministic approximation), the conclusion has consistently been that hitchhiking should have profound effects on expected heterozygosity and allele frequencies at the neutral locus, if selection is strong and linkage is tight.

In 1977 Thomson considered the impact of hitchhiking on linkage disequilibrium (LD). Most of her analysis focused on the association between the selected alleles and those at a single neutral locus. Thomson also considered the impact of hitchhiking (of a heterotic polymorphism) on the LD between alleles at two linked neutral loci. Based on numerical examples she concluded that hitchhiking creates LD between neutral loci within the same genomic domain in which it affects levels of heterozygosity. The impact of simple directional selection of rare variants rapidly going to fixation was not considered by Thomson (1977). Except for this initial study and a few rather targeted applications (Robinson *et al.* 1991; Grote *et al.* 1998) no attempts have been made to extend this simple two-locus hitchhiking model of Maynard Smith and Haigh (1974) to multiple loci. This is curious, as the hitchhiking effect has played a major role in molecular population genetics for more than 15 years.

Based on the analyses of Thomson and her colleagues, it has been believed that hitchhiking not only affects polymorphisms at individual sites (or loci), but also the association between polymorphisms. Using a three-locus model with one selected and two neutral loci, she showed that hitchhiking that has a strong effect on heterozygosity can also generate strong LD between the neutral loci. Without paying close attention to Thomson's writing, the important role of hitchhiking in generating LD has been reiterated in textbooks and publications by many authors. It was not until recently that a few authors reported quite the opposite results (Gillespie 1997; Kim and Stephan 2002; Kim and Nielsen 2004). Analyzing "shift" models that are similar to the hitchhiking model described above, Gillespie (1997) concluded that "linked selection can reduce variation without building up high levels of linkage disequilibrium, contrary to our intuition". These latter studies focused on average effects observable in simulated data. In small-sample coalescent simulations, Kim and Nielsen (2004) found increased LD between alleles at two neutral loci on the same side of the selected locus at the time of fixation, and reduced LD across the site of selection. Furthermore, they provide heuristic arguments to explain this pattern. These different and somewhat contradictory views of the relationship between genetic hitchhiking and LD motivated us to pursue an analytic investigation of the question. We followed the deterministic approach of Maynard Smith and Haigh (1974), extending their model to three loci as did Thomson (1977). To analyze this model, we used the framework of Barton and Turelli (1991), which provides a natural setting for the investigation of the directional hitchhiking, yielding transparent mathematical expressions that illuminate the rather surprising dynamics of LD under the hitchhiking effect.

THE THREE-LOCUS HITCHHIKING MODEL

We consider a three-locus model with two neutral loci and one selected locus. For each locus, we assume that there are only two allele types, denoted by 0 and 1. The selected locus may be between the two neutral ones or on either side of them (see Figure 1). We denote by L and R the left and right neutral loci, respectively, and by S the selected locus. The corresponding recombination fractions between loci are r_{LR}, etc. We assume positive directional selection according to the following fitness scheme:

Genotype at the selected locus	11	10	00
Relative fitness	$1+s$	$1+hs$	1

where s is the selection coefficient of the selected allele (type 1) and h the dominance coefficient. Note that we follow here the notation of Maynard Smith and Haigh (1974), not

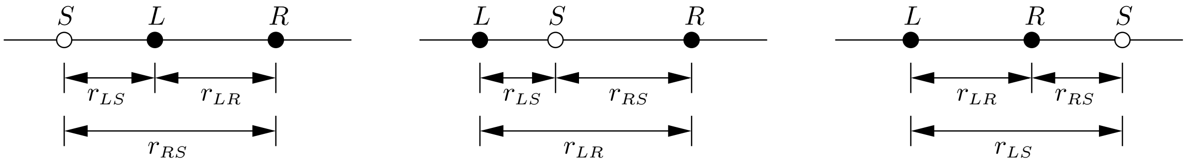


FIGURE 1: Three possible configurations. The selected locus is denoted by S, whereas the left and the right neutral loci are denoted by L and R, respectively. The recombination rate r_{LR} between L and R is given by adding or subtracting r_{LS} and r_{RS}, depending on the configuration.

the definition of KAPLAN *et al.* (1989). Effective population size N_e is assumed to be very large ($N_e s \gg 1$), such that a deterministic analysis of the model is appropriate.

The system of full recursions: To derive the recursions for the marginal (type 1) allele frequencies p_L, p_R, and p_S at the three loci and the LDs (central moments) of the second (C_{LR}, C_{LS}, C_{RS}) and third (C_{LRS}) orders measured with respect to type 1 alleles, we follow the approach of BARTON and TURELLI (1991). In that approach, coefficients which appear in the recursions are recombination rates and generalized selection coefficients, the latter denoted by $\tilde{a}_{X,Y}$ and $\tilde{a}_{X,\varnothing}$, where X and Y are non-empty subsets of loci. (Generalized selection coefficients are defined as coefficients which appear in expressing the relative fitness in terms of certain quantities related to LDs.) In our case, non-zero selection coefficients which appear in the recursions at time t are

$$\begin{aligned}\tilde{a}_{S,\varnothing}(t) &= \frac{s[h+(1-2h)p_S(t)]}{1+s\,p_S(t)[2h+(1-2h)p_S(t)]},\\ \tilde{a}_{S,S}(t) &= \frac{(1-2h)s}{1+s\,p_S(t)[2h+(1-2h)p_S(t)]}.\end{aligned}$$

For $h = 1/2$, these simplify to

$$\tilde{a}_{S,\varnothing}(t) = \frac{s}{2[1+s\,p_S(t)]} \quad \text{and} \quad \tilde{a}_{S,S}(t) \equiv 0.$$

Following BARTON and TURELLI (1991), we define $r_{LRS} = r_{LS,R} + r_{RS,L} + r_{LR,S}$ and use $r_{X,Y}$ to denote the rate of recombination events that partition the loci into two non-empty sets X and Y. In our work, we ignore double-crossover recombination events, so we define $r_{X,Y} = 0$ if the partition X, Y corresponds to a double-crossover event. For example, if the selected locus S is between the neutral loci, then $r_{LS,R} = r_{RS}$, $r_{RS,L} = r_{LS}$, and $r_{LR,S} = 0$. Further, we assume that recombination rates are additive. The recombination rate r_{LR} is therefore given by adding or subtracting r_{LS} and r_{RS}, depending on the configuration (see Figure 1).

To simplify notation, we hereafter omit writing the dependence on time t. We define $q_S = 1 - p_S$, $q_L = 1 - p_L$, and $q_R = 1 - p_R$. Marginal allele frequencies satisfy the following recursions:

$$\Delta p_S = \tilde{a}_{S,\varnothing}\, p_S q_S, \tag{1}$$

$$\Delta p_L = \tilde{a}_{S,\varnothing}\, C_{LS}, \tag{2}$$

$$\Delta p_R = \tilde{a}_{S,\varnothing}\, C_{RS}. \tag{3}$$

The LDs satisfy the recursions

$$\Delta C_{LS} = \tilde{\Delta} C_{LS} - \Delta p_L \Delta p_S = \tilde{\Delta} C_{LS} - \tilde{a}^2_{S,\varnothing} p_S q_S C_{LS}, \tag{4}$$

$$\Delta C_{RS} = \tilde{\Delta} C_{RS} - \Delta p_R \Delta p_S = \tilde{\Delta} C_{RS} - \tilde{a}^2_{S,\varnothing} p_S q_S C_{RS}, \tag{5}$$

$$\Delta C_{LR} = \tilde{\Delta} C_{LR} - \Delta p_L \Delta p_R = \tilde{\Delta} C_{LR} - \tilde{a}^2_{S,\varnothing} C_{LS} C_{RS}, \tag{6}$$

$$\begin{aligned} \Delta C_{LRS} &= \tilde{\Delta} C_{LRS} - \Delta p_S \tilde{\Delta} C_{LR} - \Delta p_L \tilde{\Delta} C_{RS} - \Delta p_R \tilde{\Delta} C_{LS} + 2 \Delta p_L \Delta p_R \Delta p_S \\ &= \tilde{\Delta} C_{LRS} - \tilde{a}_{S,\varnothing} \left[p_S q_S \tilde{\Delta} C_{LR} + C_{LS} \tilde{\Delta} C_{RS} + C_{RS} \tilde{\Delta} C_{LS} \right] + 2 \tilde{a}^3_{S,\varnothing} p_S q_S C_{LS} C_{RS}, \end{aligned} \tag{7}$$

where

$$\tilde{\Delta} C_{LS} = g(r_{LS}) C_{LS}, \tag{8}$$

$$\tilde{\Delta} C_{RS} = g(r_{RS}) C_{RS}, \tag{9}$$

$$\tilde{\Delta} C_{LR} = -r_{LR} C_{LR} + \tilde{a}_{S,\varnothing}(1 - r_{LR}) C_{LRS} + \tilde{a}_{S,S} r_{LR}\, C_{LS} C_{RS}, \tag{10}$$

$$\begin{aligned} \tilde{\Delta} C_{LRS} = {} & \left[-r_{LRS} + \tilde{a}_{S,\varnothing}(1 - r_{LRS})(1 - 2p_S) + \tilde{a}_{S,S} r_{LR,S} p_S q_S \right] C_{LRS} \\ & - \tilde{a}_{S,\varnothing}(r_{LS,R} + r_{RS,L}) p_S q_S C_{LR} \\ & - \left[\tilde{a}_{S,\varnothing}(2 - r_{LS,R} - r_{RS,L}) - \tilde{a}_{S,S}(1 - 2p_S)(r_{LS,R} + r_{RS,L}) \right] C_{LS} C_{RS}, \end{aligned} \tag{11}$$

with

$$g(r) := -r + \tilde{a}_{S,\varnothing}(1 - r)(1 - 2p_S) + \tilde{a}_{S,S}\, r\, p_S q_S.$$

The above general recursions apply to all three configurations shown in Figure 1. Depending on the particular configuration being considered, recombination rates $r_{X,Y}$ need to be defined appropriately.

The system of truncated recursions: We explore the behavior of the recursion equations (1)–(7) in the region $r, s \ll 1$ and $h = 1/2$ (see Maynard Smith and Haigh 1974).

Here, r may be any recombination parameter appearing in the equations (1)–(7). Keeping only the terms linear in s or r leads to the following set of recursions:

$$\begin{aligned}
\Delta p_S &\approx \frac{s}{2}\, p_S(1-p_S), && (12)\\
\Delta p_L &\approx \frac{s}{2}\, C_{LS}, && (13)\\
\Delta p_R &\approx \frac{s}{2}\, C_{RS}, && (14)\\
\Delta C_{LS} &\approx -\left[r_{LS} + \frac{s}{2}(2p_S - 1)\right] C_{LS}, && (15)\\
\Delta C_{RS} &\approx -\left[r_{RS} + \frac{s}{2}(2p_S - 1)\right] C_{RS}, && (16)\\
\Delta C_{LR} &\approx -r_{LR}C_{LR} + \frac{s}{2}C_{LRS}, && (17)\\
\Delta C_{LRS} &\approx -\left[r_{LRS} + \frac{s}{2}(2p_S - 1)\right] C_{LRS} - sC_{LS}C_{RS}. && (18)
\end{aligned}$$

These equations agree to first order in r and s with those of THOMSON (1977) [compare her eqs. (30iii), (30iv), and (31)]. Since the hitchhiking effect can be best observed when $r \ll s$ (such that simultaneously $N_e s \gg 1$ holds), we may approximate the truncated recursions by the following ordinary differential equations (ODEs; see MAYNARD SMITH and HAIGH 1974):

$$\begin{aligned}
\frac{dp_S}{dt} &= \frac{s}{2}\, p_S[1-p_S], && (19)\\
\frac{dp_L}{dp_S} &= \frac{C_{LS}}{p_S(1-p_S)}, && (20)\\
\frac{dp_R}{dp_S} &= \frac{C_{RS}}{p_S(1-p_S)}, && (21)\\
\frac{dC_{LS}}{dp_S} &= -\left(\frac{2r_{LS}}{s} + 2p_S - 1\right)\left[\frac{1}{p_S(1-p_S)}\right] C_{LS}, && (22)\\
\frac{dC_{RS}}{dp_S} &= -\left(\frac{2r_{RS}}{s} + 2p_S - 1\right)\left[\frac{1}{p_S(1-p_S)}\right] C_{RS}, && (23)\\
\frac{dC_{LR}}{dp_S} &= -\frac{2r_{LR}}{s}\left[\frac{1}{p_S(1-p_S)}\right] C_{LR} + \frac{1}{p_S(1-p_S)}C_{LRS}, && (24)\\
\frac{dC_{LRS}}{dp_S} &= -\left(\frac{2r_{LRS}}{s} + 2p_S - 1\right)\left[\frac{1}{p_S(1-p_S)}\right] C_{LRS} - \frac{2}{p_S(1-p_S)}C_{LS}C_{RS}. && (25)
\end{aligned}$$

Here we have introduced time t (in generations) into the equation for p_S and parameterized the other quantities by p_S (which is a monotonically increasing function of t).

Structure of equations: Several features of the dynamical system become apparent from these two sets of equations. Most importantly, selection acts on the alleles at the neutral loci indirectly and in a strictly hierarchical fashion: on the marginal neutral allele frequencies via the pairwise LDs C_{LS} and C_{RS}, on the LD between the neutral sites by the third moment, and on the latter by a fourth-order term (i.e. the product of the two pairwise LDs C_{LS} and C_{RS}).

Analytical solutions when the selected locus is between the neutral loci: The ODEs for the LDs are first-order linear differential equations. The ODEs for the pairwise LDs C_{LS} and C_{RS} are homogeneous. The equations for C_{LR} and C_{LRS} contain the higher-order moments as inhomogeneous terms that act as "driving forces" of the dynamics. However, except for this impact of the higher-order terms, the equations are decoupled and can be solved successively. We have the following results:

The frequency of the selected allele at locus S is

$$p_S(t) = \frac{p_S(0)}{p_S(0) + q_S(0)e^{-st/2}}, \tag{26}$$

whereas marginal allele frequencies at the neutral loci are

$$\begin{aligned} p_L(t) &= p_L(0) + 2\frac{C_{LS}(0)}{H_S(0)}\left(\frac{p_S(0)}{1-p_S(0)}\right)^{2r_{LS}/s} \int_{p_S(0)}^{p_S(t)} \left(\frac{1-z}{z}\right)^{2r_{LS}/s} \mathrm{d}z, & (27) \\ p_R(t) &= p_R(0) + 2\frac{C_{RS}(0)}{H_S(0)}\left(\frac{p_S(0)}{1-p_S(0)}\right)^{2r_{RS}/s} \int_{p_S(0)}^{p_S(t)} \left(\frac{1-z}{z}\right)^{2r_{RS}/s} \mathrm{d}z, & (28) \end{aligned}$$

where

$$H_S(t) := 2p_S(t)[1 - p_S(t)], \tag{29}$$

which corresponds to the heterozygosity at the selected locus. The LDs $C_{LS}(t)$ and $C_{RS}(t)$ can be written as

$$\begin{aligned} C_{LS}(t) &= C_{LS}(0)\,\frac{H_S(t)}{H_S(0)}\,e^{-r_{LS}t}, & (30) \\ C_{RS}(t) &= C_{RS}(0)\,\frac{H_S(t)}{H_S(0)}\,e^{-r_{RS}t}. & (31) \end{aligned}$$

Given these solutions for $C_{LS}(t)$ and $C_{RS}(t)$, the coupled ODEs (24) and (25) admit simple exact solutions when the selected locus is between the two neutral loci. More specifically, the 3rd order LD is given by

$$C_{LRS}(t) = \left\{-4\,C_{LS}(0)C_{RS}(0)\left[\frac{p_S(t) - p_S(0)}{H_S(0)}\right] + C_{LRS}(0)\right\}\frac{H_S(t)}{H_S(0)}e^{-r_{LRS}t}, \tag{32}$$

and the LD between the neutral loci can be written as

$$C_{LR}(t) = \left\{ -4\,C_{LS}(0)C_{RS}(0)\left[\frac{p_S(t)-p_S(0)}{H_S(0)}\right]^2 + 2\,C_{LRS}(0)\left[\frac{p_S(t)-p_S(0)}{H_S(0)}\right] + C_{LR}(0) \right\} e^{-r_{LR}t}, \tag{33}$$

where $r_{LR} = r_{LRS} = r_{LS} + r_{RS}$.

Analytical solutions when the selected locus is outside the neutral loci: Solutions to the ODEs (19)–(23) do not depend on whether the selected locus is inside or outside the two neutral loci. For example, the allele frequency $p_S(t)$ and the LDs $C_{LS}(t)$ and $C_{RS}(t)$ are given by (26), (30) and (31), respectively, in all cases. However, the dynamics of the ODEs (24) and (25) for $C_{LR}(t)$ and $C_{LRS}(t)$, respectively, depend crucially on the position of the selected locus S with respect to the neutral loci L and R. As we elaborate presently, the dynamics of $C_{LR}(t)$ when S is between L and R exhibits radically different behavior than when S is outside.

In what follows, suppose that S is to the right of R, which implies $r_{LS} = r_{LRS} = r_{LR} + r_{RS}$. The case where S is to the left of L can be handled in a similar vein, with r_{RS} replaced with r_{LS}. Now, the ODE (25) for the 3rd order LD $C_{LRS}(t)$ does not admit a closed-form solution; a general solution can be obtained in terms of the incomplete beta function $B(z;x,y)$, defined as $B(z;x,y) = \int_0^z u^{x-1}(1-u)^{y-1}\mathrm{d}u$. However, noting that $B(z;1-2r_{RS}/s, 1+2r_{RS}/s) \approx z$ if $r_{RS} \ll s$, we obtain the following simple approximate solution:

$$C_{LRS}(t) \approx \left\{ -4\,C_{LS}(0)C_{RS}(0)\left[\left(\frac{p_S(0)}{1-p_S(0)}\right)^{2r_{RS}/s}\frac{p_S(t)-p_S(0)}{H_S(0)}\right] + C_{LRS}(0) \right\} \frac{H_S(t)}{H_S(0)} e^{-r_{LRS}t}. \tag{34}$$

Further, using this solution and the approximation $B[z;2-2r_{RS}/s, 1+2r_{RS}/s] \approx z^2$ for $r_{RS} \ll s$, we obtain the following approximate solution to (24):

$$\begin{aligned} C_{LR}(t) \ \approx\ & \left\{ -4\,C_{LS}(0)C_{RS}(0)\left[\left(\frac{p_S(0)}{1-p_S(0)}\right)^{2r_{RS}/s}\frac{p_S(t)-p_S(0)}{H_S(0)}\right]^2 \right. \\ & \left. +2\,C_{LRS}(0)\left[\left(\frac{p_S(0)}{1-p_S(0)}\right)^{2r_{RS}/s}\frac{p_S(t)-p_S(0)}{H_S(0)}\right] + C_{LR}(0) \right\} e^{-r_{LR}t}. \end{aligned} \tag{35}$$

Note the striking resemblance of (34) and (35) to (32) and (33), respectively. For $r_{RS} = 0$, the two sets of equations agree exactly, and hence our solutions for different regions form one continuous solution for the entire domain. The only difference between the two sets of equations is that, in (34) and (35), an extra factor $[p_S(0)/(1-p_S(0))]^{2r_{RS}/s}$ appears together with $[p_S(t)-p_S(0)]/H_S(0)$. This simple difference leads to important observable differences in the dynamics of the LDs.

TABLE 1: Comparison of the exact $C_{LR}(t)$ (obtained by solving the full recursions numerically) with the analytic approximation (33) when the selected locus is between the two neutral loci. The values shown are in units of 10^{-2}. We used $s = 0.01, p_S(0) = 0.00005, p_L(0) = 0.38, p_R(0) = 0.41$ and $C_{LR}(0) = 0.0242$. In the exact numerical computation, $p_S(t) = 1 - p_S(0)$ at $t = 3982$.

	$r_{LS} = r_{RS} = 0.01s$		$r_{LS} = r_{RS} = 0.05s$		$r_{LS} = r_{RS} = 0.1s$	
t	exact	eq.(33)	exact	eq.(33)	exact	eq.(33)
0	2.42	2.42	2.42	2.42	2.42	2.42
500	2.21	2.21	1.48	1.48	0.90	0.90
1000	2.18	2.18	0.98	0.98	0.36	0.36
1500	3.67	3.70	1.11	1.12	0.25	0.25
2000	6.91	6.89	1.39	1.39	0.19	0.19
2500	1.59	1.53	0.21	0.21	0.02	0.02
3000	0.14	0.13	0.01	0.01	0.00	0.00
3500	0.01	0.01	0.00	0.00	0.00	0.00
3982	0.00	0.00	0.00	0.00	0.00	0.00

Comparison of approximate analytic solutions with numerical solutions to the full recursions: We have written a computer program to solve the full recursions (1)–(11) numerically. Comparison of our analytic solutions (33) and (35) with numerical solutions to the full recursions are shown in Table 1 and Table 2, respectively. As these tables show, our analytic solutions are a good approximation to the exact dynamics. In obtaining our analytic solution (35) for the case in which the selected locus is outside the neutral loci, recall that we assumed $r_{RS} \ll s$ to approximate the incomplete beta function. As expected, Table 2 shows that (35) becomes less accurate as r_{RS} increases, but it is a good approximation as long as $r_{RS} \ll s$.

THE DYNAMICS OF LD

In this section, we consider the dynamics of the LD between the two neutral loci. We utilize our analytic solutions from the previous section to study several important aspects of the dynamics.

TABLE 2: Comparison of the exact $C_{LR}(t)$ with the analytic approximation (35) when the selected locus is to the right of locus R. The values shown are in units of 10^{-2}. We used $r_{LR} = 0.02s$ and the same set of initial conditions as in Table 1. As expected, (35) is quite accurate for $r_{RS}/s \ll 1$.

	$r_{RS} = 0.01s$		$r_{RS} = 0.05s$		$r_{RS} = 0.1s$		$r_{RS} = 0.2s$		$r_{RS} = 0.3s$	
t	exact	eq.(35)	exact	eq.(35)	exact	eq.(35)	exact	eq.(35)	exact	eq.(35)
0	2.42	2.42	2.42	2.42	2.42	2.42	2.42	2.42	2.42	2.42
500	2.21	2.20	2.20	2.20	2.20	2.19	2.20	2.19	2.20	2.19
1000	2.17	2.15	2.12	2.06	2.07	2.01	2.02	1.99	2.00	1.98
1500	3.47	3.39	2.83	2.55	2.36	2.08	1.96	1.83	1.85	1.80
2000	6.94	6.93	5.48	5.15	3.59	3.15	2.05	1.85	1.72	1.65
2500	4.33	4.35	6.04	5.98	3.96	3.76	1.96	1.83	1.57	1.52
3000	3.26	3.27	5.54	5.51	3.64	3.52	1.78	1.68	1.42	1.38
3500	2.90	2.90	5.01	5.00	3.30	3.20	1.61	1.52	1.29	1.25
3982	2.63	2.63	4.55	4.54	2.99	2.90	1.46	1.38	1.17	1.13

Vanishing LD: In the domain $r \ll s$, the term $-sC_{LS}C_{RS}$ dominates the recursion for the third moment [see (18)] and hence influences also the LD between the neutral sites. If the selected site is between the two neutral sites and linkage is sufficiently tight ($r_{LR} < 0.1s$), $|C_{LR}|$ quickly increases after the favored mutation has entered the population and, after transiently reaching a peak, rapidly decays to zero before the selected phase ends. To show this, define t_f as the time satisfying $p_S(t_f) = 1 - p_S(0)$. Henceforward, we loosely refer to this time as the *fixation time*. Using (26), one can show that

$$t_f = \frac{4}{s} \log\left(\frac{1 - p_S(0)}{p_S(0)}\right). \tag{36}$$

We wish to show that, if the selected locus is between the two neutral loci, then $C_{LR}(t_f) \approx 0$ for all possible initial conditions of interest. Common to all initial conditions is that $p_S(0) = 1/(2N)$, with N being the population size. For $N = 10^4 \sim 10^6$, $1/(2N) \ll 1$, and therefore

$$\frac{p_S(t_f) - p_S(0)}{H_S(0)} = \frac{1 - 2p_S(0)}{2p_S(0)[1 - p_S(0)]} \approx \frac{1}{2p_S(0)}, \tag{37}$$

which implies that (33) at t_f can be written approximately as follows:

$$C_{LR}(t_f) \approx \left\{ -C_{LS}(0)C_{RS}(0) \left[\frac{1}{p_S(0)}\right]^2 + C_{LRS}(0)\frac{1}{p_S(0)} + C_{LR}(0) \right\} e^{-r_{LR}t_f}. \tag{38}$$

We use $\{000, 001, 010, 011, 100, 101, 110, 111\}$ to denote gametic types. Their frequencies are denoted by $\{f_{000}, f_{001}, f_{010}, f_{011}, f_{100}, f_{101}, f_{110}, f_{111}\}$, which are related to the marginal frequencies and the LDs as follows:

$$\begin{aligned}
p_{_L} &= f_{100} + f_{101} + f_{110} + f_{111}, \\
p_{_R} &= f_{001} + f_{011} + f_{101} + f_{111}, \\
p_{_S} &= f_{010} + f_{011} + f_{110} + f_{111}, \\
C_{_{LR}} &= f_{101} + f_{111} - p_{_L}p_{_R}, \\
C_{_{LS}} &= f_{110} + f_{111} - p_{_L}p_{_S}, \\
C_{_{RS}} &= f_{011} + f_{111} - p_{_R}p_{_S}, \\
C_{_{LRS}} &= f_{111} - (p_{_L}p_{_R}p_{_S} + p_{_L}C_{_{RS}} + p_{_R}C_{_{LS}} + p_{_S}C_{_{LR}}).
\end{aligned}$$

Using these relations, we obtain

$$-C_{_{LS}}(0)C_{_{RS}}(0)\left[\frac{1}{p_{_S}(0)}\right]^2 + C_{_{LRS}}(0)\frac{1}{p_{_S}(0)} + C_{_{LR}}(0) = \frac{f_{010}(0)f_{111}(0) - f_{011}(0)f_{110}(0)}{p_{_S}(0)^2}. \quad (39)$$

At $t = 0$, a new favored mutation occurs on only one of the following gametic types: 000, 001, 100 or 101. For instance, if the mutation occurs on a gamete of type 101, $f_{010}(0) = f_{011}(0) = f_{110}(0) = 0$ and $f_{111}(0) = p_{_S}(0) = 1/(2N)$. More generally, only one of $f_{010}(0)$, $f_{111}(0), f_{011}(0), f_{110}(0)$ is supposed to be non-zero. This implies that the right hand side of (39) must be equal to zero. Hence, the coefficient of $\exp(-r_{_{LR}}t_f)$ in (38) is exactly zero. If the approximation (37) were not used, then the coefficient of $\exp(-r_{_{LR}}t_f)$ in $C_{_{LR}}(t_f)$ would not be exactly zero, but would still be very small. Note that $\exp(-r_{_{LR}}t_f)$ may not be very small. For example, $\exp(-r_{_{LR}}t_f) = 0.452813$ for $p_{_S}(0) = 0.00005$, $r_{_{LR}} = 0.00002$, and $s = 0.001$. This shows that, for small recombination rates ($r_{_{LR}} \ll 1$), selection rather than recombination is the dominant force that causes $C_{_{LR}}(t)$ to vanish before the fixation time. For large recombination rates, the contribution $\exp(-r_{_{LR}}t)$ from recombination should dominate over selection effects.

This behavior may be explained as follows. Under the scenario of tight linkage and strong selection, a low-frequency gamete on which the favored mutation landed is quickly dragged into intermediate to high frequency. If the recombination rates are non-zero, this gamete may undergo recombination, thereby creating the two types of single recombinants that also carry the selected allele and thus increase in frequency. This reduces the LD between L and R created by the hitchhiking effect in the first half of the selected phase. This hitchhiking effect on the recombinants is stronger, the greater the linkage between the selected site and the two neutral sites is, and thus also the product $C_{LS}C_{RS}$.

Figure 2a shows another important observation: LD may vanish very quickly in the selected phase, while relative heterozygosity approaches a finite (i.e. non-zero) equilibrium value. Thus, LD does not vanish because of the variation-reducing effect of hitchhiking *per se*, but as a consequence of secondary hitchhiking effects on the recombinants created in the selected phase (described above).

A selected mutation occurring outside the two neutral sites on a low-frequency gamete may also lead to a transient peak of C_{LR}, if both neutral polymorphic sites are less than $0.1s$ recombination distances away from the selected site (see Figure 2b,c). Although this peak vanishes faster than under neutral conditions (i.e., with the selected site far away from the neutral sites, as in Figure 2d), the decay rate is not as high as when the favored mutation occurs between the neutral sites (Figure 2a). We analyze this behavior in more detail below.

Shown in Figure 3 is a plot of $C_{LR}(n)$ for varying position of the selected locus. As in Figure 2, the distance between the two neutral loci is fixed at $r_{LR} = 0.0002$, and the same set of initial conditions are used. Note that this plot is symmetric about the plane $r = 0$; we return to this point later in the paper. We stress that our conclusions described above do not depend on the particular values of neutral marginal allele frequencies $p_L(0)$ and $p_R(0)$ used for illustration. Even for low values of $p_L(0)$ and $p_R(0)$, for example, the same conclusions hold.

An alternative illustration of the above discussion is provided in Figure 4, which shows pairwise LD plots for a region containing 100 neutral loci and a single selected locus. The two plots shown correspond to two different time points. The selected locus is located in the middle of the region and LD values below a cutoff value are not plotted.

The maximum LD at t_f: Consider the case in which the selected locus S is to the right of locus R. Viewing $C_{LR}(t_f)$ as a function of r_{RS}, whether there exists a local optimum in the domain $r_{RS} \ll s$ depends on initial conditions. The example shown in Figure 3 has a local maximum at $r_{RS}/s \approx 0.039$. Differentiating the analytic solution (35), it is possible to determine whether there is a critical point $r_{RS} = r^*_{RS}$ that satisfies

$$\left.\frac{dC_{LR}(t_f)}{dr_{RS}}\right|_{r_{RS}=r^*_{RS}} = 0. \tag{40}$$

Suppose that, at $t = 0$, the new gamete carrying a selected allele (of type 1) at locus S is of type $ij1$, with i being the allele at locus L and j the allele at locus R. Then, given that

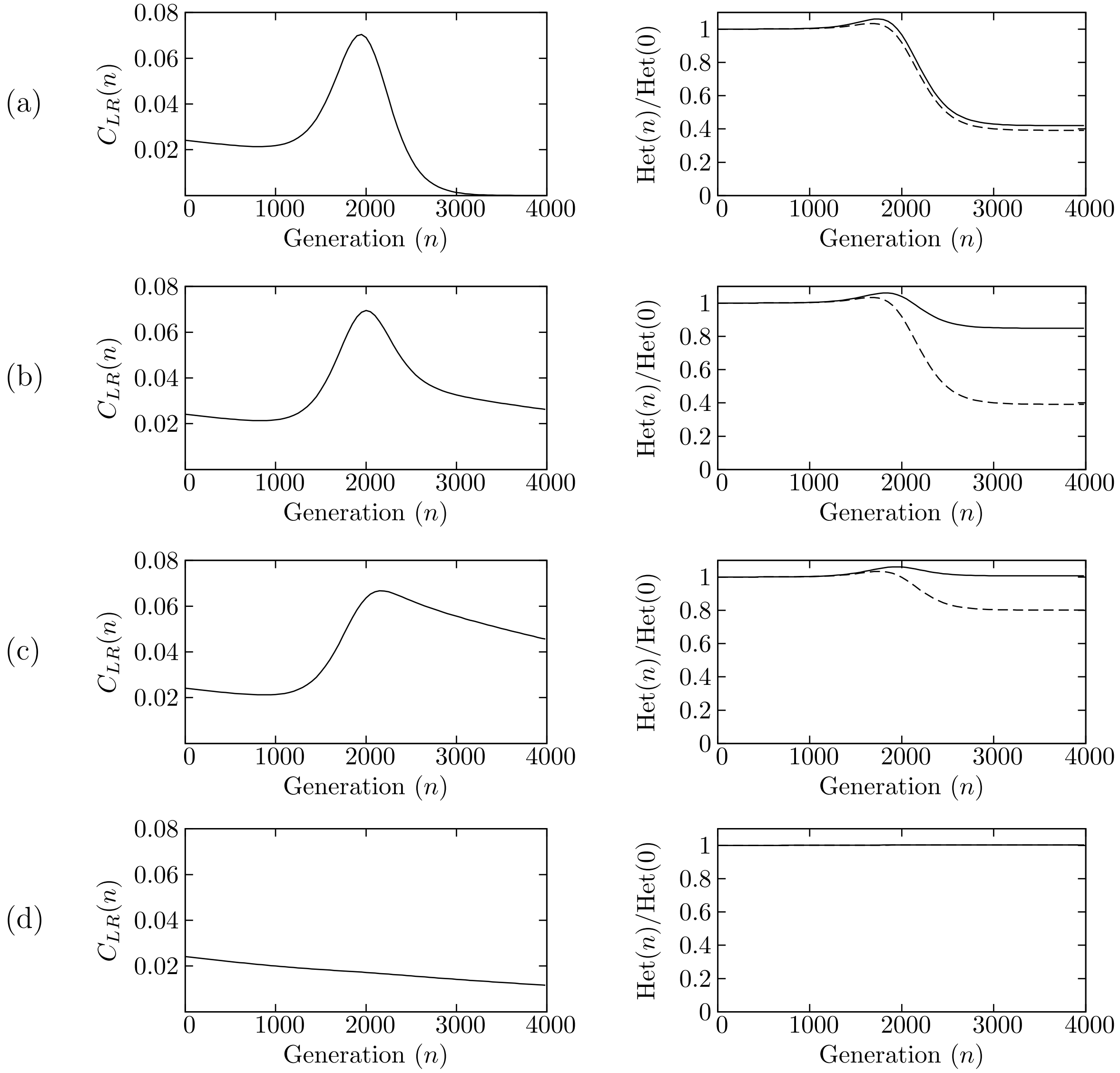


FIGURE 2: Example trajectories of $C_{LR}(n)$ are shown on the left hand side, and that of normalized heterozygosity Het(n)/Het(0) are shown on the right hand side, where solid (resp. dashed) lines correspond to locus L (resp. R). (a) The selected locus is at the midpoint between the neutral loci, i.e., $r_{LS} = r_{RS} = 0.01s$. (b) The selected locus is to the right of locus R and $r_{RS} = 0.01s$. (c) The selected locus is to the right of locus R and $r_{RS} = 0.03s$. (d) The selected locus is to the right of locus R and $r_{RS} = 0.3s$. Exact recursions (1)–(11) were used, with $s = 0.01, r_{LR} = 0.02s, p_S(0) = 0.00005, p_L(0) = 0.38, p_R(0) = 0.41$ and $C_{LR}(0) = 0.0242$. At generation $n = 3982$, $p_S(n) = 1 - p_S(0)$. Initial conditions were chosen so that $C_{LR}(0) \neq 0$, for two reasons; to demonstrate that, when the selected locus is between the neutral loci (as in case a), $C_{LR}(n)$ quickly vanishes in the selected phase even if the initial value $C_{LR}(0)$ is not zero, and to contrast the effect of recombination (see case d) with that of selection.

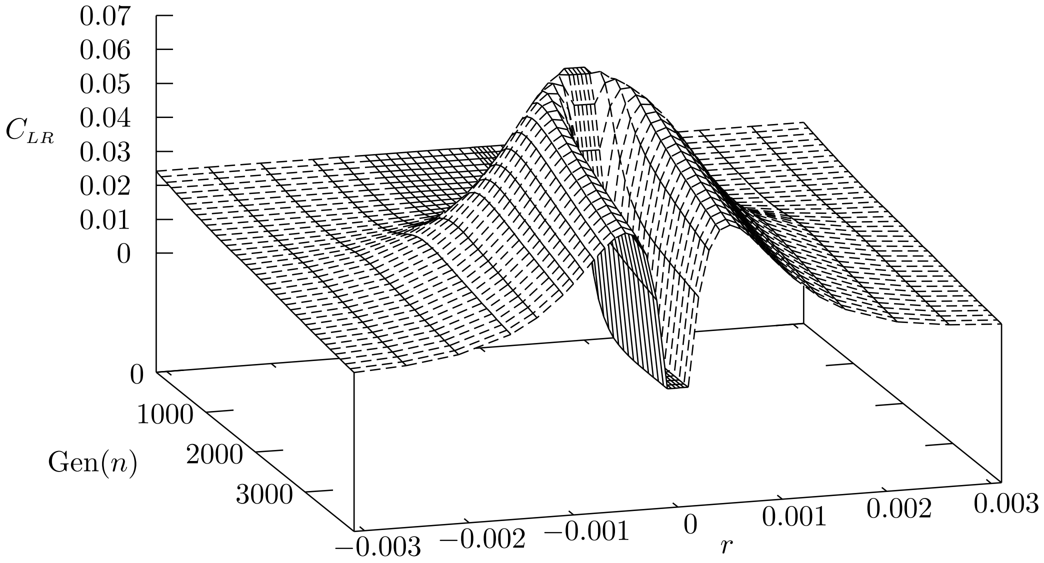


FIGURE 3: The LD C_{LR} between the neutral loci as a function of the position of the selected locus, for $r_{LR} = 0.0002$ and $s = 0.01$. Here, r is the position of the selected locus S, and the value $r = 0$ corresponds to the midpoint between the neutral loci. Locus L is fixed at $r = -0.0001$, whereas locus R is fixed at $r = 0.0001$. The same set of initial conditions as in Figure 2 was used for all r. Exact recursions (1)–(11) were used to generate this plot.

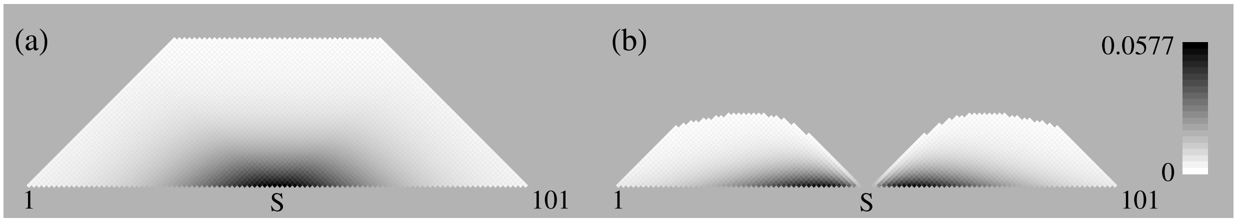


FIGURE 4: Truncated pairwise LD plots for a region consisting of 100 neutral loci and 1 selected locus located in the middle. The recombination rate between any two adjacent loci is $0.4s/100$, and hence the recombination rate between the leftmost and the rightmost neutral loci is $0.4s$. We used $s = 0.01, p_S(0) = 0.00005, p_R(0) = p_L(0) = 0.5$ and $C_{LR}(0) = 0$. The plots were truncated so that LD values less than 1% of the largest value 0.0577 were not included. (a) Pairwise LD plot at $t = t_f/2$, where the time of fixation t_f satisfies $p_S(t_f) = 1 - p_S(0)$. (b) Pairwise LD plot at $t = t_f$.

$[\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)] > C_{\scriptscriptstyle LR}(0)$, where δ_{ab} is 1 if $a = b$ or 0 if $a \neq b$, we obtain

$$r^*_{\scriptscriptstyle RS} \approx \frac{s}{2} \log \left[\frac{1}{2} \left(1 - \frac{C_{\scriptscriptstyle LR}(0)}{[\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)]} \right) \right] \frac{1}{\log[1/(2N)]} \tag{41}$$

and

$$C_{\scriptscriptstyle LR}(t_f)|_{r_{RS} = r^*_{RS}} \approx \frac{\{C_{\scriptscriptstyle LR}(0) + [\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)]\}^2}{4[\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)]} e^{-r_{LR} t_f}.$$

The value of $r^*_{\scriptscriptstyle RS}$ in (41) may be very large for some initial conditions. In such a case, as our analytic solution (35) is valid only in the domain $r_{\scriptscriptstyle RS} \ll s$, all we can say for sure is that $C_{\scriptscriptstyle LR}(t_f)$ has no critical point in the domain $r_{\scriptscriptstyle RS} \ll s$ (i.e., $C_{\scriptscriptstyle LR}(t_f)$ is either a monotonically increasing or a monotonically decreasing function of $r_{\scriptscriptstyle RS}$ in that domain). Further, if $[\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)] \leq C_{\scriptscriptstyle LR}(0)$, there is no real-valued $r^*_{\scriptscriptstyle RS}$ such that our approximate analytic solution (35) satisfies (40). Hence, noting that $C_{\scriptscriptstyle LR}(t_f)$ approaches $C_{\scriptscriptstyle LR}(0)e^{-r_{LR}t_f}$ as $r_{\scriptscriptstyle RS}/s$ increases, we conclude that, in the domain $r_{\scriptscriptstyle RS} \ll s$,

$$\max |C_{\scriptscriptstyle LR}(t_f)| \approx \begin{cases} \max(|X|, |C_{\scriptscriptstyle LR}(0)|) \times e^{-r_{LR}t_f}, & \text{if } [\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)] > C_{\scriptscriptstyle LR}(0), \\ |C_{\scriptscriptstyle LR}(0)| \times e^{-r_{LR}t_f}, & \text{otherwise,} \end{cases}$$

where

$$X = \frac{\{C_{\scriptscriptstyle LR}(0) + [\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)]\}^2}{4[\delta_{i1} - p_{\scriptscriptstyle L}(0)][\delta_{j1} - p_{\scriptscriptstyle R}(0)]} \tag{42}$$

and $e^{-r_{LR}t_f} \approx \left(\frac{1}{2N}\right)^{4r_{LR}/s}$. Note that the maximum possible value of X is $1/4$. The critical point $r_{\scriptscriptstyle LS} = r^*_{\scriptscriptstyle LS}$ for the case in which the selected locus is to the left of locus L, is also given by (41).

Invariance of $C_{\scriptscriptstyle LR}$ and $C_{\scriptscriptstyle LRS}$ when the selected locus is between the two neutral loci: When the favored mutation occurs between the two neutral sites, the dynamics of the system of truncated recursions for $C_{\scriptscriptstyle LR}$ and $C_{\scriptscriptstyle LRS}$ does not depend on the position of the selected locus. This can immediately be seen from equations (17) and (18), which depend only on the sum $r_{\scriptscriptstyle LS} + r_{\scriptscriptstyle RS}$ and not on any individual recombination parameter. We show in Appendix that this invariance also (nearly) holds for the system of full recursions.

INITIAL CONDITIONS AND THE PARAMETER SPACE OF THE MODEL

In this section, we assume that the selected locus is to the right of locus R. For such a case, recall that the LD (measured with respect to type 1 alleles) between the neutral loci is given by (35). We use ijk to denote gametic types, with i being for locus L, j for locus R, and k for locus S. By *the gamete of origin*, we mean the new gamete at $t = 0$ carrying a selected allele (of type 1) at locus S. We include ij in superscript (i.e., we write $C^{ij}_{LR}(t)$) if the gamete of origin is of type $ij1$. Using (35), we obtain

$$C^{ij}_{LR}(t) = [1-\alpha(t, r_{RS}/s)]\left\{C_{LR}(0) + \alpha(t, r_{RS}/s)[\delta_{i1} - p_L(0)][\delta_{j1} - p_R(0)]\right\} e^{-r_{LR}t}, \qquad (43)$$

where, as before, δ_{ab} is 1 if $a = b$ or 0 if $a \neq b$, and $\alpha(t, y)$ is defined as

$$\alpha(t, y) := \left[\frac{p_S(0)}{1-p_S(0)}\right]^{2y} \frac{p_S(t) - p_S(0)}{1-p_S(0)}. \qquad (44)$$

Recall that marginal (type 1) allele frequencies $p_L(t)$ and $p_R(t)$ at the neutral loci are given by (27) and (28), respectively. For $p_S(0) \ll 1$ and $r/s \ll 1$, we can use the approximation

$$\int_{p_S(0)}^{1-p_S(0)} \left(\frac{1-z}{z}\right)^{2r/s} \mathrm{d}z \approx 1 - 2p_S(0),$$

from which it follows that

$$\begin{aligned} p_L(t_f) &= p_L(0) + \frac{C_{LS}(0)}{p_S(0)}\alpha(t_f, r_{LS}/s), \\ p_R(t_f) &= p_R(0) + \frac{C_{RS}(0)}{p_S(0)}\alpha(t_f, r_{RS}/s), \end{aligned}$$

where $\alpha(t_f, r_{RS}/s)$ is defined as in (44). Similar to $C^{ij}_{LR}(t)$, we use $p^{ij}_L(t)$ and $p^{ij}_R(t)$ to denote $p_L(t)$ and $p_R(t)$, respectively, if the gamete of origin is of type $ij1$. It is straightforward to show that

$$p^{ij}_L(t_f) = p_L(0) + [\delta_{i1} - p_L(0)]\alpha(t_f, r_{LS}/s), \qquad (45)$$

$$p^{ij}_R(t_f) = p_R(0) + [\delta_{j1} - p_R(0)]\alpha(t_f, r_{RS}/s). \qquad (46)$$

Frequency averaged LD and the range of the hitchhiking effect: In what follows, we use x_{ijk} to denote the frequency of the gametic type ijk at time $t = 0$ and define $x_{ij\cdot} = x_{ij0} + x_{ij1}$. The type of the gamete of origin could be any of $001, 011, 101$ and 111. Suppose that the probability of the gamete of origin being of type $ij1$ is equal to the frequency of the gametic type $ij0$ just before time $t = 0$ (note that this frequency is equal to $x_{ij\cdot}$). Then, the average value of $C_{LR}(t)$ with respect to this probability is given by $\sum_{i,j} x_{ij\cdot} C^{ij}_{LR}(t)$, which we call a *frequency averaged* LD. We show below that, contrary to people's common intuition, the effect of selection on such an averaged LD does not depend on haplotype diversity $x_{ij\cdot}$ at $t = 0$.

Using (43), we can show that $\sum_{i,j} x_{ij\cdot} C^{ij}_{LR}(t)$ is given by

$$\sum_{i,j} x_{ij\cdot} C^{ij}_{LR}(t) = C_{LR}(0)\left\{1 - [\alpha(t, r_{RS}/s)]^2\right\} e^{-r_{LR}t}.$$

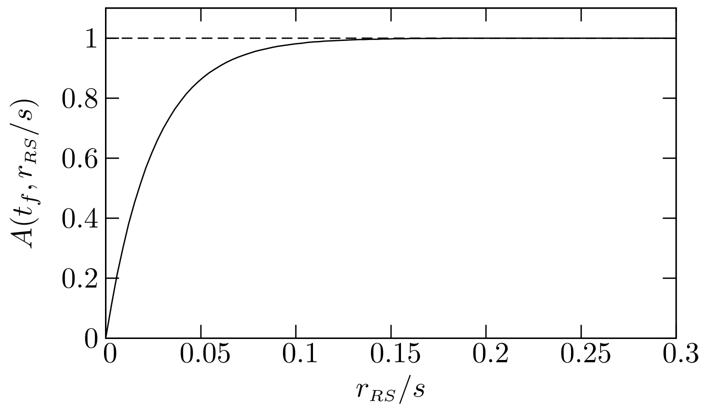


FIGURE 5: A plot of $A(t_f, r_{RS}/s)$ for $p_S(0) = 0.00005$. The function $A(t_f, r_{RS}/s)$, defined in (47), captures the effect of selection on both relative frequency averaged LD and relative frequency averaged heterozygosity (see (47) and (48)).

If $C_{LR}(0) = 0$, then $\sum_{i,j} x_{ij}.C^{ij}_{LR}(t) = 0$ for all t. For $C_{LR}(0) \neq 0$, we define

$$A(t, r_{RS}/s) := \frac{\sum_{i,j} x_{ij}.C^{ij}_{LR}(t)}{C_{LR}(0)e^{-r_{LR}t}} = 1 - [\alpha(t, r_{RS}/s)]^2 .$$

Note that r_{LR} need not be much smaller than s for our analytic solution (35) to be valid (recall that only $r_{RS} \ll s$ is required). Assuming $r_{LR} \gg \frac{1}{N}$, we can ignore genetic drift and regard $C_{LR}(0)e^{-r_{LR}t}$ as the behavior of LD under neutrality. Thus, $A(t, r_{RS}/s)$ can be viewed as the ratio of the frequency averaged LD in the presence of selection to that in the absence of selection. At the time t_f of fixation, $p_S(t_f) = 1 - p_S(0)$ and therefore

$$A(t_f, r_{RS}/s) = \frac{\sum_{i,j} x_{ij}.C^{ij}_{LR}(t_f)}{C_{LR}(0)e^{-r_{LR}t_f}} = 1 - \left[\frac{p_S(0)}{1 - p_S(0)}\right]^{4r_{RS}/s} \left[\frac{1 - 2p_S(0)}{1 - p_S(0)}\right]^2 . \qquad (47)$$

For given r_{RS}/s, $A(t_f, r_{RS}/s)$ only depends on $p_S(0) = 1/(2N)$; it has no dependence on other initial conditions. A plot of $A(t_f, r_{RS}/s)$ is shown in Figure 5 for $p_S(0) = 0.00005$.

We now compare $A(t_f, r_{RS}/s)$ with relative frequency averaged heterozygosity. Let us focus on the right neutral locus R and define $H^{ij}_R(t) = 2p^{ij}_R(t)[1 - p^{ij}_R(t)]$. For $H_R(0) = 2p_R(0)[1 - p_R(0)] \neq 0$, one can use (46) to show that

$$\frac{\sum_{i,j} x_{ij}.H^{ij}_R(t_f)}{H_R(0)} = 1 - [\alpha(t_f, r_{RS}/s)]^2 = 1 - \left[\frac{p_S(0)}{1 - p_S(0)}\right]^{4r_{RS}/s} \left[\frac{1 - 2p_S(0)}{1 - p_S(0)}\right]^2 . \qquad (48)$$

For $p_S(0) = 1/(2N)$, this is approximately equal to $1 - 1/(2N)^{4r_{RS}/s}$, which is equivalent to eq. (14d) of STEPHAN *et al.* (1992) (the factor 2 in the exponent of that formula needs to be replaced by 4 because of the different definition of the selection coefficient). In the absence

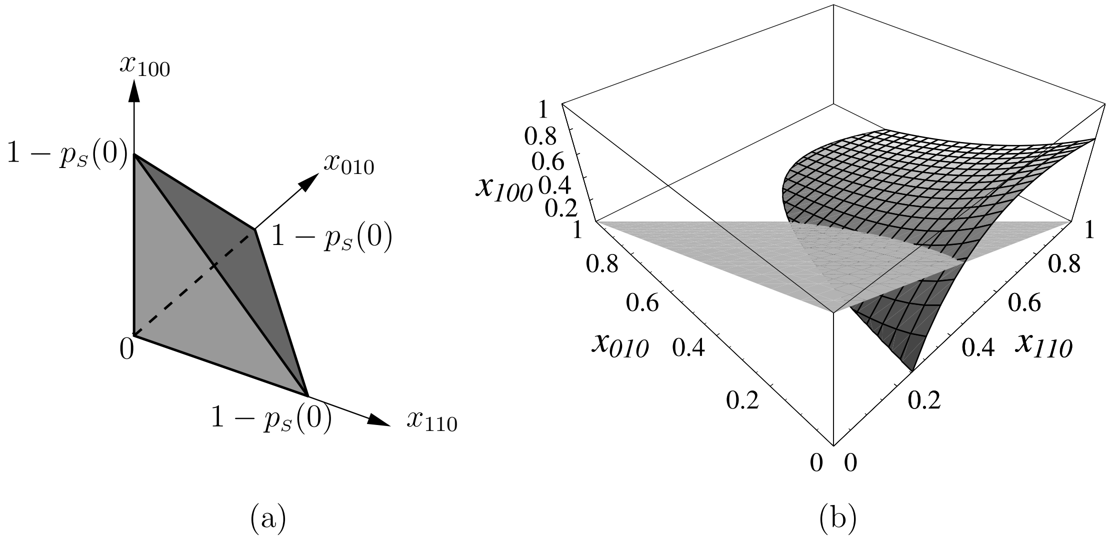


FIGURE 6: Illustration of initial conditions leading to the same $C_{LR}^{00}(t_f)$. (a) The illustrated tetrahedron Δ corresponds to the set of all possible initial conditions, given that the gamete of origin is of type 001. Δ is defined by (49), and a point in Δ corresponds to a particular initial condition. (b) The intersection of Δ and surface Ξ, defined by (50), corresponds to the set of all initial conditions leading to a fixed value c of $C_{LR}^{00}(t_f)$. For visual clarity, only one face F of Δ is shown. The intersection of Δ and Ξ is the part of Ξ below F.

of genetic drift, $H_R^{ij}(t_f) = H_R(0)$ for $s = 0$. Therefore, (48) can be regarded as the ratio of the frequency averaged heterozygosity in the presence of selection to that in the absence of selection. Surprisingly, this ratio is exactly equal to the analogous ratio for LD shown in (47). The function $A(t_f, r_{RS}/s)$ plays a special role in the sense that it encodes the effect of selection on two different frequency averaged quantities.

For site heterozygosity, the hitchhiking effect is generally only profound when, provided that $N_e s \gg 1$, the recombination distance r between the selected and neutral sites satisfies $r < 0.1s$ (MAYNARD SMITH and HAIGH 1974). The term determining this effect is $(2N)^{-4r/s}$, assuming that the initial frequency $p_S(0)$ of the selected allele is $1/(2N)$. Our above analysis shows that the range of a substantial reduction of LD due to hitchhiking (determined by $(2N)^{-4r_{RS}/s}$) is exactly equal to that for variation (determined by $(2N)^{-4r/s}$). (see Figure 5.)

Characterization of equivalent initial conditions:

We now find the set of all initial conditions that lead to the same value of C_{LR} at the time of fixation, i.e., $C_{LR}(t_f) = c$, where c is some fixed constant.

To be concrete, suppose that the gamete of origin is of type 001, in which case $x_{101} = x_{011} = x_{111} = 0$ and $x_{001} = p_S(0) = 1/(2N)$. First, note that $x_{000}+x_{010}+x_{100}+x_{110}+p_S(0) = 1$

implies

$$0 \leq x_{010} + x_{100} + x_{110} \leq 1 - p_S(0), \tag{49}$$

which defines a tetrahedron Δ as depicted in Figure 6a. Second, using (43), one can show that $C^{00}_{LR}(t_f) = c$ implies

$$c = [1-\alpha(t_f, r_{RS}/s)] \times \{x_{110} - [1-\alpha(t_f, r_{RS}/s)](x_{100} + x_{110})(x_{010} + x_{110})\} \times \left[\frac{p_S(0)}{1-p_S(0)}\right]^{4r_{LR}/s}, \tag{50}$$

which defines a surface Ξ in a 3-dimensional Euclidean space with $(x_{110}, x_{010}, x_{100})$ as coordinates. The intersection of surface Ξ with tetrahedron Δ, illustrated in Figure 6b, corresponds to the set of initial conditions such that $C^{00}_{LR}(t_f) = c$. A case in which the gamete of origin is of type other than 001 can be handled in a similar vein.

Probability distributions of $C_{LR}(t_f)$: Recall that $C_{LR}(t)$ for $t > 0$ depends on initial conditions. In what follows, we regard initial gametic frequencies as being random and consider the probability distribution of $C_{LR}(t_f)$. The squared correlation coefficient $R^2(t_f)$ is addressed later in DISCUSSION. We assume that all initial gametic frequency configurations $x_{000}, x_{010}, x_{100}, x_{110}$ are equally likely and satisfy $x_{000} + x_{010} + x_{100} + x_{110} + p_S(0) = 1$. Under this assumption of uniform distribution, it is possible to compute the probability distribution $\mathbb{P}[C_{LR}(t_f) < c]$ for fixed r_{LR}/s and r_{RS}/s. The key idea is to utilize the characterization of equivalent initial conditions described above. More precisely, as c changes, the surface defined by $C_{LR}(t_f) = c$ changes in a smooth fashion, sweeping out a region in three dimensions. The probability $\mathbb{P}[C_{LR}(t_f) < c]$ is equal to the volume of the region corresponding to $C_{LR}(t_f) < c$ inside Δ, normalized by the total volume of Δ.

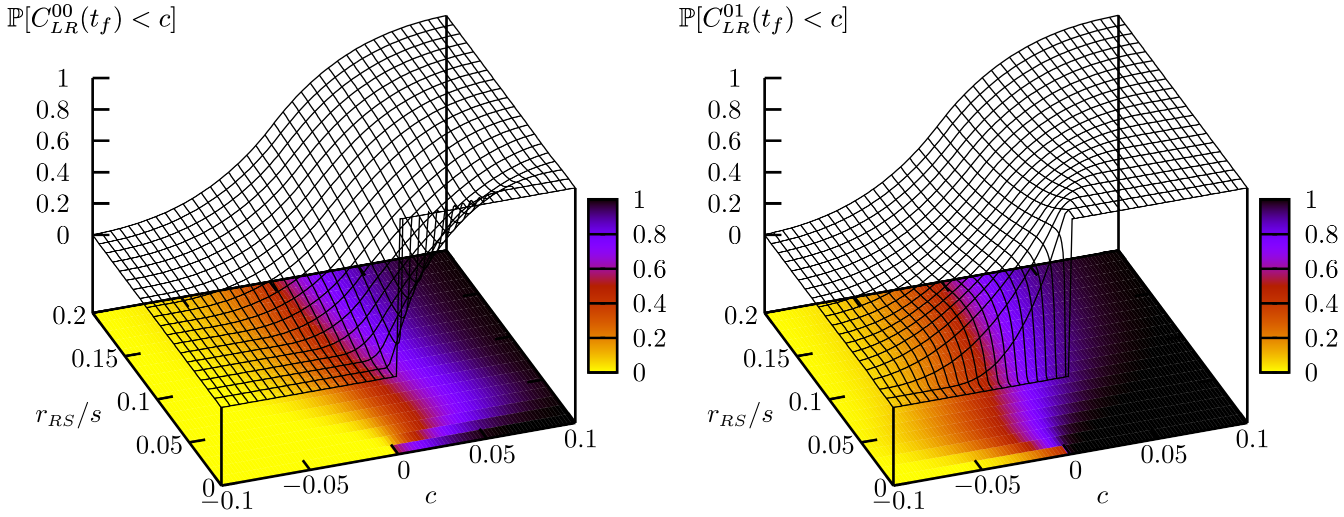
Our main result, illustrated in Figure 7, is

$$\mathbb{P}[C^{00}_{LR}(t_f) < c] = \mathbb{P}[C^{11}_{LR}(t_f) < c] = \mathbb{P}[C < c],$$
$$\mathbb{P}[C^{01}_{LR}(t_f) < c] = \mathbb{P}[C^{10}_{LR}(t_f) < c] = 1 - \mathbb{P}[C < -c],$$

where $\mathbb{P}[C < c]$ is defined below. Let

$$a = \left(\frac{1}{2N}\right)^{2r_{RS}/s} \quad \text{and} \quad b = \left(\frac{1}{2N}\right)^{4r_{LR}/s}.$$

Cases with $c > 0$ and $c < 0$ are treated separately below.



(a) (b)

FIGURE 7: Probability distributions of C^{ij}_{LR} for uniformly distributed initial gametic frequency configurations. These plots were generated using our analytic formulae for $\mathbb{P}[C^{ij}_{LR} < c]$, with $r_{LR}/s = 0.02$ and $1/(2N) = 0.00005$. As r_{RS}/s increases, $\mathbb{P}[C^{ij}_{LR} < c]$ for all ij become identical. (a) $\mathbb{P}[C^{00}_{LR}(t_f) < c]$ and $\mathbb{P}[C^{11}_{LR}(t_f) < c]$. (b) $\mathbb{P}[C^{01}_{LR}(t_f) < c]$ and $\mathbb{P}[C^{10}_{LR}(t_f) < c]$.

• For $c > 0$:

If $c \le b/4$, $\sqrt{b^2 - 4bc} > b(2a-1)$ and $c \le ab(1-a)$, then

$$\begin{aligned}\mathbb{P}[C < c] \;=\; & 1 - \frac{1}{4(1-a)^3 b^2}\left\{12ab(ab-2c)\log\left[\frac{a^2 b}{ab-c}\right]\right.\\ & +b\left(b + 6a^2 b - 8a^3 b + \sqrt{b^2-4cb}\right) - 2c\left(6b - 6ab + 5\sqrt{b^2-4cb}\right)\\ & \left. + 12c^2\log\left[\frac{2c^2}{(ab-c)(b-2c-\sqrt{b^2-4cb})}\right]\right\}.\end{aligned}$$

If $c \le b/4$, $\sqrt{b^2 - 4bc} > b(2a-1)$ and $c > ab(1-a)$, then

$$\mathbb{P}[C < c] \;=\; 1 - \frac{1}{2(1-a)^3 b^2}\left\{(b-10c)\sqrt{b^2-4bc} - 6c^2\log\left[\frac{2c-b+\sqrt{b^2-4bc})}{2c-b-\sqrt{b^2-4bc})}\right]\right\}.$$

If either $c > b/4$ or $\sqrt{b^2-4bc} \le b(2a-1)$, then $\mathbb{P}[C < c] = 1$.

• For $c = -|c| < 0$:

If $|c| \leq (1-a)^2 b/4$, then

$$\begin{aligned}\mathbb{P}[C < -|c|] \;\; = \;\; & \frac{1}{2(1-a)^3 b^2} \Big\{ \big[b(1-5a-2a^2) - 10|c| \big] \sqrt{(1-a)^2 b^2 - 4b|c|} \\ & + 6(2a^2b^2 + 4ab|c| + |c|^2) \log \left[\frac{(1+a)b + \sqrt{(1-a)^2 b^2 - 4b|c|}}{2\sqrt{b(ab+|c|)}} \right] \\ & - 6|c|^2 \log \left[\frac{b[(1-a)^2 b + (a-3)|c|] + [|c| - (1-a)b]\sqrt{(1-a)^2 b^2 - 4b|c|}}{2|c|\sqrt{b(ab+|c|)}} \right] \Big\} .\end{aligned}$$

If $|c| > (1-a)^2 b/4$, then $\mathbb{P}[C < -|c|] = 0$.

Polarization: Polarized LDs are measured with respect to major alleles. To determine the polarized LD $C_\omega(t_f)$ between the neutral loci, we compute

$$\sigma = [p_L^{ij}(t_f) - q_L^{ij}(t_f)] \times [p_R^{ij}(t_f) - q_R^{ij}(t_f)], \tag{51}$$

the main point being that $C_\omega(t_f) = C_{LR}(t_f)$ if $\sigma > 0$ and $C_\omega(t_f) = -C_{LR}(t_f)$ if $\sigma < 0$. (Recall that $C_{LR}(t)$ is measured with respect to type 1 alleles.)

First, for $r_{LR} = r_{RS} = 0$, note that

$$\sigma \approx (2\delta_{i1} - 1)(2\delta_{j1} - 1) = \begin{cases} +1, & \text{if } (i,j) = (0,0) \text{ or } (i,j) = (1,1), \\ -1, & \text{if } (i,j) = (0,1) \text{ or } (i,j) = (1,0). \end{cases} \tag{52}$$

Second, for fixed initial marginal frequencies, we need to determine for what values of r_{LR} and r_{RS}, σ changes sign. Using (45) and (46), we can obtain the following results:

For $i = 0$, $\sigma \approx 0$ if $p_L(0) > 1/2$ and

$$r_{RS} = \frac{s}{2\log(1/2N)} \log\left[\frac{p_L(0) - \frac{1}{2}}{p_L(0)}\right] - r_{LR}. \tag{53}$$

For $i = 1$, $\sigma \approx 0$ if $q_L(0) > 1/2$ and

$$r_{RS} = \frac{s}{2\log(1/2N)} \log\left[\frac{q_L(0) - \frac{1}{2}}{q_L(0)}\right] - r_{LR}. \tag{54}$$

For $j = 0$, $\sigma \approx 0$ if $p_R(0) > 1/2$ and

$$r_{RS} = \frac{s}{2\log(1/2N)} \log\left[\frac{p_R(0) - \frac{1}{2}}{p_R(0)}\right]. \tag{55}$$

For $j = 1$, $\sigma \approx 0$ if $q_R(0) > 1/2$ and

$$r_{RS} = \frac{s}{2\log(1/2N)} \log\left[\frac{q_R(0) - \frac{1}{2}}{q_R(0)}\right]. \tag{56}$$

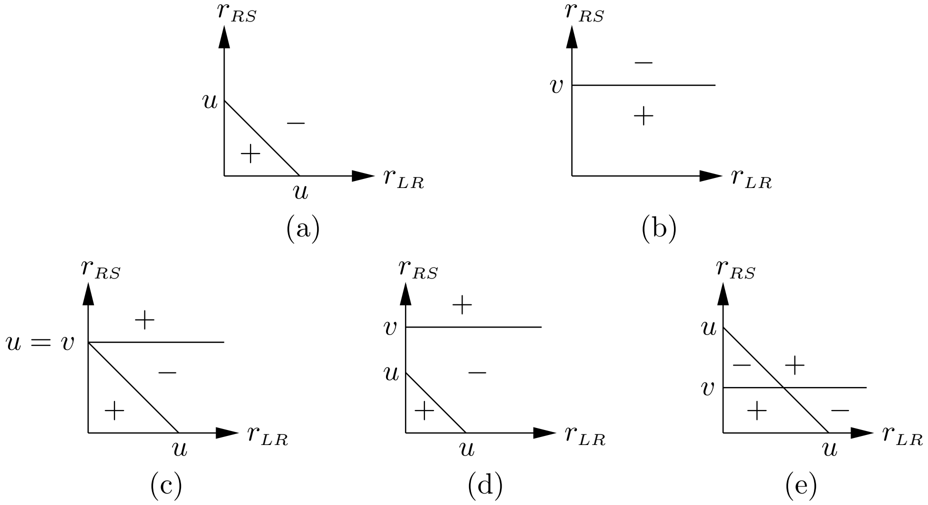


FIGURE 8: The sign of $\sigma = [p_L^{ij}(t_f) - q_L^{ij}(t_f)] \times [p_R^{ij}(t_f) - q_R^{ij}(t_f)]$ for $(i,j) = (0,0)$. The polarized LD $C_\omega(t_f)$ between the neutral loci is equal to $C_{LR}(t_f)$ if $\sigma > 0$ or $-C_{LR}(t_f)$ if $\sigma < 0$. The r_{LR},r_{RS} domain is partitioned into two, three, or four blocks, depending on $p_L(0)$ and $p_R(0)$. Note that σ tends to be positive in the neighborhood of $(r_{LR}, r_{RS}) = (0,0)$. The size and shape of this neighborhood depends on u and v, given by $u = \frac{s}{2\log(1/2N)} \log\left[\frac{p_L(0)-\frac{1}{2}}{p_L(0)}\right]$ and $v = \frac{s}{2\log(1/2N)} \log\left[\frac{p_R(0)-\frac{1}{2}}{p_R(0)}\right]$. (a) $p_L(0) > 1/2$ and $p_R(0) < 1/2$. (b) $p_L(0) < 1/2$ and $p_R(0) > 1/2$. (c) $p_L(0) = p_R(0) > 1/2$. (d) $p_L(0) > p_R(0) > 1/2$. (e) $p_R(0) > p_L(0) > 1/2$.

Combined with (52), these equations completely determine whether $C_\omega(t_f) = C_{LR}(t_f)$ or $C_\omega(t_f) = -C_{LR}(t_f)$ for given parameter values. For example, suppose that $(i,j) = (0,0)$. If $p_L(0) < 1/2$ and $p_R(0) < 1/2$, then there is no real-valued solution to the condition $\sigma = 0$, and therefore $C_\omega(t_f) = C_{LR}(t_f)$ for all values of r_{LR} and r_{RS}. If $p_L(0) > 1/2$ and $p_R(0) < 1/2$, or if $p_L(0) < 1/2$ and $p_R(0) > 1/2$, then σ changes sign as illustrated in Figure 8a,b, where

$$u = \frac{s}{2\log(1/2N)} \log\left[\frac{p_L(0) - \frac{1}{2}}{p_L(0)}\right] \quad \text{and} \quad v = \frac{s}{2\log(1/2N)} \log\left[\frac{p_R(0) - \frac{1}{2}}{p_R(0)}\right].$$

Note that u takes its minimum value at $p_L(0) = 1$ and that $u \to \infty$ as $p_L(0) \to 1/2$. Similarly, v takes its minimum value at $p_R(0) = 1$ and $v \to \infty$ as $p_R(0) \to 1/2$. If both $p_L(0) > 1/2$ and $p_R(0) > 1/2$, then there are three possibilities, depicted in Figure 8c,d,e.

Regions of positive $C_\omega(t_f)$: To determine the regions of positive $C_\omega(t_f)$, we need to know how the sign of $C_{LR}^{ij}(t_f)$ depends on r_{RS}; (43) implies that the sign of $C_{LR}^{ij}(t_f)$ does not depend on r_{LR}. For concreteness, suppose that $(i,j) = (0,0)$, in which case we can obtain the following results from using (43):

1. If $C_{LR}(0) \geq 0$, then $C_{LR}^{00}(t_f) \geq 0$ for all r_{RS}, and therefore the sign of $C_\omega(t_f)$ is com-

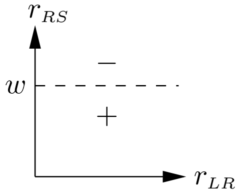


FIGURE 9: The sign of $C_{LR}^{00}(t_f)$ for $C_{LR}(0) < 0$. If $C_{LR}(0) \geq 0$, then $C_{LR}^{00}(t_f)$ is non-negative for all r_{RS} and r_{LR}. If $C_{LR}(0) < 0$, however, the sign of $C_{LR}^{00}(t_f)$ can change at $r_{RS} = w$, where $w = \frac{s}{2\log(1/2N)} \log\left[\frac{|C_{LR}(0)|}{p_L(0)p_R(0)}\right]$. In general, $C_{LR}^{00}(t_f)$ tends to be positive near $(r_{LR}, r_{RS}) = (0, 0)$.

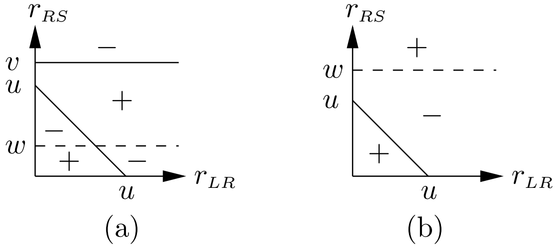


FIGURE 10: Examples of the sign of the polarized LD $C_\omega(t_f)$ when the gamete of origin is of type 001 and $C_{LR}(0) < 0$. Note that $C_\omega(t_f)$ is positive if and only if $C_{LR}^{00}(t_f)$ and σ are either both positive or both negative. In general, $C_\omega(t_f)$ tends to be positive near $(r_{LR}, r_{RS}) = (0, 0)$. (a) $p_L(0) > p_R(0) > 1/2$ and $w < u$. (b) $p_L(0) > 1/2$, $p_R(0) < 1/2$ and $w > u$.

pletely determined by that of σ.

2. If $C_{LR}(0) < 0$, then $C_{LR}^{00}(t_f)$ changes sign as illustrated in Figure 9, where

$$w = \frac{s}{2\log(1/2N)} \log\left[\frac{|C_{LR}(0)|}{p_L(0)p_R(0)}\right].$$

   Note that w takes its minimum value of zero at $|C_{LR}(0)| = p_L(0)p_R(0)$ and that it increases monotonically as $|C_{LR}(0)|/[p_L(0)p_R(0)]$ decreases. The polarized LD $C_\omega(t_f)$ is positive if and only if $C_{LR}^{00}(t_f)$ and σ are either both positive or both negative. Examples are shown in Figure 10.

As shown in Figure 8, σ tends to be positive in the neighborhood of $(r_{LR}, r_{RS}) = (0, 0)$. The size and shape of this neighborhood depends on u and v. Likewise, as shown in Figure 9, $C_{LR}^{00}(t_f)$ tends to be positive in the neighborhood of $(r_{LR}, r_{RS}) = (0, 0)$, with the size of the neighborhood depending on w. As a consequence, the polarized LD $C_\omega(t_f)$ also tends to be positive near $(r_{LR}, r_{RS}) = (0, 0)$.

More generally, if the gamete of origin is of type $ij1$, the sign of $C^{ij}_{LR}(t_f)$ (respectively, σ) can be analyzed using (43) (respectively, (52)-(56)). For all ij, the polarized LD $C_\omega(t_f)$ tends to be positive near $(r_{LR}, r_{RS}) = (0, 0)$.

An exact symmetry when the selected locus is outside the two neutral loci: Suppose that the selected locus is outside the two neutral loci, and that geometric configuration and recombination fractions are fixed. Let $\{p_S(0), p_L(0), p_R(0), C_{LS}(0), C_{RS}(0), C_{LR}(0), C_{LRS}(0)\}$ and $\{p'_S(0), p'_L(0), p'_R(0), C'_{LS}(0), C'_{RS}(0), C'_{LR}(0), C'_{LRS}(0)\}$ denote two different sets of initial conditions. At generation $n > 1$, we use "prime" to refer to the marginal allele frequencies and LDs obtained using the second set of initial conditions. In Appendix, we show that if $C_{LS}(0) = C'_{RS}(0)$ and $C_{RS}(0) = C'_{LS}(0)$, while $p_S(0) = p'_S(0)$, $C_{LR}(0) = C'_{LR}(0)$ and $C_{LRS}(0) = C'_{LRS}(0)$, then the system of full recursions (1)–(11) implies

$$C_{LR}(n) = C'_{LR}(n) \qquad \text{and} \qquad C_{LRS}(n) = C'_{LRS}(n)$$

for all $n \geq 1$. This is an exact symmetry result that holds for an arbitrary dominance coefficient h.

An application of this general result is the explanation of the symmetry of Figure 3 with respect to reflection about the $r = 0$ plane, for those regions corresponding to the selected locus being outside the neutral loci. Note that what is depicted in Figure 3 is different from the obviously symmetric case in which initial conditions $C_{LS}(0)$ and $C_{RS}(0)$ get exchanged when locus S is reflected about $r = 0$. In that figure, initial conditions remain fixed, while the geometric configuration of the loci and recombination fractions change upon reflection. That situation is related to changing initial conditions as described above, while keeping the geometric configuration and recombination fractions fixed.

DISCUSSION

To understand the forces that shape genomic variation in natural populations and the divergence between species, observed patterns must be compared to predictions of the models that faithfully represent the mechanisms through which such forces may work. While much of the natural selection of organismic phenotypes may be effectively approximated by deterministic single locus equations, interactive and stochastic forces are thought to play a significant role. Until recently genetic drift has been considered the primary stochastic process determining the temporal, geographic and genomic distribution of the vast majority of DNA sequence polymorphism and divergence. Gillespie has repeatedly demonstrated and emphasized fundamental differences between constant fitness models and stochastically varying selection, despite their superficial similarities (Gillespie 1994). Recently emerging results of surveys of genomic regions of low crossing over per physical length indicate that linked selection rather than genetic drift can dominate the levels of polymorphism within populations (Aguadé *et al.* 1989; Stephan and Langley 1989; Begun and Aquadro 1992). The hitchhiking effect not only reduces the average level of heterozygosity in the surrounding genomic regions, but it also leaves a skewed frequency spectrum (Braverman *et al.* 1995). The early study by Thomson (1977) indicated that linked selection can create linkage disequilibrium. Several subsequent papers have addressed specific cases (Robinson *et al.* 1991; Grote *et al.* 1998) or noted some temporal and spatial patterns (Kim and Nielsen 2004). Here we have demonstrated that the hitchhiking effect involves a number of strong and surprisingly distinct dynamics and patterns of linkage disequilibrium. We believe that the approach we have taken to address the impact of selection can be extended further to address more complex selection schemes and genetic interactions.

The technological capacity of molecular population genomics is increasing rapidly. For example, the HapMap Project (The International HapMap Consortium 2005) provides extensive genotypic survey results on more than one million SNPs in almost 300 individual humans. At this scale of observation one can anticipate much more powerful inferences about the role of direct selection, linked selection, crossing over, gene conversion, mutation, and geographic demography. Indeed, based on such new data, genomic variation in the rate of crossing over has been proposed as the primary determinant of the patterns of linkage disequilibrium in human populations (McVean *et al.* 2004).

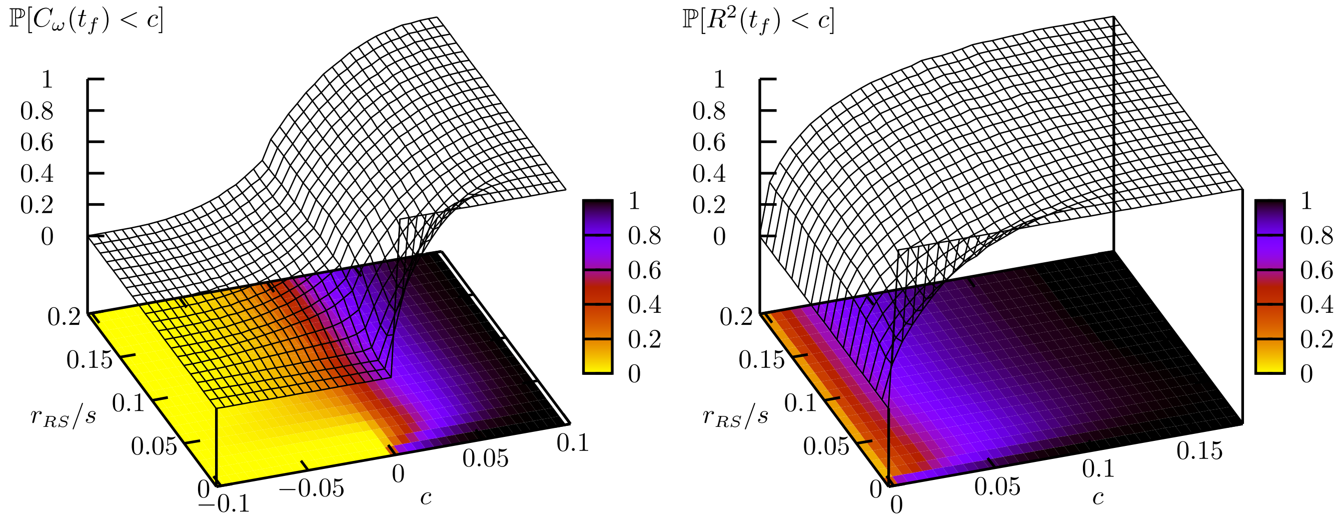
Several representations/notations have been developed to analyze the dynamics of multilocus systems (Bürger 2000). Through a series of papers Barton and Turelli have elaborated and applied their method based on the explicit representations of the moments of allele frequencies (Barton and Turelli 1987, 1991; Turelli and Barton 1990, 1994). Their representation proved surprisingly tractable and transparent in the analysis of the hitchhiking effect on linkage disequilibrium.

Our analysis begins with the full representation of the three locus dynamics using the notation of Barton and Turelli. These equations suggest the familiar approximation, "truncated equations," in which $r, s \ll 1$ and small higher order terms can be dropped. The truncated equations immediately expose much of the fundamental structure and their differential analogs, ordinary differential equations, allow approximate analytic solutions. Comparisons of these ODE dynamics with those of the Barton and Turelli representation indicate that the approximations remain quite accurate as long as $r/s \ll 1$ (see Table 1 and Table 2).

Particularly fortuitous and important are the role of the three locus LD C_{LRS} in driving the dynamics of the LD C_{LR} between the two linked neutral loci and the dependence of C_{LRS} on the product of the two locus LDs C_{LS} and C_{RS}.

This systematic investigation of the dynamics of LD under hitchhiking reveals four important features. First, and quite generally, hitchhiking indeed generates LD during the initial half of the hitchhiking time course. As Figure 3 shows LD (positive in this instance) reaches a maximum shortly before the originally rare selected allele reaches 0.5. This result is consistent with Thomson's analysis of hitchhiking caused by the dynamics of an initially rare allele under balancing selection in that its frequency reaches an equilibrium closer to 0.5 than to 1.0. But what is truly surprising is that from several important perspectives the hitchhiking effect on LD is one of reduction. In Figure 3 it is obvious that in the second half of the hitchhiking period the large peak of LD (positive in this case) decreases rapidly. Figure 2 shows several configurations of initial conditions and demonstrates that the decline in the magnitude of LD is not attributable to decline in the heterozygosity at the two neutral loci. A second and striking result is that preexisting LD is completely destroyed when the selected locus is situated between the neutral sites. This geometric relationship produces a striking pattern when all pairwise associations are plotted together as in Figure 4. This is probably the mechanism behind the pattern noted by Kim and Nielsen (2004). This LD reducing effect of hitchhiking is also evident when the selected site is outside the neutral pair since much of the LD generated during the initial phase is destroyed in the latter phase. A third unexpected property of the hitchhiking on LD is that the averaging over the frequencies of the gametes with which the rare selected variant can be associated indicates that the net effect of hitchhiking would be to reduce preexisting average LD. This is despite the fact that hitchhiking does tend to increase the variance in LD (see below). Notice in Figure 3 that there is considerably increased LD in both regions flanking the two neutral sites (i.e., when the selected site is outside and is close to the two neutral sites). When the rare favored allele appears on two of the other three haplotypes (10 or 01) the final LD is strongly negative. Thus the average (weighted by the frequencies of the four gametes) will remain at zero if there is no LD and tend toward zero if initially different from zero. The rate of approach to zero is greater than or equal to that expected in the absence of hitchhiking.

The fourth notable LD hitchhiking effect is on the expected LD when this association is polarized by the marginal allele frequencies. Langley and Crow (1974) noted that with molecular polymorphism data the sign of LD is typically arbitrary. They proposed to orient LD such that it reflected the deviation for the expected most frequent gametic type and



(a)

(b)

FIGURE 11: Probability distributions of the polarized LD $C_\omega(t_f)$ and the squared correlation coefficient $R^2(t_f)$, obtained from numerical simulations using $r_{LR}/s = 0.02$ and $1/(2N) = 0.00005$. The polarized LD $C_\omega(t_f)$ tends to be positive for small r_{RS}/s. (a) $\mathbb{P}[C_\omega(t_f) < c]$. (b) $\mathbb{P}[R^2(t_f) < c]$.

demonstrated that under quadratic stabilizing selection this measure of LD, denoted C_ω, is negative. Under hitchhiking the average C_ω tends to be positive. This can be understood as the consequence of the fact that the neutral alleles at each site on the initially selected haplotype tend to rise to frequencies greater than 0.5 and the LD between those alleles is positive. A bias in the distribution of C_ω either regionally or across the genome could be interpreted as evidence that hitchhiking is shaping LD. Thus the frequency averaged hitchhiking effect on LD is to drive it to zero. But as shown in Figure 11a there is a bias with respect to marginal frequencies at the two neutral sites; $C_\omega(t_f)$ tends to be positive for small r/s. And, of course, there is a broad range of r in which the variance of LD is increased when the selected site is outside the two neutral sites. Figure 11b shows that the projection of the probability distribution of the squared correlation coefficient $R^2(t_f)$ also has a peak for small r/s, near 0.02.

The genomic scale over which hitchhiking has a significant effect on heterozygosity and the frequency spectrum has been considered previously. Beyond the obvious $r \ll s$ inherent in the approximation, STEPHAN *et al.* (1992) showed that the reduction in heterozygosity was approximately proportional to $1 - (2N)^{-4r/s}$. Simulations of BRAVERMAN *et al.* (1995) indicated a similar scale and shape to the skewness in the frequency spectrum as measured by Tajima's D (also see DURRETT and SCHWEINSBERG 2005). In studying the expectation of the linkage disequilibrium caused by hitchhiking we notice a striking common function $A(t, r/s)$ that relates the averages of both heterozygosity and LD to what they would be in a large population in the absence of hitchhiking. We are tempted to speculate that this simple function may be fundamental to average dynamics of other moments of allele frequencies under the hitchhiking scenario.

While the effect of hitchhiking on the average C_{LR} is to drive it toward zero, this is clearly not expected for C_{LR}^2, R^2 or the absolute value of C_{LR}. We have not obtained an analytic expression for such expectations but simulated results such as those shown in Figures 3, 4 and 11 indicate that the hitchhiking effect on magnitude of C_{LR} between neutral sites on the same side of the selected site can be substantial.

Our results were derived using a deterministic three-locus model of hitchhiking. Similar results hold for the pseudohitchhiking model (GILLESPIE 2000 and unpublished results). We have compared both models. The recursion equations of the pseudohitchhiking model are a good approximation of the dynamics of the three-locus model if the selected locus is outside the two neutral loci and the distance between the selected locus to either one of the neutral loci is much larger than the distance between the two neutral loci. In this parameter region, LD predicted by both models decays more quickly than under neutrality. How might these conclusions about the theoretical hitchhiking dynamics of LD influence the interpretation of population genomic polymorphism and divergence? Certainly it seems to inform any effort to identify regions in the genomes of natural populations in which there has been very recent selected substitution of newly arising mutations or otherwise rare variants. Hitchhiking may not increase LD in the neighborhood of a selected site as it has been widely thought, rather it can decrease it especially when the neutral sites are on opposite sides of the selected locus (see Figure 4). More generally LD that is built up by hitchhiking shortly after the occurrence of a favored mutation is quickly destroyed (even before fixation is reached). As a consequence, genomic regions around targets of recent positive directional selection are expected to exhibit a lack of LD, which is not simply due to the variation-reducing force of hitchhiking. This local dip in the magnitude of LD may be of use in the localization of targets of positive selection in the genome. Given the current debate of how various variation-reducing forces can be distinguished (in particular, bottlenecks from selective sweeps; GLINKA *et al.* 2003; HADDRILL *et al.* 2005), there is merit in attempting to include the specific pattern of LD predicted by these analyses into the methods for identifying targets of selection by selective sweeps (e.g. KIM and STEPHAN 2002; KIM and NIELSEN 2004). Because populations that have undergone population size bottlenecks should show elevated genome-wide levels of LD, regions lacking LD around targets of selection may be more easily distinguishable from the rest of the loci than when statistics that are solely based on the reduction of variation are used.

We have not attempted to extend our results to situations in which recurring and genomically randomly distributed hitchhiking events occur. The significant impediment to the analysis of the effect of such recurrent hitchhiking on heterozygosity may be the impact of simultaneous events within the same genomic region. But if selection is strong and events sufficiently rare such occurrences may be negligible (KAPLAN *et al.* 1989; DURRETT and SCHWEINSBERG 2005). While this issue of the dynamic interaction of simultaneous linked hitchhiking events may well remain for the analysis of the impact of hitchhiking on LD, there is clearly a second considerable issue. While in large populations the heterozygosity does not change in between hitchhiking events, that is not true of LD which, of course, decays in magnitude at rate r. If the rate of recurrent and randomly distributed hitchhiking event were sufficiently rare and there were no other force causing LD, the results given above are applicable, since LD would decay to zero throughout the genomic region before the next event. Given that LD is, in fact, commonly present on some scale in the various studied species, further analysis and/or simulations are warranted to make a general prediction of the genomic pattern.

We thank Michael Turelli for teaching us the multilocus formalism and John Gillespie for supplying us with an unpublished manuscript on pseudohitchhiking. We wish to acknowledge support from the following sources: the Volkswagen-Foundation grant I/78 815 (WS); the National Science Foundation grants EIA-0220154 (YSS and CHL) and IIS-0513910 (YSS and CHL); and the National Human Genome Research Institute (NIH) grants 5R01HG002107-03 and 5R01HG002942-02 (CHL).

LITERATURE CITED

Aguadé, M., N. Miyashita and C. H. Langley, 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

Barton, N. H. and M. Turelli, 1987 Adaptive landscapes, genetic distance and evolution of quantitative characters. Genet. Res. **49:** 157–173.

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

Begun, D. J. and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site-frequency spectrum of DNA polymorphism. Genetics **140:** 783–796.

Bürger, R., 2000 *The Mathematical Theory of Selection, Recombination, and Mutation*, John Wiley and Sons, Chichester.

Durret, R. and J. Schweinsberg, 2005 A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Processes Appl. **115:** 1628–1657.

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

Gillespie, J. H., 1994 *The Causes of Molecular Evolution*, Oxford University Press, Oxford.

Gillespie, J. H., 1997 Junk ain't what junk does: neutral alleles in a selected context. Gene **205:** 291–299.

Gillespie, J. H., 2000 Genetic drift in an infinite population: The pseudohitchhiking model. Genetics **155:** 909–919.

Glinka, S., L. Ometto, S. Mousset, W. Stephan and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in Drosophila melanogaster: A multi-locus approach. Genetics **165:** 1269–1278.

Grote, M., W. Klitz, and G. Thomson, 1998 Constrained disequilibrium values and hitchhiking in a three-locus system. Genetics **150:** 1295–1307.

Haddrill, P. R., K. R. Thornton, B. Charlesworth and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res., **15:** 790–799.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167**, 1513–1524.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Langley, C. H. and J. F. Crow, 1974 The direction of linkage disequilibrium. Genetics **78:** 937–941.

McVean, G.A.T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

Robinson, W.P., A. Cambon-Thomsen, N. Borot, W. Klitz, and G. Thomson, 1991 Selection, hitchhiking and disequilibrium analysis at three linked loci with application to HLA data. Genetics **129:** 931–948.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res., Camb. **23:** 23–35.

Stephan, W. and C. H. Langley, 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the vermilion and forked loci. Genetics **121:** 89–99.

Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theor. Pop. Biol. **41:** 237–254.

The International HapMap Consortium, 2005 A haplotype map of the human genome. Nature **437**:1299–1320.

Thomson, G., 1977 The effect of a selected locus on linked neutral loci. Genetics **85:** 753–788.

Turelli, M. and N. H. Barton, 1990 Dynamics of polygenic characters under selection. Theor. Pop. Biol. **38:** 1–57.

Turelli, M. and N. H. Barton, 1994 Genetic and statistical analyses of strong selection on polygenic traits. What, me normal? Genetics **138:** 913–941.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

APPENDIX

Quasi-invariance (embedded selected locus case): In this section, we examine in more detail the case where the selected locus is between the neutral loci. More exactly, we wish to keep the sum $r_{LS} + r_{RS}$ fixed to some value, say ρ, and consider varying r_{LS} and r_{RS} while satisfying that condition. To avoid being long-winded, we call this kind of translation of the selected locus a *constrained S-translation.* We wish to show that the dynamics of certain linkage disequilibria is *quasi-invariant*, as we clarify presently, under the constrained S-translation.

First, note that the dynamics of p_S does not depend at all on the position of the selected locus. Then, as $r_{LR} = r_{LS} + r_{RS}$ for the case under consideration, recursions (6) and (10) do not change under the constrained S-translation. Since $r_{LS,R} = r_{RS}$ and $r_{RS,L} = r_{LS}$, we have $r_{LS,R}+r_{RS,L} = r_{RS}+r_{LS}$ and $r_{LRS} = r_{LR,S}+r_{RS}+r_{LS}$. If no double-crossovers are allowed, then $r_{LR,S}$ is identically zero. Note that $C_{LS}\tilde{\Delta}C_{RS}+C_{RS}\tilde{\Delta}C_{LS} = [g(r_{LS})+g(r_{RS})]C_{LS}C_{RS}$, and that the sum $g(r_{LS}) + g(r_{RS})$ does not change under the constrained S-translation. Therefore, recursions (7) and (11) do not change under the constrained S-translation, as long as $r_{LR,S}$ does not depend on where in-between the neutral loci the selected locus is located.

We now turn to the product $C_{LS}C_{RS}$. For ease of notation, we define

$$f(r;k) := [1 - p_S(k)\tilde{a}_{S,\varnothing}(k)][1 + q_S(k)\tilde{a}_{S,\varnothing}(k)] - r\left\{1 + \tilde{a}_{S,\varnothing}(k)[q_S(k) - p_S(k)] - \tilde{a}_{S,S}(k)p_S(k)q_S(k)\right\} \tag{A1}$$

Then, (4), (5), (8) and (9) imply

$$C_{LS}(k+1) = f(r_{LS};k)C_{LS}(k) \quad \text{and} \quad C_{RS}(k+1) = f(r_{RS};k)C_{RS}(k).$$

The product $C_{LS}C_{RS}$ satisfies the recursion

$$C_{LS}(k+1)C_{RS}(k+1) = f(r_{LS};k)f(r_{RS};k)C_{LS}(k)C_{RS}(k).$$

The quantity $f(r_{LS};k)f(r_{RS};k)$ can be written as

$$\alpha(k) + (r_{LS} + r_{RS})\beta(k) + r_{LS}r_{RS}\gamma(k),$$

where

$$\begin{aligned}
\alpha(k) &:= [1 - p_S(k)\tilde{a}_{S,\varnothing}(k)]^2[1 + q_S(k)\tilde{a}_{S,\varnothing}(k)]^2, \\
\beta(k) &:= [1 - p_S(k)\tilde{a}_{S,\varnothing}(k)][1 + q_S(k)\tilde{a}_{S,\varnothing}(k)]\left\{1 + \tilde{a}_{S,\varnothing}(k)[q_S(k) - p_S(k)] - \tilde{a}_{S,S}(k)p_S(k)q_S(k)\right\}, \\
\gamma(k) &:= \left\{1 + \tilde{a}_{S,\varnothing}(k)[q_S(k) - p_S(k)] - \tilde{a}_{S,S}(k)p_S(k)q_S(k)\right\}^2.
\end{aligned}$$

Under the restriction that $r_{LS} + r_{RS} = \rho$, the maximum value of $r_{LS}r_{RS}$ is $\rho^2/4$, whereas the minimum is 0. Since $\gamma(k)$ is positive definite, the maximum variation of $C_{LS}(n)C_{RS}(n)$, as the selected locus moves between the neutral loci, can be obtained by comparing $C_{LS}(n)C_{RS}(n)$ at $r_{LS}r_{RS} = 0$ with that at $r_{LS}r_{RS} = \rho^2/4$. Define *maximal relative variation* $\varepsilon(n)$ as

$$\varepsilon(n) := \frac{C_{LS}(n)C_{RS}(n)\ |_{r_{LS}=\frac{\rho}{2}, r_{RS}=\frac{\rho}{2}} \ - C_{LS}(n)C_{RS}(n)\ |_{r_{LS}=0, r_{RS}=\rho}}{C_{LS}(n)C_{RS}(n)\ |_{r_{LS}=0, r_{RS}=\rho}}.$$

It is straightforward to show that

$$\varepsilon(n) \quad = \quad \frac{\rho^2}{4}\sum_{k=1}^{n-1}\frac{\gamma(k)}{\alpha(k)+\rho\,\beta(k)} + \cdots ,$$

where "$\cdots$" represents terms proportional to ρ^m, $m \geq 4$. In the case of directional selection, $\gamma(k)/(\alpha(k)+\rho\,\beta(k))$ is of order 1 for all values of $s, h, p_S(k)$ and ρ. Therefore, $\varepsilon(n) = O(\rho^2 n)$, and we conclude that relative variation increases as time passes.

The dynamics of C_{LR} and C_{LRS} is almost (or quasi) invariant under the constrained S-translation in the following sense: the range of ρ in which selection has observable influence on the dynamics of C_{LR} and C_{LRS} is where $\rho \ll 1$. In that case, it is possible to maintain $\varepsilon(n) = O(\rho^2 n) \ll 1$ throughout the entire period from the initial generation to the fixation generation. We would then observe almost no variation in C_{LR} or C_{LRS} as the location of the selected locus is varied between the neutral loci. For large ρ, selection has little influence on C_{LR} and C_{LRS}, so their dynamics should be approximately invariant under translation of the selected locus.

An exact symmetry: Suppose that the selected locus is outside the two neutral loci, and that recombination fractions $r_{LS}, r_{RS}, r_{LR}, r_{LRS}, r_{LR,S}, r_{RS,L}$, and $r_{LS,R}$ appearing in the system of full recursions (1)–(11) are fixed. In what follows, the dominance coefficient h is assumed to be arbitrary. Let $\{p_S(0), p_L(0), p_R(0), C_{LS}(0), C_{RS}(0), C_{LR}(0), C_{LRS}(0)\}$ and $\{p'_S(0), p'_L(0), p'_R(0), C'_{LS}(0), C'_{RS}(0), C'_{LR}(0), C'_{LRS}(0)\}$ denote two different sets of initial conditions. At generation $n > 1$, we use "prime" to refer to the allele frequencies and LDs obtained using the second set of initial conditions.

We first consider the 2nd order LDs involving the selected locus.

Lemma 1 *Suppose that $C_{LS}(0) = C'_{RS}(0)$ and $C_{RS}(0) = C'_{LS}(0)$. Then, for all $n \geq 1$,*

$$C_{LS}(n)C_{RS}(n) = C'_{LS}(n)C'_{RS}(n).$$

Proof: This result follows from induction on n. Recall that

$$C_{LS}(n) = f(r_{LS}; n-1)C_{LS}(n-1) \quad \text{and} \quad C_{RS}(n) = f(r_{RS}; n-1)C_{RS}(n-1),$$

where the function $f(r,k)$ is defined as in (A1). Similarly,

$$C'_{LS}(n) = f(r_{LS}; n-1)C'_{LS}(n-1) \quad \text{and} \quad C'_{RS}(n) = f(r_{RS}; n-1)C'_{RS}(n-1).$$

If $C_{LS}(0) = C'_{RS}(0)$ and $C_{RS}(0) = C'_{LS}(0)$, then

$$\begin{aligned} C_{LS}(1)C_{RS}(1) &= [f(r_{LS};0)C_{LS}(0)] \times [f(r_{RS};0)C_{RS}(0)] \\ &= [f(r_{LS};0)C'_{RS}(0)] \times [f(r_{RS};0)C'_{LS}(0)] \\ &= [f(r_{LS};0)C'_{LS}(0)] \times [f(r_{RS};0)C'_{RS}(0)] = C'_{LS}(1)C'_{RS}(1). \end{aligned}$$

Suppose that the claim is true for all $1 \le n \le k$. Then, for $n = k+1$,

$$\begin{aligned} C_{LS}(k+1)C_{RS}(k+1) &= [f(r_{LS};k)C_{LS}(k)] \times [f(r_{RS};k)C_{RS}(k)] \\ &= f(r_{LS};k)f(r_{RS};k)C_{LS}(k)C_{RS}(k) \\ &= f(r_{LS};k)f(r_{RS};k)C'_{LS}(k)C'_{RS}(k) \\ &= [f(r_{LS};k)C'_{LS}(k)] \times [f(r_{RS};k)C'_{RS}(k)] \\ &= C'_{LS}(k+1)C'_{RS}(k+1), \end{aligned}$$

where the third line follows from the induction hypothesis. ■

Using the above lemma, we can obtain the following result regarding the 3rd order LD and the LD between the neutral loci:

Proposition 1 *Suppose that* $p_S(0) = p'_S(0)$*,* $C_{LS}(0) = C'_{RS}(0)$*,* $C_{RS}(0) = C'_{LS}(0)$*,* $C_{LR}(0) = C'_{LR}(0)$*, and* $C_{LRS}(0) = C'_{LRS}(0)$*. Then,*

$$C_{LR}(n) = C'_{LR}(n) \qquad \textit{and} \qquad C_{LRS}(n) = C'_{LRS}(n)$$

for all $n \ge 1$*.*

Proof: First, note that $p_S(0) = p'_S(0)$ implies $p_S(n) = p'_S(n)$, $\tilde{a}_{S,\varnothing}(n) = \tilde{a}'_{S,\varnothing}(n)$ and $\tilde{a}_{S,S}(n) = \tilde{a}'_{S,S}(n)$ for all $n \ge 1$. Therefore, since $C'_{LS}C'_{RS} = C_{LS}C_{RS}$ by Lemma 1, we obtain

$$\tilde{\Delta}C'_{LR} = -r_{LR}C'_{LR} + \tilde{a}_{S,\varnothing}(1-r_{LR})C'_{LRS} + \tilde{a}_{S,S}r_{LR}\,C_{LS}C_{RS},$$

which is equivalent to (10), and

$$\Delta C'_{LR} = \tilde{\Delta} C'_{LR} - \tilde{a}^2_{S,\varnothing} C'_{LS} C'_{RS} = \tilde{\Delta} C'_{LR} - \tilde{a}^2_{S,\varnothing} C_{LS} C_{RS},$$

which is equivalent to (6). Similarly,

$$\begin{aligned} \tilde{\Delta} C'_{LRS} &= \left[-r_{LRS} + \tilde{a}_{S,\varnothing}(1 - r_{LRS})(1 - 2p_S) + \tilde{a}_{S,S} r_{LR,S} p_S q_S\right] C'_{LRS} \\ &\quad - \tilde{a}_{S,\varnothing}(r_{LS,R} + r_{RS,L}) p_S q_S C'_{LR} \\ &\quad - \left[\tilde{a}_{S,\varnothing}(2 - r_{LS,R} - r_{RS,L}) - \tilde{a}_{S,S}(1 - 2p_S)(r_{LS,R} + r_{RS,L})\right] C_{LS} C_{RS} \end{aligned}$$

and

$$\begin{aligned} \Delta C'_{LRS} &= \tilde{\Delta} C'_{LRS} - \tilde{a}_{S,\varnothing} \left[p_S q_S \tilde{\Delta} C'_{LR} + C'_{LS} \tilde{\Delta} C'_{RS} + C'_{RS} \tilde{\Delta} C'_{LS}\right] + 2\tilde{a}^3_{S,\varnothing} p_S q_S C'_{LS} C'_{RS} \\ &= \tilde{\Delta} C'_{LRS} - \tilde{a}_{S,\varnothing} \left\{p_S q_S \tilde{\Delta} C'_{LR} + [g(r_{LS}) + g(r_{RS})] C'_{LS} C'_{RS}\right\} + 2\tilde{a}^3_{S,\varnothing} p_S q_S C'_{LS} C'_{RS} \\ &= \tilde{\Delta} C'_{LRS} - \tilde{a}_{S,\varnothing} \left\{p_S q_S \tilde{\Delta} C'_{LR} + [g(r_{LS}) + g(r_{RS})] C_{LS} C_{RS}\right\} + 2\tilde{a}^3_{S,\varnothing} p_S q_S C_{LS} C_{RS} \end{aligned}$$

are equivalent to (11) and (7), respectively. Hence, C'_{LR} and C'_{LRS} satisfy exactly the same set of recursions as do C_{LR} and C_{LRS}. Since $C'_{LR}(0) = C_{LR}(0)$ and $C'_{LRS}(0) = C_{LRS}(0)$, it thus follows that $C'_{LR}(n) = C_{LR}(n)$ and $C'_{LRS}(n) = C_{LRS}(n)$ for all $n \geq 1$. ■