DOI: 10.1534/genetics.107.081497

Linkage Disequilibrium Under Genetic Hitchhiking in Finite Populations

P. Pfaffelhuber,[1] A. Lehnert and W. Stephan

Ludwig-Maximilians University, Biocenter, 82152 Planegg, Germany

Manuscript received September 3, 2007
Accepted for publication February 22, 2008

ABSTRACT

The model of genetic hitchhiking predicts a reduction in sequence diversity at a neutral locus closely linked to a beneficial allele. In addition, it has been shown that the same process results in a specific pattern of correlations (linkage disequilibrium) between neutral polymorphisms along the chromosome at the time of fixation of the beneficial allele. During the hitchhiking event, linkage disequilibrium on either side of the beneficial allele is built up whereas it is destroyed across the selected site. We derive explicit formulas for the expectation of the covariance measure D and standardized linkage disequilibrium σ_D^2 between a pair of polymorphic sites. For our analysis we use the approximation of a star-like genealogy at the selected site. The resulting expressions are approximately correct in the limit of large selection coefficients. Using simulations we show that the resulting pattern of linkage disequilibrium is quickly—*i.e.*, in $<0.1N$ generations—destroyed after the fixation of the beneficial allele for moderately distant neutral loci, where N is the diploid population size.

THE detection of targets of positive selection using polymorphism data is an important research topic. There are two major patterns in DNA data that help to identify these targets. First, the fast fixation of a beneficial allele causes a reduction of neutral diversity at closely linked neutral loci and a distortion of the site-frequency spectrum. Second, the fast fixation of the beneficial allele causes an increased level of linkage disequilibrium (LD) around the selected site. Both patterns have been used to construct statistical tests to reject neutrality (Hudson *et al.* 1994; Kelly 1997; Depaulis and Veuille 1998; Fay and Wu 2000; Kim and Nielsen 2004).

While the diversity-reducing effect of genetic hitchhiking is well described on a quantitative level (*e.g.*, Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 1998; Etheridge *et al.* 2006), investigations of patterns of LD only started with Kim and Nielsen (2004), using numerical simulations. Analytical expressions for measures of LD after a selective sweep have been obtained by Stephan *et al.* (2006), who use differential equations to derive an expression for the covariance measure D [defined in (2)] between a pair of neutral alleles linked to a beneficial allele. This study was complemented by a genealogical (*i.e.*, backward in time) perspective in Pfaffelhuber and Studeny (2007) and McVean (2007).

The aim of this article is threefold: first, we describe a genealogical perspective of the joint genealogy of two neutral loci linked to a beneficial allele at the time of its fixation, which is accurate for large selection coefficients. Second, using the genealogical perspective, we derive an explicit analytic expression for standardized LD σ_D^2 [defined in (3)] at the end of a selective sweep. Our main result is given in (10). Third, we use simulations to see in which time frame before and after fixation we can observe a specific pattern of LD.

In our genealogical perspective we rely on the frequently used assumption that the genealogy at the selected site is exactly star-like at the end of the selective sweep. We show that genetic hitchhiking can lead to perfectly associated (*i.e.*, $\sigma_D^2 = 1$) alleles close to the selected site if both neutral loci are on the same side of the beneficial allele. If they are on different sides, LD is eliminated during the sweep. Interestingly, standardized LD σ_D^2 in a finite sample is much higher than in the whole population. All results on σ_D^2 at the time of fixation of the beneficial allele can be obtained from the explicit expressions that are found in Equation 10. Finally, our simulations show that the pattern of LD changes drastically shortly before and after fixation of the beneficial allele.

MODELS AND MEASURES OF LINKAGE DISEQUILIBRIUM

If a new beneficial allele B enters a population of N sexually reproducing diploid individuals, it might increase in frequency until it fixes in the population. If the fitness advantage of each copy of the B-allele is s and $Ns \gg 1$, the frequency X of the beneficial allele in

[1] *Corresponding author:* Mathematical Institute, Albert-Ludwigs University, Eckerstrasse 1, D-79104 Freiburg, Germany.
E-mail: peter.pfaffelhuber@stochastik.uni-freiburg.de

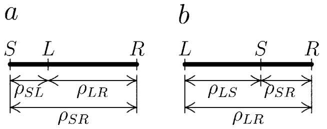

Figure 1.—The two possible geometries (a and b) of the selected (S) and the two neutral loci (L and R). The scaled recombination rates between the two loci are given by ρ_{SL}, ρ_{LR}, ρ_{LS}, and ρ_{SR}.

the population can be described by the differential equation

$$\dot{X} = \alpha X(1 - X), \quad X_0 = \varepsilon \tag{1}$$

(see, *e.g.*, Kaplan *et al.* 1989; Stephan *et al.* 1992), where $\alpha := 2Ns$ and time is measured in $2N$ generations. The process stops at time $T = 2\log(1/\varepsilon - 1)/\alpha$ when $X_T = 1 - \varepsilon$. In the following, we choose $\varepsilon = 1/\alpha$ since the fixation time of a beneficial allele is $\sim 2\log(\alpha)/\alpha$ if genetic drift is taken into account (Hermisson and Pennings 2005). In particular, we set $T := 2\log(\alpha)/\alpha$.

Maynard Smith and Haigh (1974) argued that neutral variants that are partially linked to the beneficial allele at $t = 0$ increase in frequency together with the beneficial allele. We extend this model to two neutral loci following Stephan *et al.* (2006). We have to take two possible geometries for the selected and the two neutral loci into account; see Figure 1. Either (a) the neutral loci are on the same side of the selected site or (b) the selected locus is in the middle of both neutral loci. Throughout we assume that mutation rates are sufficiently small that at most two alleles are segregating at both loci. At the selected S-locus we call b the wild-type and B the beneficial allele. For the other loci, we call the alleles L, ℓ at the first and R, r at the second neutral locus. The neutral loci are called the L/ℓ- and R/r-loci or, in short, the L- and R-loci.

During reproduction, recombination events might occur. If a recombination event occurs between two loci, they have different ancestors. Taking the recombination probability per generation between the two loci as r and measuring time in units of $2N$ generations, a recombination event splits the ancestry of the two loci at rate $\rho := 2Nr$. These scaled recombination rates between all pairs of loci are given in Figure 1. Note that $\rho_{SR} = \rho_{SL} + \rho_{LR}$ for geometry a and $\rho_{LR} = \rho_{LS} + \rho_{SR}$ for geometry b.

Let us denote the allelic frequencies at the neutral loci by q_L, q_ℓ, q_R, q_r, q_{LR}, q_{Lr}, $q_{\ell R}$, $q_{\ell r}$; *e.g.*, q_{LR} gives the fraction of the total population carrying both the L-allele at the L-locus and the R-allele at the R-locus.

Several statistics have been proposed to measure correlations, *i.e.*, LD, between two loci. Two of them are

$$D = q_{LR} - q_L q_R, \quad r^2 = \frac{D^2}{q_L(1 - q_L)q_R(1 - q_R)}. \tag{2}$$

Usually, data are obtained from samples only while these equations are based on population frequencies. As a consequence, measures for LD need to be corrected for finite sample size (Hudson 1985). Denoting allelic frequencies in the sample by $\hat{q}_L, \hat{q}_R, \ldots$, we obtain sample measures of LD by exchanging population frequencies with sample frequencies in (2), which results in the sample measures $\widehat{D}$ and $\widehat{r^2}$.

While one can obtain moments of the random variables $\widehat{D}$ for various demographic scenarios, even the expectation of $\widehat{r^2}$ is hard to obtain under a standard neutral model. (Note, however, the recent advances in Song and Song 2007.) It was argued by Hudson (1985) that standardized LD, introduced by Ohta and Kimura (1969),

$$\widehat{\sigma_D^2} = \frac{\mathbb{E}[\widehat{D}^2]}{\mathbb{E}[\hat{q}_L(1 - \hat{q}_L)\hat{q}_R(1 - \hat{q}_R)]}, \tag{3}$$

provides a good approximation of $\mathbb{E}[\widehat{r^2}]$ as long as low-frequency variants are ignored.

The star-like approximation: To approximate polymorphism patterns at the end of the selective sweep we use a genealogical perspective and introduce the star-like approximation. In this approximation we assume throughout that the selective sweep is so short that no new neutral mutations occur during fixation of the beneficial allele.

We proceed in three steps. First, we consider the selected site only; then we add a single neutral locus; finally, we add a second neutral locus. The latter approximation allows us to derive explicit expressions for $\mathbb{E}[D]$ and $\widehat{\sigma_D^2}$ at the end of the selective sweep in Equations 5 and 10.

The genealogy at the selected site: Consider a sample of beneficial alleles taken from the population at time T. Apparently, there is a single haploid individual at time 0 that is the ancestor of all individuals in the sample. In our analysis we make the assumption that this individual at time 0 is in fact the *most recent* common ancestor of all possible samples. Consequently, the genealogy at the selected site is star-like.

The assumption of a star-like genealogy at the selected site is frequently used in the analysis of selective sweeps (Maynard Smith and Haigh 1974; Fay and Wu 2000; McVean 2007). Moreover, it has been shown that it is accurate as long as $\log(\alpha)$ is large (Durrett and Schweinsberg 2004).

The genealogy at a linked neutral locus: If DNA sequences did not recombine the whole chromosome would share the same ancestry with the beneficial allele. However, by recombination, common ancestry is broken up. Let us consider the allele at a single neutral locus linked to the selected site carrying the beneficial allele B. It might be that an ancestor of this allele was linked to a wild-type allele b and only by recombination merged with a beneficial allele B. Following ancestral lines this means

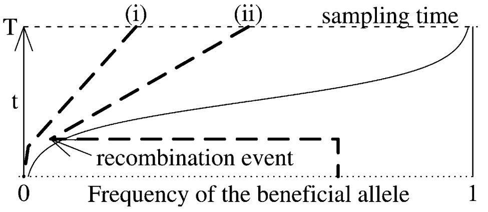

Figure 2.—Possible ancestries of a single neutral locus. For the allele at a neutral locus linked to the beneficial allele, either (i) it shares the ancestry of the linked beneficial allele or (ii) its ancestor at $t = 0$ was linked to a wild-type allele.

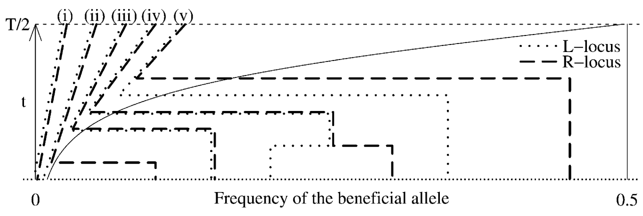

Figure 4.—Possible ancestries of two linked neutral loci for geometry a. For the alleles at the *L*- and *R*-loci there are five possible ancestries according to geometry a. Their probabilities are given in Table 1.

that the ancestral line changes its background from the beneficial to the wild-type background. Assuming that ρ is the scaled recombination rate between the beneficial and the neutral locus and the frequency of the beneficial allele is *X*, the instantaneous rate of changing backgrounds is ρ(1 − *X*). The probability that the ancestral line does not change backgrounds is thus (recall $\alpha := 2Ns$)

$$p(\rho) := \exp\left(-\int_0^T \rho(1 - X(t))\,dt\right) = \exp\left(-\frac{\rho}{\alpha}\log(\alpha)\right) \tag{4}$$

(Kaplan *et al.* 1989; Barton 1998). This event is shown in Figure 2 in case (i). With probability $1 - p(\rho)$ there was a recombination event and the neutral allele is linked to a wild-type one at time $t = 0$; this happened to line (ii) in Figure 2. We also say that the line escaped the sweep (backward in time). By the star-like approximation, each line of a finite sample escapes the sweep independently of the others. It has been shown that other events, *e.g.*, back-recombination into the beneficial background, occur only with low probability (Durrett and Schweinsberg 2004; Etheridge *et al.* 2006). Hence, we ignore such events here.

The joint genealogy at two linked neutral loci: To derive expressions for LD between two neutral sites we have to extend the star-like approximation. During the selective phase several recombination events might happen. To distinguish them, we speak, *e.g.*, of an *SL*-recombination event if it falls between the *S*- and the *L*-locus.

For both geometries we divide the time of the selective sweep into two halves. Toward the end of the selected phase, we assume that no recombinations to the wild-type background occur. The only events that occur in this phase are *LR*-recombination events to the effect that the alleles at both loci are linked to different beneficial alleles; see Figure 3. The probability that the ancestries of the alleles at the *L*- and *R*-loci do not split in this second half is approximately

$$\exp\left(-\rho\int_{T/2}^{T} X(t)\,dt\right) = \exp\left(-\frac{\rho}{\alpha}\int_{1/2}^{1-1/\alpha} \frac{1}{1 - X}\,dX\right) \approx p(\rho_{LR})$$

[recall (4)], where we used the fact that the contribution of times when $0 < X < \frac{1}{2}$ to the last integral is small. This case is shown for line (i) of Figure 3. With probability $1 - p(\rho_{LR})$ the alleles at the *L*- and *R*-loci have different ancestors, which both carry the beneficial allele at time $T/2$ as shown in line (ii) of Figure 3. In the latter case the ancestral lines of the alleles at the *L*- and the *R*-locus independently escape the sweep as in the case of a single neutral locus in Figure 2.

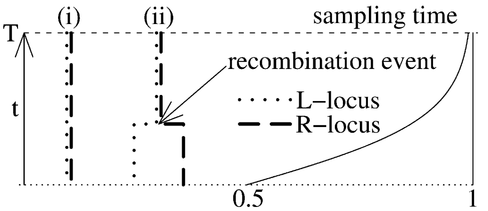

Figure 3.—Possible split of two linked neutral loci. Two alleles at the neutral loci linked to the beneficial allele either (i) have a common ancestor at time $T/2$ or (ii) have two different ancestors that are both linked to a beneficial allele.

For the joint genealogy of both neutral loci during the starting phase of the selective sweep we have to distinguish between geometries a and b. We set $p_\square := p(\rho_\square)$ for $\square = SL, LR, SR, LS, SR$. Let us first consider geometry a, where the selected locus is outside both neutral loci; see also Figure 4. All cases are listed in Table 1.

Consider line (i) as an example. We assume that all recombination events that split the alleles at the two loci such that both remain in the beneficial background already occurred in the late phase of the sweep. Hence, all recombination events automatically bring at least one allele to the wild-type background and both alleles stay linked in the beneficial background only if neither an *SL*- nor an *LR*-recombination event occurs. Since recombination events between both pairs occur independently and the probability that no recombination event brings an allele in scaled recombination distance ρ to the wild-type background is $p(\rho)$, it follows that case (i) has probability $p_{SL}p_{LR}$.

Observe that the effect for both lines (iv) and (v) is that the alleles at both loci are unlinked in the wild-type background. To produce one of these events there must be one *SL*- and one *LR*-recombination event. In line (iv) the first recombination event (backward in time) occurs between *S* and *L* and the second only between *L* and *R*,

TABLE 1

Probabilities of several events happening between times 0 and $T/2$ for geometry a; see Figure 4

Case	Event	Probability
(i)	No recombination event	$p_{SL}p_{LR}$
(ii)	An *LR*-recombination event makes the allele at the *R*-locus escape the sweep without the allele at the *L*-locus.	$p_{SL}(1-p_{LR})$
(iii)	By an *SL*-recombination event the line escapes the sweep and the alleles at the *L*- and the *R*-locus stay linked.	$(1-p_{SL})p_{LR}$
(iv)	An *SL*-recombination event brings the alleles at the *L*- and *R*-loci linked into the wild-type background; here, the ancestry of both alleles is split by an *LR*-recombination.	$\mathbb{P}[\text{(iv) or (v)}]$ $=(1-p_{SL})(1-p_{LR})$
(v)	An *LR*- and an *SL*-recombination event bring first the allele at the *R*-locus and then the allele at the *L*-locus into the wild-type background.	

All events are described backward in time.

while in line (v) the order is reversed. Altogether, either of the two events happens if and only if there is both an *SL*- and an *LR*-recombination event that results in the given probability of $(1-p_{SL})(1-p_{LR})$.

The genealogy for geometry b can be obtained similarly. Figure 5 and Table 2 give all the details. Observe that for geometry b it is not possible that an allele at the *L*- and one at the *R*-locus are linked in the wild-type background at $t=0$.

Again, by the star-like approximation, the ancestry of each line of a finite sample behaves independently of the other lines.

RESULTS

We are now in a position to obtain analytical results on measures of LD at the end of a selective sweep.

$\mathbb{E}[D(T)]$: Writing $D(0)$ and $D(T)$ for the LD measures at the beginning and end of the selective sweep, we obtain (using the star-like approximation)

$$\text{(a)}\ \mathbb{E}[D(T)] = p_{LR}^2(1-p_{SL}^2)\mathbb{E}[D(0)] \quad \text{(b)}\ \mathbb{E}[D(T)] = 0 \tag{5}$$

for geometries a and b, respectively. Note that (5) agrees approximately with Equation 47 in Stephan *et al.* (2006) for large values of α.

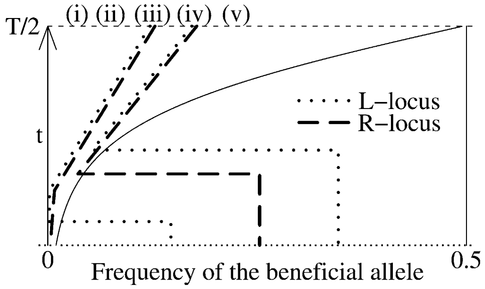

Figure 5.—The same as Figure 4 for geometry b. The lines (i), (ii), and (v) are the same as in Figure 4. The corresponding probabilities are given in Table 2.

To derive (5), consider a pair of one allele at the *L*- and one allele at the *R*-locus at the end of the sweep. In the case that the alleles are linked (*i.e.*, taken from the same individual at the end of the sweep) we denote the probability that both have the same ancestor (*i.e.*, their ancestors are linked) at the beginning of the sweep by *d*. Moreover, if the alleles are unlinked at the end of the sweep, we denote the probability that their ancestors are linked at the beginning of the sweep by *e*. Assuming no new mutations during the sweep,

$$\mathbb{E}[q_{LR}(T)] = d\cdot\mathbb{E}[q_{LR}(0)] + (1-d)\cdot\mathbb{E}[q_L(0)q_R(0)]$$

$$\mathbb{E}[q_L(T)q_R(T)] = e\cdot\mathbb{E}[q_{LR}(0)] + (1-e)\cdot\mathbb{E}[q_L(0)q_R(0)],$$

such that

$$\mathbb{E}[D(T)] = (d-e)\mathbb{E}[D(0)], \tag{6}$$

which leaves us with the task to compute *d* and *e* for geometries a and b. We have

$$\text{(a)}\ e = p_{SL}p_{SR} = p_{SL}^2p_{LR} \quad \text{(b)}\ e = p_{LS}p_{SR} = p_{LR} \tag{7}$$

because the ancestors of a pair of alleles that are unlinked at the end of the sweep can be linked at the beginning of the sweep only if none of the two ancestral lines recombines out of the sweep. Moreover, for *d*, we have two cases: either the two linked alleles split between *T*and $T/2$ like line (ii) in Figure 3 or they do not. If this happens, the probability for a common ancestor is the same as for the unlinked case, *d*. If the two alleles do not split between T and $T/2$, there must not be a recombination event separating them between $T/2$ and 0. So,

$$\text{(a)}\ d = (1-p_{LR})e + p_{LR}p_{LR} \quad \text{(b)}\ d = (1-p_{LR})e + p_{LR}p_{LS}p_{SR} = p_{LR}. \tag{8}$$

Combining (7) and (8) with (6) shows (5).

$\widehat{\sigma_D^2}$: To formulate our result on $\widehat{\sigma_D^2}$, we need the three quantities

TABLE 2

Probabilities for geometry b; see Figure 5

Line	Event	Probability
(i)	No recombination event	$p_{LS}p_{SR}$
(ii)	An *SR*-recombination event makes the allele at the *R*-locus escape the sweep without the allele at the *L*-locus.	$p_{LS}(1 - p_{SR})$
(iii)	An *LS*-recombination event makes the allele at the *L*-locus escape the sweep without the allele at the *R*-locus.	$(1 - p_{LS})p_{SR}$
(iv)	An *LS*-recombination event followed by an *SR*-recombination event brings the alleles at the *L*- and the *R*-locus into the wild-type background.	$\mathbb{P}[(\text{iv}) \text{ or } (\text{v})]$ $= (1 - p_{LS})(1 - p_{SR})$
(v)	Same as (iv) but in reverse order of the *LS*- and *SR*-recombination events.	

$$\begin{aligned}\mathcal{X}_t &:= \mathbb{E}[q_L(t)(1 - q_L(t))q_R(t)(1 - q_R(t))]\\ \mathcal{Y}_t &:= \mathbb{E}[D(t)(1 - 2q_L(t))(1 - 2q_R(t))]\\ \mathcal{Z}_t &:= \mathbb{E}[(D(t))^2] \qquad (9)\end{aligned}$$

for $0 \leq t \leq T$. At the end of the selective sweep, we show that if the sample size n is large enough such that terms of order $1/n^2$ can be ignored, we have

$$\begin{aligned}&(\text{a})\ \widehat{\sigma_D^2} = \frac{p_{LR}^4(1 - p_{SL})(p_{SL}^2(\mathcal{X}_0 + \mathcal{Y}_0) + (1 + 2p_{SL})\mathcal{Z}_0) + (1/n)\zeta}{(1 - p_{SR})((1 + p_{SL} + p_{SR})\mathcal{X}_0 + p_{SL}p_{SR}(\mathcal{X}_0 + \mathcal{Y}_0)) + (1/n)\chi},\\ &(\text{b})\ \widehat{\sigma_D^2} = \frac{1}{n - 2}, \qquad (10)\end{aligned}$$

where $\mathcal{X}_0, \mathcal{Y}_0$, and $\mathcal{Z}_0$ denote the three quantities (9) at the beginning of the sweep. Moreover, ζ and χ are corrections according to the finite sample size for geometry a and are given by

$$\begin{aligned}\zeta := {}&((1 - p_{SR}^2)(1 + p_{SL}) + 4p_{SR}^4 - 3p_{SR}^2p_{LR}^2(1 - p_{SL}))\mathcal{X}_0\\ &+ p_{LR}((1 - p_{SR})p_{SL}^2 + p_{LR}(1 + p_{SL}(1 - 2p_{SL})(1 - 2p_{SR}))\\ &\qquad - 3p_{SR}^2p_{LR}(1 - p_{SL}))\mathcal{Y}_0\\ &+ p_{LR}^3(4p_{SL}^2 - 3p_{LR}(1 - p_{SL})(1 + 2p_{SL}))\mathcal{Z}_0,\\ \chi := {}&(4p_{SR}^3 - 2(1 - p_{SR}^2)(1 + p_{SL}))\mathcal{X}_0\\ &+ p_{LR}(p_{LR}(1 + p_{SL}(1 - 2p_{SL})(1 - 2p_{SR}))\\ &\qquad - 2p_{SL}^2(1 - p_{SR}))\mathcal{Y}_0\\ &+ 4p_{SR}^2p_{LR}\mathcal{Z}_0. \qquad (11)\end{aligned}$$

If the population at time 0 is in equilibrium, both loci mutate with probability u and we set $\theta := 4Nu$. Ohta and Kimura (1969) have shown that

$$\begin{aligned}\mathcal{X}_0 &= \frac{1}{4}\cdot\frac{\theta^2}{1 + \theta}\\ &\quad\cdot\frac{(5 + 2\theta + \rho_{LR})(3 + 2\theta + 2\rho_{LR}) - 4}{(1 + \theta)(3 + 2\theta + 2\rho_{LR})(5 + 2\theta + \rho_{LR}) - 2(3 + 2\theta)}\\ \mathcal{Y}_0 &= \frac{\theta^2}{1 + \theta}\\ &\quad\cdot\frac{1}{(1 + \theta)(3 + 2\theta + 2\rho_{LR})(5 + 2\theta + \rho_{LR}) - 2(3 + 2\theta)}\\ \mathcal{Z}_0 &= \frac{1}{4}\cdot\frac{\theta^2}{1 + \theta}\\ &\quad\cdot\frac{2\theta + \rho_{LR} + 5}{(1 + \theta)(3 + 2\theta + 2\rho_{LR})(5 + 2\theta + \rho_{LR}) - 2(3 + 2\theta)}. \qquad (12)\end{aligned}$$

Assuming that the population was in neutral equilibrium when the sweep started, we predict the pattern of LD for $\alpha = 1000$, $n = 20$, a per-site mutation rate of $\theta = 0.005$, and $\rho = 0.025$ between two adjacent bases shown in Figures 6 and 7. Note that selection coefficients in the order of $\alpha = 1000$ are observed in practice (Beisswanger *et al.* 2006). Significant amounts of LD build up on each side of the selected site, but there is no LD for a pair of polymorphisms from both sides of the selected site. In Figure 7 we assume that two neutral polymorphisms have a fixed distance and consider the dependence of LD on their distance to the selected site. We see here that the finite sample size has a profound effect on the level of LD. Moreover, even for $\rho_{LR} = 50$ a twofold increase of LD relative to neutral expectations can be expected if both neutral loci are in a 2-kb distance from the selected site.

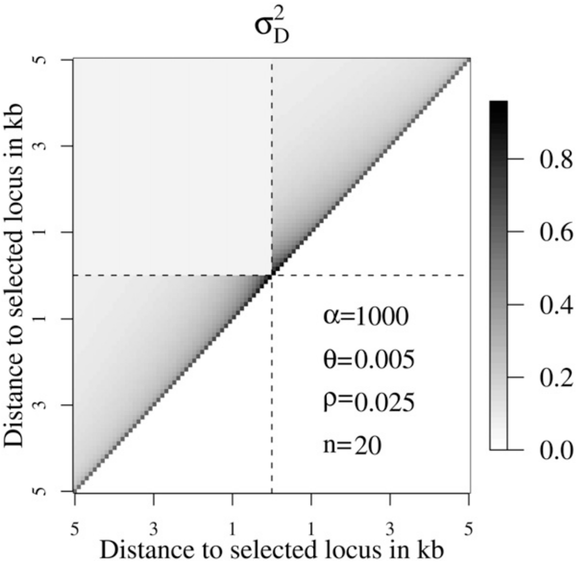

Figure 6.—Plot of analytical results from (10) and (12). If the population was in equilibrium before the selective sweep, we see a distinct pattern of LD around the genomic target of selection at the time of fixation $t = T$. Here, a 10-kb stretch of DNA is shown and $\rho = 2Nr$ is the scaled recombination rate between two adjacent sites.

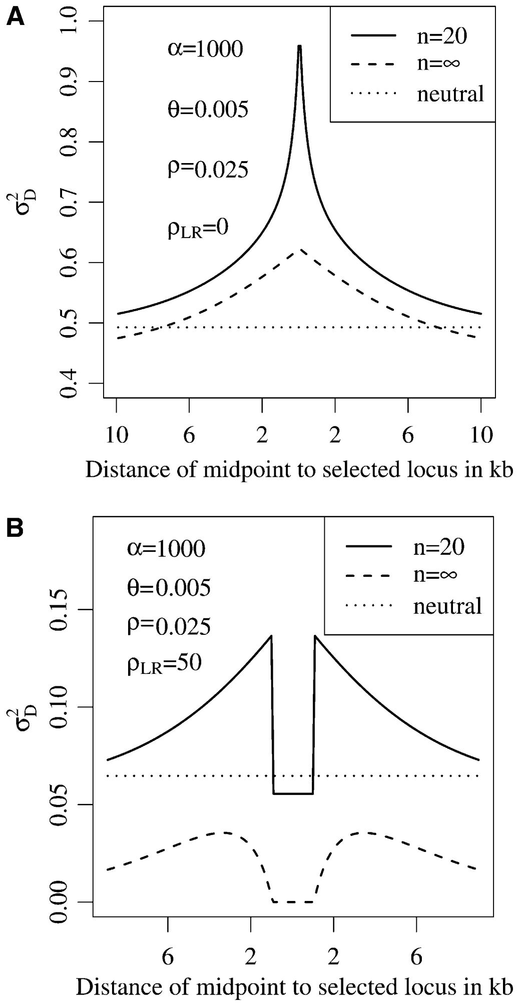

Figure 7.—Plot of results from (10) and (12). With the same parameter values as in Figure 6 we look at two loci that are (A) 0 kb and (B) 2 kb apart. The *x*-axis shows the distance between the selected site and the midpoint between the loci under consideration. The dotted curve shows standardized LD in a sample of $n = 20$ under a standard neutral model.

The big effect of the finite sample size ($n = 20$ in the numerical example) close to the selected site on $\widehat{\sigma_D^2}$ can be seen from (10). Note that for $\rho_{SL} \approx \rho_{SR} \approx \rho_{LR} \approx 0$ we have $p_{SL} \approx p_{SR} \approx p_{LR} \approx 1$ and we find that $\widehat{\sigma_D^2} \approx \zeta/\chi \approx 1$. However, mutations in the region close to the selected site are rarely observed.

To derive (10) we show that for geometry a

$$\begin{aligned}\mathcal{X}_T &= (1-p_{SL})(1-p_{SR})\\ &\quad\cdot((1+p_{SL}+p_{SR})\mathcal{X}_0 + p_{SL}p_{SR}(\mathcal{X}_0+\mathcal{Y}_0)),\\ \mathcal{Y}_T &= (1-p_{SL})p_{LR}^2(4p_{SL}p_{SR}(p_{SL}\mathcal{X}_0+\mathcal{Z}_0)\\ &\qquad + (1+p_{SL}(1-2p_{SL})(1-2p_{SR}))\mathcal{Y}_0),\\ \mathcal{Z}_T &= p_{LR}^4(1-p_{SL})^2((p_{SL}^2(\mathcal{X}_0+\mathcal{Y}_0)+(1+2p_{SL})\mathcal{Z}_0)\end{aligned} \tag{13}$$

and for geometry b

$$\begin{aligned}\mathcal{X}_T &= (1-p_{LS}^2)(1-p_{SR}^2)\mathcal{X}_0 + p_{LS}(1-p_{LS})p_{SR}(1-p_{SR})\mathcal{Y}_0,\\ \mathcal{Y}_T &= 0\\ \mathcal{Z}_T &= 0.\end{aligned} \tag{14}$$

These results, together with (A4) and $\widehat{\sigma_D^2} = \widehat{\mathcal{Z}}_T/\widehat{\mathcal{X}}_T$, then imply (10).

There is a close relationship between $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ and pairwise heterozygosities as described in the appendix. There are three measures for pairwise heterozygosity we have to take into account. Consider two pairs of one allele at the *L*- and one allele at the *R*-locus each, taken from the population at time *T*. Let f_T, g_T, and h_T be the probabilities that both pairs are heterozygous if both pairs are linked, only one pair is linked, and both pairs are unlinked, respectively; see also the appendix. The quantities f_0, g_0, and h_0 are defined analogously for the population at time 0. Moreover, we take $f_{T/2}$, $g_{T/2}$, and $h_{T/2}$ as the corresponding pairwise heterozygosities if the two pairs of one allele at the *L*- and one allele at the *R*-locus each are taken from the *beneficial* background at time $T/2$. To obtain (13) and (14) we consider a sample taken at time *T*. First, splits of linked alleles at the *L*- and *R*-loci in the beneficial background are generated between *T*and $T/2$. For both geometries, we obtain

$$\begin{pmatrix} f_T \\ g_T \\ h_T \end{pmatrix} = C \cdot \begin{pmatrix} f_{T/2} \\ g_{T/2} \\ h_{T/2} \end{pmatrix},$$

$$C := \begin{pmatrix} p_{LR}^2 & 2p_{LR}(1-p_{LR}) & (1-p_{LR})^2 \\ 0 & p_{LR} & 1-p_{LR} \\ 0 & 0 & 1 \end{pmatrix}. \tag{15}$$

To see this, consider two linked pairs of alleles at the *L*- and *R*-loci as an example. These are heterozygous at both the *L*- and the *R*-locus if none of them splits (which occurs with probability p_{LR}^2), one of them splits [probability $2p_{LR}(1 - p_{LR})$], or both split [probability $(1 - p_{LR})^2$] and the resulting pairs of *L*- and *R*-loci are heterozygous.

Furthermore, using the star-like approximation we can compute $f_{T/2}$, $g_{T/2}$, and $h_{T/2}$. For example, consider a linked pair of one allele at the *L*- and one allele at the *R*-locus in the beneficial background at time $T/2$. One possibility that it is heterozygous at both loci is that their ancestors at time 0 are a linked pair of one *L*- and one *R*-locus at time 0 and these are heterozygous. The probability for this event (which is denoted a_{11} below) is for geometry a given as $(1 - p_{SL}^2)p_{LR}^2$ since at least one *SL*-recombination event and no *LR*-recombination event must occur. In other words, one of the two lines is like (iii) in Figure 4 while the other is either like (i) or (iii) in the same figure. For geometry b this probability is 0 because it is not possible to have a linked pair of one *L*- and one *R*-locus in the wild-type background at time 0 as can be seen from Figure 5.

As a second example consider a_{23} for geometry a. This is the probability that one linked and one unlinked pair of one allele at the L- and one allele at the R-locus each, taken from the beneficial background at time $T/2$, have four different ancestors at time 0. Either the ancestral line of exactly one allele at the L-locus stays in the beneficial background [probability $2p_{SL}(1-p_{SL})$] and both alleles at the R-locus escape the sweep [probability $(1-p_{LR})(1-p_{SR})$] or both alleles at the L-locus are linked to a wild-type allele at the beginning of the sweep [probability $(1-p_{SL})^2$] and the linked pair is split by an LR-recombination event [probability $(1-p_{LR})$].

Altogether we have

$$(f_{T/2}, g_{T/2}, h_{T/2})^\top = A\cdot(f_0, g_0, h_0)^\top \tag{16}$$

with $A=(a_{ij})_{1\le i,j\le 3}$. For geometry a, A has the form

$$\begin{aligned}
a_{11} &= (1-p_{SL}^2)p_{LR}^2,\\
a_{12} &= 2(1-p_{SL}^2)p_{LR}(1-p_{LR}),\\
a_{13} &= (1-p_{SL}^2)(1-p_{LR})^2,\\
a_{21} &= (1-p_{SL})p_{SR}^2,\\
a_{22} &= p_{SR}(1-p_{SL})(1-p_{SR}) + (1-p_{SL})p_{LR}(1-p_{SL}p_{SR})\\
&\quad + 2p_{SL}p_{SR}(1-p_{SL})(1-p_{LR})\\
&= p_{LR}(1-p_{SL})(p_{SL}(1-p_{SR}) + 1 - p_{SL}p_{SR}\\
&\quad + 2p_{SL}^2(1-p_{LR}))\\
&= p_{LR}(1-p_{SL})(1+p_{SL}-4p_{SL}p_{SR}+2p_{SL}^2),\\
a_{23} &= 2p_{SL}(1-p_{SL})(1-p_{LR})(1-p_{SR}) + (1-p_{SL})^2(1-p_{LR})\\
&= (1-p_{SL})(1-p_{LR})(1+p_{SL}-2p_{SL}p_{SR}),\\
a_{31} &= 0,\\
a_{32} &= 4(1-p_{SL})p_{SL}(1-p_{SR})p_{SR},\\
a_{33} &= 2(1-p_{SL})p_{SL}(1-p_{SR})^2 + 2(1-p_{SL})^2(1-p_{SR})p_{SR}\\
&\quad + (1-p_{SL})^2(1-p_{SR})^2\\
&= (1-p_{SL})(1-p_{SR})(1+p_{SL}+p_{SR}-3p_{SL}p_{SR}).
\end{aligned}$$

For geometry b, LS- and SR-recombination events occur independently, leading to

$$\begin{aligned}
a_{11} &= a_{21} = a_{31} = 0\\
a_{12} &= a_{22} = a_{32} = 4p_{LS}(1-p_{LS})p_{SR}(1-p_{SR})\\
a_{13} &= a_{23} = a_{33}\\
&= (1-p_{LS})(1-p_{SR})\\
&\quad\cdot(2p_{LS}(1-p_{SR}) + 2(1-p_{LS})p_{SR} + (1-p_{SL})(1-p_{SR})).
\end{aligned}$$

Combining (A3), (15), and (16) we can write for both geometries

$$(\mathcal{X}_T, \mathcal{Y}_T, \mathcal{Z}_T)^\top = B\cdot C\cdot A\cdot B^{-1}(\mathcal{X}_0, \mathcal{Y}_0, \mathcal{Z}_0)^\top.$$

For geometry a

$$\begin{aligned}
&B\cdot C\cdot A\cdot B^{-1}\\
&= \frac{1}{4}\begin{pmatrix} 0 & 0 & 1\\ 0 & 4p_{LR} & -4p_{LR}\\ p_{LR}^2 & -2p_{LR}^2 & p_{LR}^2\end{pmatrix}\\
&\quad\cdot\begin{pmatrix} 4(1-p_{SL}^2) & 2(1-p_{SL}^2)p_{LR} & 4(1-p_{SL}^2)p_{LR}^2\\ 4(1-p_{SL})(1+p_{SL}-p_{SR}^2) & (1-p_{SL})p_{LR}(1+p_{SL}-2p_{SL}p_{SR}+2p_{SL}^2) & 4(1-p_{SL})p_{SR}^2\\ 4(1-p_{SL}^2)(1-p_{SR}^2) & 4(1-p_{SL})p_{SL}(1-p_{SR})p_{SR} & 0\end{pmatrix}\\
&= (1-p_{SL})\\
&\quad\cdot\begin{pmatrix} (1+p_{SL})(1-p_{SR}^2) & p_{SL}(1-p_{SR})p_{SR} & 0\\ 4p_{SR}^3 & p_{LR}^2(1+p_{SL}(1-2p_{SL})(1-2p_{SR})) & 4p_{LR}p_{SR}^2\\ p_{LR}^4(1-p_{SL})p_{SL}^2 & p_{LR}^4(1-p_{SL})p_{SL}^2 & p_{LR}^4(1-p_{SL})(1+2p_{SL})\end{pmatrix},
\end{aligned}$$

which shows (13). For geometry b we have similarly

$$\begin{aligned}
&B\cdot C\cdot A\cdot B^{-1}\\
&= (1-p_{LS})(1-p_{SR})\\
&\quad\cdot\begin{pmatrix} 0 & 0 & 1\\ 0 & 4p_{LR} & -4p_{LR}\\ p_{LR}^2 & -2p_{LR}^2 & p_{LR}^2\end{pmatrix}\\
&\quad\cdot\begin{pmatrix} (1+p_{LS})(1+p_{SR}) & p_{LS}p_{SR} & 0\\ (1+p_{LS})(1+p_{SR}) & p_{LS}p_{SR} & 0\\ (1+p_{LS})(1+p_{SR}) & p_{LS}p_{SR} & 0\end{pmatrix}\\
&= \begin{pmatrix} (1-p_{LS}^2)(1-p_{SR}^2) & p_{LS}(1-p_{LS})p_{SR}(1-p_{SR}) & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},
\end{aligned}$$

which gives (14).

Simulations: We use the program SSW (Kim and Stephan 2002) to simulate data under a selective sweep and compare these simulations to our predictions for $\widehat{\sigma_D^2}$ from (10). We changed the program to set $\varepsilon = 1/\alpha$ in (1). The parameter values in our simulations coincide with those taken for Figures 6 and 7. We consider a 20-kb stretch of DNA in a sample of $n = 20$ taken at the time a beneficial mutation with $\alpha = 1000$ has fixed. Here, the sweep region where levels of polymorphism are reduced by at least 50% consists of ~10 kb.

The heuristics of Hudson (1985) that $\mathbb{E}[r^2]$ and σ_D^2 coincide approximately if we ignore low-frequency variants are also valid at the end of a selective sweep (consult the supporting online supplemental material to see numerical results). Moreover, in Figure 8 we compare simulated data to predictions from (10) for $n = 20$. Here, we divide the 20-kb stretch of DNA into 100 bins of 0.2 kb each and measure LD between SNPs of two different bins. In Figure 8A, we use adjacent bins while Figure 8B shows results for bins that are 2 kb apart. We see that LD is highly elevated for the closely linked pair that is also seen in (10). The fit between simulated data and our predictions is worse for smaller values of α and larger n (see supporting online supplemental material). While the deviation from the numerical results is as large as 25% in Figure 8B, *i.e.*, for $\alpha =$

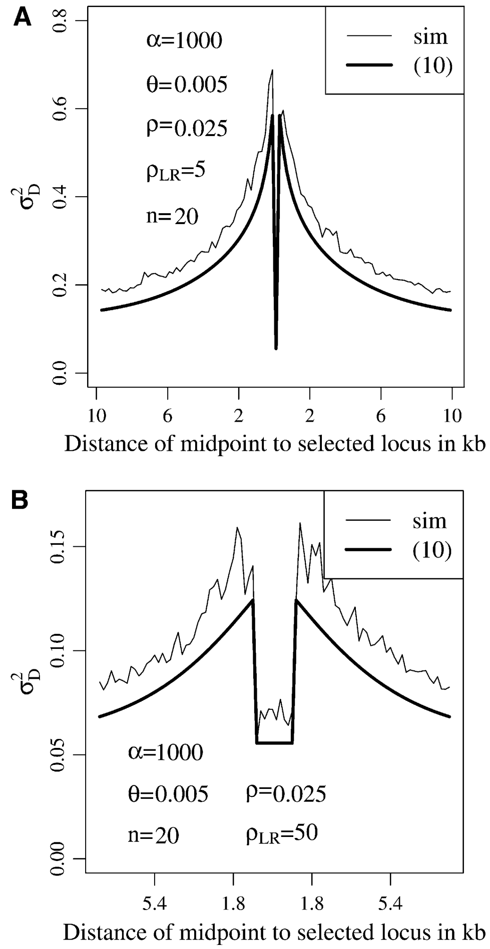

Figure 8.—Comparison of simulations and prediction using the star-like approximation from (10) and (12). The neutral loci in the simulation fall in windows that are (A) 0.2 kb and (B) 2 kb apart. Every curve is based on 10^3 simulations.

1000, it increases to 30% for $\alpha = 500$ and decreases to 20% for $\alpha = 2000$ with the same values for ρ_{LR}/α, respectively. The worse fit of the analytical results for the larger sample size can also be explained. The genealogy of larger samples is more complex and thus may differ from the star-like approximation in several ways.

For data analysis it is most important to see how long such a pattern can be observed. In Figure 9 we analyze the pattern of LD in the sweep region at three time points: before fixation when the frequency of the beneficial allele is 0.95 (which is for the given parameters the time $t \approx T - 0.01N$), at the time of fixation, and $0.1N$ generations afterward. Two observations can be made here. First, LD between both sides of the selected site is destroyed only at the very end of a selective sweep. Second, while LD for closely linked (0.2 kb, which equals $\rho_{LR} = 5$) neutral variants is still elevated after $0.1N$ generations, the effect of selection on LD completely vanishes for more distant (2 kb, which equals $\rho_{LR} = 50$) neutral loci. A closer analysis reveals that the decay of LD is fastest directly after the selective sweep (see supporting online supplemental material).

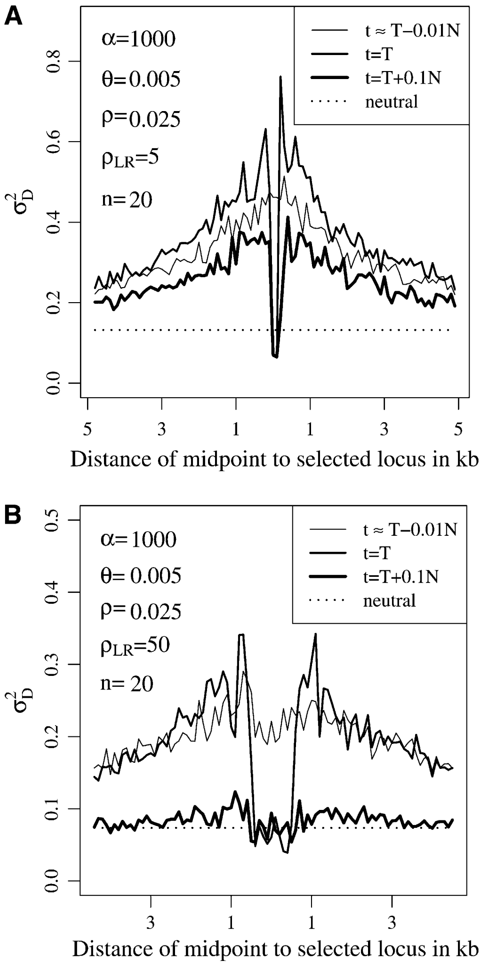

Figure 9.—The pattern of LD at different time points. The $t \approx T - 0.01N$ curve gives standardized LD at the time the beneficial allele has reached 0.95. The $t = T$ and $t = T + 0.1N$ curves describe the pattern at the time of fixation and $0.1N$ generations afterward, respectively. Both neutral loci are (A) 0.2 kb and (B) 2 kb apart. Every curve is based on 10^3 simulations of the 10-kb fragment.

DISCUSSION

Recently several statistical tests to infer selection using patterns of LD have been developed (Hudson *et al.* 1994; Depaulis and Veuille 1998; Sabeti *et al.* 2002; Toomajian *et al.* 2003; Kim and Nielsen 2004; Hanchard *et al.* 2006; Wang *et al.* 2006). The heuristics behind these tests are as follows: if a beneficial allele enters the population and increases in frequency, neutral variants increase in frequency by genetic hitchhiking. Recombination did not have much time during the selective sweep to break up linkage between these neutral polymorphisms. As a consequence, we see alleles that have both high frequency—typical for old alleles under neutrality—and long-range associations with other alleles, which is typical for young alleles (Sabeti *et al.* 2006).

In a simulation study, Jensen *et al.* (2007) carry out a power analysis of the test developed in Kim and Nielsen (2004). They show that distinct patterns of LD vanish within $0.1N$ generations after fixation of the beneficial allele. Such a signal is too weak to produce significant results using the overall pattern of LD. However, using the increased level of LD between tightly linked polymorphisms it might be possible to distinguish recurrent sweeps from neutrality or other demographic scenarios, for example, population bottlenecks.

On a fine scale, the effect of genetic hitchhiking on LD at the time of fixation can be described as follows (see also Figure 6): on either side of the beneficial allele, correlations between existent polymorphisms are built up, leading to long-range LD. Between the two sides of the beneficial allele LD is destroyed. This destruction can be explained heuristically: the observation of polymorphisms on any side of the beneficial allele (assuming no new mutations in the sweep) requires a recombination event between the beneficial allele and the neutral polymorphisms. By this recombination event a large haplotype is introduced into the population, leading to strong LD on each side of the beneficial allele. The existence of two neutral polymorphisms on both sides of the beneficial allele requires two independent recombination events, one on each side of the beneficial allele. By the independence of these events, LD vanishes when the beneficial allele fixes.

Looking at the pattern of LD at the end of a selective sweep, one might be tempted to conclude that there must be a hotspot of recombination at the selected site. This has been investigated by Reed and Tishkoff (2006), who indeed found out that hitchhiking may confound tests for recombination hotspots. However, only hitchhiking can reduce sequence diversity, which helps to make a clear distinction between genetic hitchhiking and recombination hotspots (McVean 2007).

In our study, we use the star-like approximation for the genealogy at the selected site to describe patterns of LD. This approximation is already implicit in the analysis of Maynard Smith and Haigh (1974) and it still inspires new methods for data analysis (*e.g.*, Nielsen *et al.* 2005). Our star-like approximation of the joint genealogy at the two neutral loci is a slight but crucial modification of the approach of McVean (2007). On the one hand, McVean does not describe splits in the wild-type background [see line (iv) in Figure 4] but implicitly accounts for these events. On the other hand, he ignores splits in the beneficial background that are shown in Figure 3. As a consequence, his star-like approximation becomes less accurate with increasing distance of both neutral loci. In addition, McVean's approximation is incompatible with the results on D obtained in Stephan *et al.* (2006). Like McVean we see a big effect of a finite sample size on patterns of LD. Since we use larger selection coefficients α, we could not reproduce his finding that neutral mutations that are more recent than the beneficial allele lead to a significant reduction in LD.

Generally, the star-like approximation gives a good approximation for $\widehat{\sigma_D^2}$ at the end of a selective sweep. It predicts correctly the increase of LD close to the selected site and the elimination of LD between both sides of the selective site. The slight underestimation of LD of the star-like approximation (see Figure 8) can also be explained: coalescence events during the selective phase lead to more complex scenarios than star-like genealogies at the selected site. In particular, these events are responsible for the fact that the star-like approximation creates a too long genealogy that then leads to an overestimation of the number of recombination events (*i.e.*, an underestimation of LD) under the star-like hypothesis. Even more complex genealogies appear if we take back-recombinations into the beneficial background into account (Barton 1998). Although such events have been shown to appear with low probability (Etheridge *et al.* 2006), the star-like approximation underestimates LD because back-recombinations can lead to common ancestry at the beginning of the selective sweep.

The star-like approximation was criticized and proposed to be replaced by the genealogy of a Yule process (Durrett and Schweinsberg 2004; Pfaffelhuber *et al.* 2006). The corresponding Yule approximation for the joint genealogy of two neutral loci using a Yule process was obtained by Pfaffelhuber and Studeny (2007). However, the star-like approximation is still useful. First, as shown by Durrett and Schweinsberg (2004) the star-like approximation for a single neutral locus gives correct results if $\log(\alpha)$ is large. Second, by the independence of all lines during the selective sweep, it allows for explicit calculations. In particular, using the star-like approximation, it is possible to obtain not only predictions for standardized LD or second moments of D—see (13) and (14)—but also higher moments of D.

Recently, the model of selective sweeps has been extended to the case of multiple origins of the beneficial allele—so-called *soft selective sweeps* (Hermisson and Pennings 2005). Such multiple origins may, for example, be due to recurrent mutation to the beneficial allele during its fixation. Together with new mutants at the selected site, new ancestral haplotypes are imported into the beneficial background. As a consequence, statistical tests based on haplotype structure, *i.e.*, LD, have most power to detect soft selective sweeps (Pennings and Hermisson 2006). Moreover, the coalescent at the selected locus was also derived (Pennings and Hermisson 2006): given that the frequency of the beneficial allele is x, an ancestral line escapes the sweep with rate $\theta_b(1-x)/(2x)$, where θ_b is the scaled mutation rate to the beneficial allele. Extending the analysis of genealogies to pairs of neutral loci close to the selective sweep, we believe that an analysis of LD under soft selective sweeps is feasible, shedding new light on the distinction between classical and soft selective sweeps.

We thank Joachim Hermisson for fruitful discussion and we are grateful to Pleuni Pennings and Nick Barton for helpful comments on our work. W.S. thanks the Deutsche Forschungsgemeinschaft for support (grant STE 325/7). P.P. thanks the BMBF for financial support via FRISYS (Kennzeichen 0313921).

LITERATURE CITED

Barton, N., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Beisswanger, S., W. Stephan and D. DeLorenzo, 2006 Evidence for a selctive sweep in the *wapl* region of *Drosophila melanogaster.* Genetics **172:** 265–274.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the number of haplotypes under an infinite sites model. Mol. Biol. Evol. **15:** 1788–1790.

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

Etheridge, A., P. Pfaffelhuber and A. Wakolbinger, 2006 An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. **15:** 685–729.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Hanchard, N., K. Rockett, C. Spencer, G. Coop, M. Pinder *et al.*, 2006 Screening for recently selected alleles by analysis of human haplotype similarity. Am. J. Hum. Genet. **78:** 153–159.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics **169:** 2335–2352.

Hudson, R., K. Bailey, D. Skarecky, J. Kwiatowski and F. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster.* Genetics **136:** 1329–1340.

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

Jensen, J. D., K. R. Thornton, C. D. Bustamante and C. F. Aquadro, 2007 On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in non-equilibrium populations. Genetics **176:** 2371–2379.

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

Kelly, J., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

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.

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

McVean, G. A., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175:** 1395–1406.

Nielsen, R., S. Williamson, Y. Kim, M. Hubisz, A. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566–1575.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63:** 229–238.

Pennings, P., and J. Hermisson, 2006 Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. **2:** e186.

Pfaffelhuber, P., and A. Studeny, 2007 Approximating genealogies for partially linked neutral loci under a selective sweep. J. Math. Biol. **55:** 299–330.

Pfaffelhuber, P., B. Haubold and A. Wakolbinger, 2006 Approximate genealogies under genetic hitchhiking. Genetics **174:** 1995–2008.

Reed, F. A., and S. A. Tishkoff, 2006 Positive selection can create false hotspots of recombination. Genetics **172:** 2011–2014.

Sabeti, P., D. R. J. Higgins, H. Levine, D. Richter, S. Schaffner *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Sabeti, P., S. Schaffner, B. Fry, J. Lohmuller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science **312:** 1614–1620.

Song, Y. S., and J. S. Song, 2007 Analytic computation of the expectation of the linkage disequilibrium coefficient r^2. Theor. Popul. Biol. **71:** 49–60.

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. Popul. Biol. **41:** 237–254.

Stephan, W., Y. Song and C. H. Langley, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics **172:** 2647–2663.

Strobeck, C., and K. Morgan, 1978 The effect of intragenic recombination on the number of alleles in a finite population. Genetics **88:** 829–844.

Toomajian, C., R. Ajioka, L. Jorde, J. Kushner and M. Kreitman, 2003 A method for detecting recent selection in the human genome from allele age estimates. Genetics **165:** 287–297.

Wang, E., G. Komada, P. Baldi and R. Moyzis, 2006 Global landscape of recent inferred Darwinian selection for *Homo sapiens.* Proc. Natl. Acad. Sci. USA **103:** 135–140.

Communicating editor: R. Nielsen

APPENDIX

Two relationships are important for the derivation of our results on LD: first, a genealogical interpretation of σ_D^2 and second, the difference between measures of LD in the population and in a sample. Both are not restricted to the analysis of selective sweeps.

Linkage disequilibrium and pairwise heterozygosities: Consider two diallelic loci (one called the *L*-locus and the other the *R*-locus) in a population with (random) allele frequencies $q_L, q_R, q_\ell, q_r, q_{LR}, \ldots$. Using (2) we write

$$\mathcal{X} := \mathbb{E}[q_L(1-q_L)q_R(1-q_R)], \quad \mathcal{Y} := \mathbb{E}[D(1-2q_L)(1-2q_R)], \quad \mathcal{Z} := \mathbb{E}[D^2] \tag{A1}$$

such that $\sigma_D^2 = \mathcal{Z}/\mathcal{X}$.

To interpret $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ as pairwise heterozygosities, we need three quantities. Consider two pairs of one allele at the *L*- and one allele at the *R*-locus each. Each pair of alleles is either linked or unlinked, *i.e.*, the allele at the *L*-locus lies on the same chromosome as the allele at the *R*-locus or they are located on different chromosomes from different individuals. Two pairs of alleles might both be linked, or one is linked while the other is unlinked, or both are unlinked. Denote by *f* the probability that two linked pairs of alleles at the *L*- and the *R*-locus are heterozygous at both the *L*- and the *R*-locus. Moreover, *g* denotes the probability that two pairs of alleles, where the first pair is linked and the second pair is unlinked, are heterozygous when picked randomly from the population. Third, *h* denotes the probability that two pairs of unlinked alleles are heterozygous.

For the allelic frequencies, some relationships hold; *e.g.*, $q_{Lr} = q_L - q_{LR}$, $q_{\ell R} = q_R - q_{LR}$, $q_{\ell r} = 1 - q_L - q_R + q_{LR}$, and $D = q_{\ell r} - q_\ell q_r$. These allow us to write

$$\begin{aligned}
\mathcal{X} &= \frac{1}{4}h \\
\mathcal{Y} &= \mathbb{E}[D(q_\ell q_r + q_L q_R - q_\ell q_R - q_L q_r)] \\
&= \mathbb{E}[q_{LR} q_\ell q_r + q_{\ell r} q_L q_R + q_{Lr} q_\ell q_r + q_{\ell R} q_L q_r - 4 q_L q_\ell q_R q_r] \\
&= g - h, \\
\mathcal{Z} &= \frac{1}{2}(\mathbb{E}[(q_{LR} - q_L q_R)(q_{\ell r} - q_\ell q_r)] + \mathbb{E}[(q_{Lr} - q_L q_r)(q_{\ell R} - q_\ell q_R)]) \\
&= \frac{1}{4}(\mathbb{E}[2 q_{LR} q_{\ell r} + 2 q_{\ell R} q_{R\ell}] + 4\mathbb{E}[q_L q_R q_\ell q_r] \\
&\qquad - 2\mathbb{E}[q_{LR} q_\ell q_r + q_{Lr} q_\ell q_R + q_{\ell R} q_L q_r + q_{\ell r} q_L q_R]) \\
&= \frac{1}{4}(f - 2g + h), \qquad \text{(A2)}
\end{aligned}$$

such that

$$\begin{pmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathcal{Z} \end{pmatrix} = \frac{1}{4} \underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 0 & 4 & -4 \\ 1 & -2 & 1 \end{pmatrix}}_{=:B} \begin{pmatrix} f \\ g \\ h \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} f \\ g \\ h \end{pmatrix} = \underbrace{\begin{pmatrix} 4 & 2 & 4 \\ 4 & 1 & 0 \\ 4 & 0 & 0 \end{pmatrix}}_{=B^{-1}} \begin{pmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathcal{Z} \end{pmatrix}. \qquad \text{(A3)}$$

Note that Strobeck and Morgan (1978) and Hudson (1985) use pairwise *homo*zygosities to derive $\mathcal{Z}$, while we use pairwise *hetero*zygosities.

Population and sample measures of linkage disequilibrium: Usually, we are given data from a finite sample and want to compute the amount of LD. Using the sample frequencies $\hat{q}_L, \hat{q}_R, \ldots$, we define the sample quantities $\widehat{\mathcal{X}}, \widehat{\mathcal{Y}}$, and $\widehat{\mathcal{Z}}$ as in (A1). By a calculation analogous to (A2)$(\widehat{\mathcal{X}}, \widehat{\mathcal{Y}}, \widehat{\mathcal{Z}})^\top = B \cdot (\hat{f}, \hat{g}, \hat{h})^\top$, where $\hat{f}, \hat{g}$, and $\hat{h}$ are the corresponding pairwise heterozygosities in the sample. Between *f*, *g*, and *h* and $\hat{f}, \hat{g}$, and $\hat{h}$, we have the relationships

$$\begin{aligned}
\hat{f} &= \left(1 - \frac{1}{n}\right) f, \\
\hat{g} &= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) g + \left(1 - \frac{1}{n}\right)\frac{1}{n} f, \\
\hat{h} &= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right) h + \left(1 - \frac{1}{n}\right)\frac{4}{n}\left(1 - \frac{2}{n}\right) g + \left(1 - \frac{1}{n}\right)\frac{2}{n}\frac{1}{n} f.
\end{aligned}$$

For example, two linked pairs of one allele at the *L*- and one allele at the *R*-locus each, taken at random (with replacement) from a sample are heterozygous, if we did not pick the same individual twice and the resulting two different lines are heterozygous at both loci.

Let *I* be the identity matrix and set

$$E = I + \frac{1}{n}\begin{pmatrix} -1 & 0 & 0 \\ 1 & -3 & 0 \\ 0 & 4 & -6 \end{pmatrix} + \frac{1}{n^2}\begin{pmatrix} 0 & 0 & 0 \\ -1 & 2 & 0 \\ 2 & -12 & 11 \end{pmatrix} + \frac{1}{n^3}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & 8 & -6 \end{pmatrix}$$

such that $(\hat{f}, \hat{g}, \hat{h})^\top = E \cdot (f, g, h)^\top$. Assuming that the sample size is large enough such that terms of order $1/n^2$ can be ignored, the above reasoning and some matrix algebra gives

$$\begin{pmatrix} \widehat{\mathcal{X}} \\ \widehat{\mathcal{Y}} \\ \widehat{\mathcal{Z}} \end{pmatrix} = B \begin{pmatrix} \hat{f} \\ \hat{g} \\ \hat{h} \end{pmatrix} = B \cdot E \cdot B^{-1} \begin{pmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathcal{Z} \end{pmatrix} \approx \left[I + \frac{1}{n}\begin{pmatrix} -2 & 1 & 0 \\ 0 & -5 & 4 \\ 1 & 1 & -3 \end{pmatrix} \right] \begin{pmatrix} \mathcal{X} \\ \mathcal{Y} \\ \mathcal{Z} \end{pmatrix}. \qquad \text{(A4)}$$