Chapter 4

Two-Locus Dynamics

The sequencing of entire genomes is becoming commonplace and multiple complete genomes are available for some species. In a few years, genomic samples will be the norm. Population genetics will have to respond with new theories that address the evolution of regions of the genome with many loci and stretches of noncoding DNA. Such theories will be similar in character to those for single loci. They will have descriptive aspects and dynamics. The mathematics, however, will be much more difficult; the sort of analytic approaches that are so important in one-locus theories will not be possible. Computer simulations will play a central role in investigations of genomic evolution.

The framework for studying the evolution of regions of the genome is already in place. The association of alleles on chromosomes is described by a quantity called the linkage disequilibrium, whose statistical properties and dynamics under random mating are well understood. The basic equations of multi-locus genetic drift and selection have been known for decades, although their analysis has been impeded by their extraordinary complexity. Already, genomic studies have suggested an exciting new form of evolution in which selection at one locus affects the dynamics of linked loci in a process called hitchhiking. This chapter will touch on each of these areas.

4.1 Linkage disequilibrium

Linkage disequilibrium is used to describe the associations of alleles on chromosomes. It appears quite naturally when describing the dynamics of gametes with random mating and recombination. The simplest model capable of showing the effects of recombination is of a diploid species with two linked loci, each with two segregating alleles. The left-hand side of Figure 4.1 illustrates the position of the two loci on the chromosome. The probability that a recombinant gamete is produced at meiosis is denoted by r, which is often called the recombination rate. (The genetic or map distance between the loci is always greater than r because it is the average number of recombinational events rather than the probability of producing a recombinant offspring.)

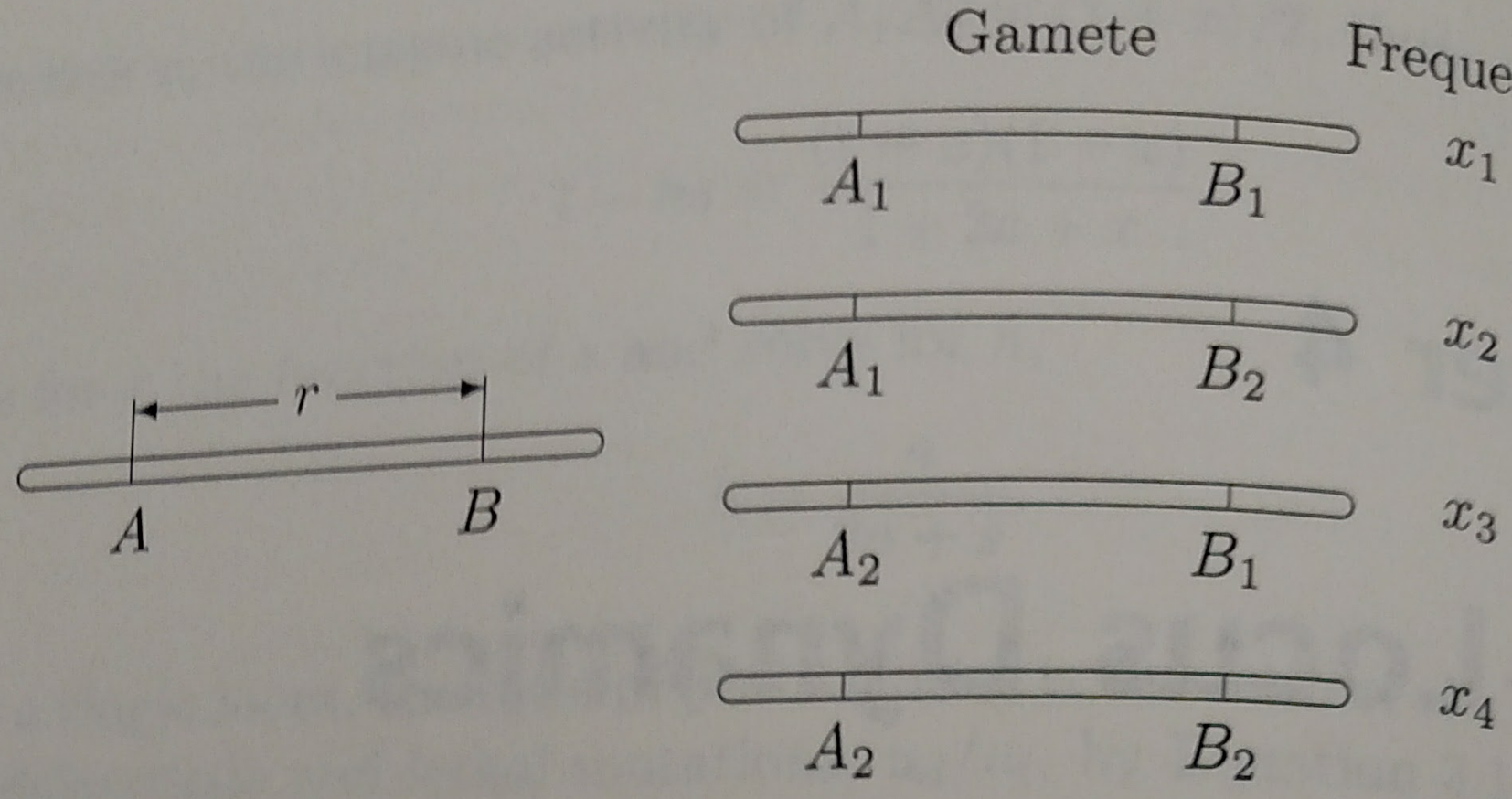

Figure 4.1: The chromosome on the left shows the position of the A and B loci. The right side illustrates the four possible gametes with their frequencies.

The right-hand side of Figure 4.1 shows that there are four gametes in the population, A_1B_1, A_1B_2, A_2B_1, and A_2B_2 with frequencies x_1, x_2, x_3, and x_4, respectively. The frequency of the A_1 allele, as a function of the gamete frequencies, is $p_1 = x_1 + x_2$. Similarly, the allele frequency of B_1 is $p_2 = x_1 + x_3$.

Recombination changes the frequencies of these gametes in a very simple way. For example, the frequency of the A_1B_1 gamete after a round of random mating, x_1', is simply

$$x_1' = (1 - r)x_1 + rp_1p_2\,. \tag{4.1}$$

This expression is best understood as a statement about the probability of choosing an A_1B_1 gamete from the population. A randomly chosen gamete will have had one of two possible histories: Either it will be a recombinant gamete (this occurs with probability r) or it won't be (this occurs with probability $1-r$). If it is not a recombinant, then the probability that it is an A_1B_1 gamete is x_1. Thus, the probability that the chosen gamete is an unrecombined A_1B_1 gamete is $(1-r)x_1$, which is the first term on the right side of Equation 4.1. If the gamete is a recombinant, then the probability that it is A_1B_1 is the probability that the A locus is A_1, which is just the frequency of A_1, p_1, times the probability that the B locus is B_1, p_2. The probability of being a recombinant gamete and being A_1B_1 is rp_1p_2, which is the rightmost term of Equation 4.1. The allele frequencies can be multiplied because the effect of a recombination is to choose the allele at the A and B loci independently.

Problem 4.1 *Derive the three equations for the frequencies of the A_1B_2, A_2B_1, and A_2B_2 gametes after a round of random mating.*

The change in the frequency of the A_1B_1 gamete in a single generation of random mating is, from Equation 4.1,

$$\Delta_r x_1 = -r(x_1 - p_1p_2). \tag{4.2}$$

The coefficient of r provides a definition for the linkage disequilibrium,

$$D = x_1 - p_1p_2,$$

which is a measure of the difference between the frequency of the A_1B_1 gamete, x_1, and the expected frequency if alleles associated randomly on chromosomes, p_1p_2. (If there were no tendency for the A_1 allele to be associated with the B_1 allele, the probability of choosing an A_1B_1 allele from the population would be the product of the frequencies of the A_1 and B_1 alleles.) The change in x_1, written as a function of D, is

$$\Delta_r x_1 = -rD.$$

The equilibrium gamete frequency is obtained by solving $\Delta_r x_1 = D = 0$,

$$\hat{x}_1 = p_1p_2.$$

From this, we conclude that recombination removes associations between alleles on chromosomes. The time scale of change of gamete frequencies due to recombination is roughly the reciprocal of the recombination rate.

The frequency of the A_1B_1 gamete may be written

$$x_1 = p_1p_2 + D,$$

which emphasizes that the departure of the gamete frequency from its equilibrium value is determined by D. The linkage disequilibrium may also be written in the more conventional form

$$D = x_1x_4 - x_2x_3, \tag{4.3}$$

which leads to the following new expressions for the gamete frequencies:

Gamete :	A_1B_1	A_1B_2	A_2B_1	A_2B_2
Frequency :	x_1	x_2	x_3	x_4
Frequency :	$p_1p_2 + D$	$p_1q_2 - D$	$q_1p_2 - D$	$q_1q_2 + D$

Problem 4.2 *Show that the gamete frequencies as a function of D are correct in the above table. Next, show that the definition of D in Equation 4.3 is consistent with the expressions for gamete frequencies by substituting $p_1p_2 + D$ for x_1 and the equivalent expressions for the other gametes into the right side of Equation 4.3, proceeding with a frenzy of cancellations, and ending with a lone D.*

The A_1B_1 and A_2B_2 gametes are often called coupling gametes because the same subscript is used for both alleles. The A_1B_2 and A_2B_1 gametes are called repulsion gametes. Linkage disequilibrium may be thought of as a measure of the excess of coupling over repulsion gametes. When D is positive, there are more coupling gametes than expected at equilibrium; when negative, there are more repulsion gametes than expected.

The value of D after a round of random mating may be obtained directly from Equation 4.1 by using $x_1 = p_1p_2 + D$,

$$p_1'p_2' + D' = (1 - r)(p_1p_2 + D) + rp_1p_2.$$

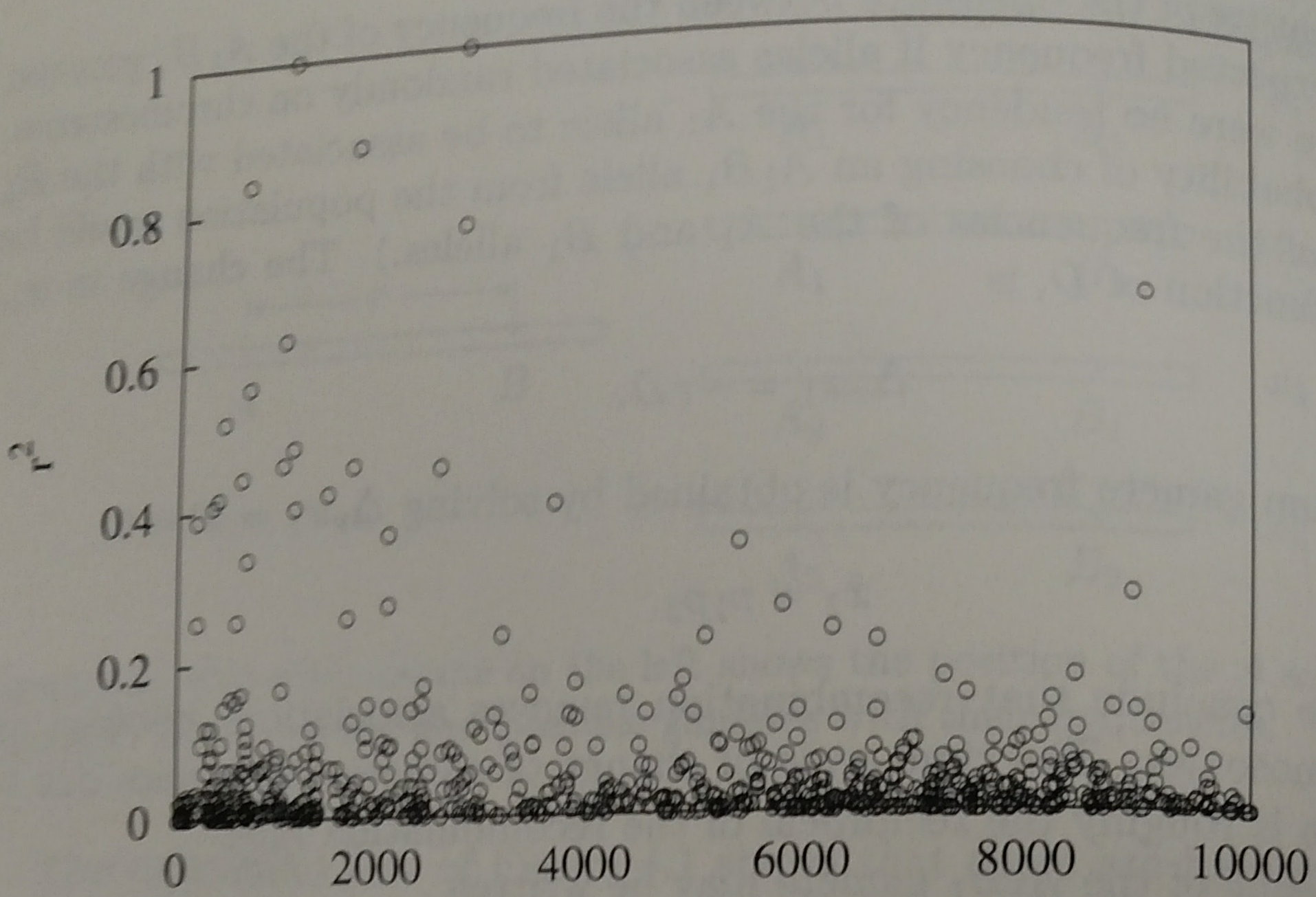

Figure 4.2: The linkage disequilibrium, as measured by r^2, between pairs of restriction sites in the *Delta* gene region of *Drosophila melanogaster*. Distance is measured in units of base pairs. The data are from Long et al. (1998).

A few quick cancellations yield

$$D' = (1 - r)D.$$

Some of the cancellations use the Hardy-Weinberg truism that allele frequencies don't change with random mating. We would be in trouble if the addition of loci affected the Hardy-Weinberg law for single loci!

The change in D in a single generation is

$$\Delta_r D = -rD,$$

which depends on the gamete frequencies only through their contributions to D. Finally,

$$D_t = (1 - r)^t D_0,$$

showing, once again, that the ultimate state of the population is $D = 0$. Note that with free recombination ($r = 1/2$) the linkage disequilibrium does not disappear in a single generation. If you find this startling, follow D for a couple of generations in a population initiated with $x_1 = x_4 = 1/2$ and $r = 1/2$.

In natural populations, the reduction in the magnitude of linkage disequilibrium by recombination is opposed by several evolutionary forces that may increase $|D|$. Natural selection will increase D if selection favors coupling gametes over repulsion gametes or decrease D if repulsion gametes are favored. Migration may increase the absolute value of D if the allele frequencies of the immigrants differ from those of the resident population. Finally, genetic drift

can lead to changes in D due to random sampling. As the efficacy of recombination increases with r, we would expect to find tightly linked loci farther from linkage equilibrium than loosely linked loci.

Linkage disequilibrium has the unfortunate property that its maximum absolute value is very sensitive to allele frequencies,

$$\max(|D|) = \min(p_1 q_2,\, q_1 p_2)$$

A measure of association that is much less sensitive to allele frequencies is

$$r^2 = \frac{D^2}{p_1 q_1 p_2 q_2}.$$

(The r here has nothing to do with the recombination rate.) r^2 may be interpreted as the square of the correlation coefficient between alleles at two loci as described on page 196. Figure 4.2 illustrates the relationship between r^2 and the distance between restriction site polymorphisms in the *Delta* gene region of *Drosophila melanogaster* using data from Long et al. (1998). All pairs of polymorphic sites within ten thousand base pairs (kb) of each other are included in the figure. As is evident, large values of r^2 are more common for pairs of sites within 5 kb of each other than between pairs of sites that are more distant. This is often expressed by saying that the scale of disequilibrium in this region is 5 kb. Note, however, that for the vast majority of pairs of sites r^2 is not very large, certainly not significantly greater than zero. Moreover, the sampling properties of r^2 are such that many of the larger values of r^2 are not significantly different from zero either. Occasionally, significant linkage disequilibrium is detected between loci that are farther apart than 5 kb nucleotides. This is often a sign that there is epistasis in fitness between these loci.

4.2 Two-locus selection

One of the most significant observations in population genetics over the past decade is the correlation between the rate of recombination in a region of a chromosome and the silent heterozygosity in that region. Figure 4.3 illustrates the relationship on the X chromosome of *Drosophila melanogaster* as described by Dave Begun and Chip Aquadro in 1992. The rate of recombination of a *melanogaster* chromosome per unit physical length is lower in the vicinity of the telomeres and the centromere than in other regions of the chromosome. It is as if the genetic map is compressed in certain regions and expanded in others. The two left-most points in the figure represent loci very close to the telomore where recombination is exceedingly low. These loci have no observed silent variation. Two explanations for this correlation spring to mind. The first and most obvious is that mutation rates are correlated with recombination rates. Less recombination means less mutation, which, in turn, means less variation. There is a simple check for this hypothesis: Are rates of substitution lower in regions of low recombination? Begun and Aquadro showed that they are not.

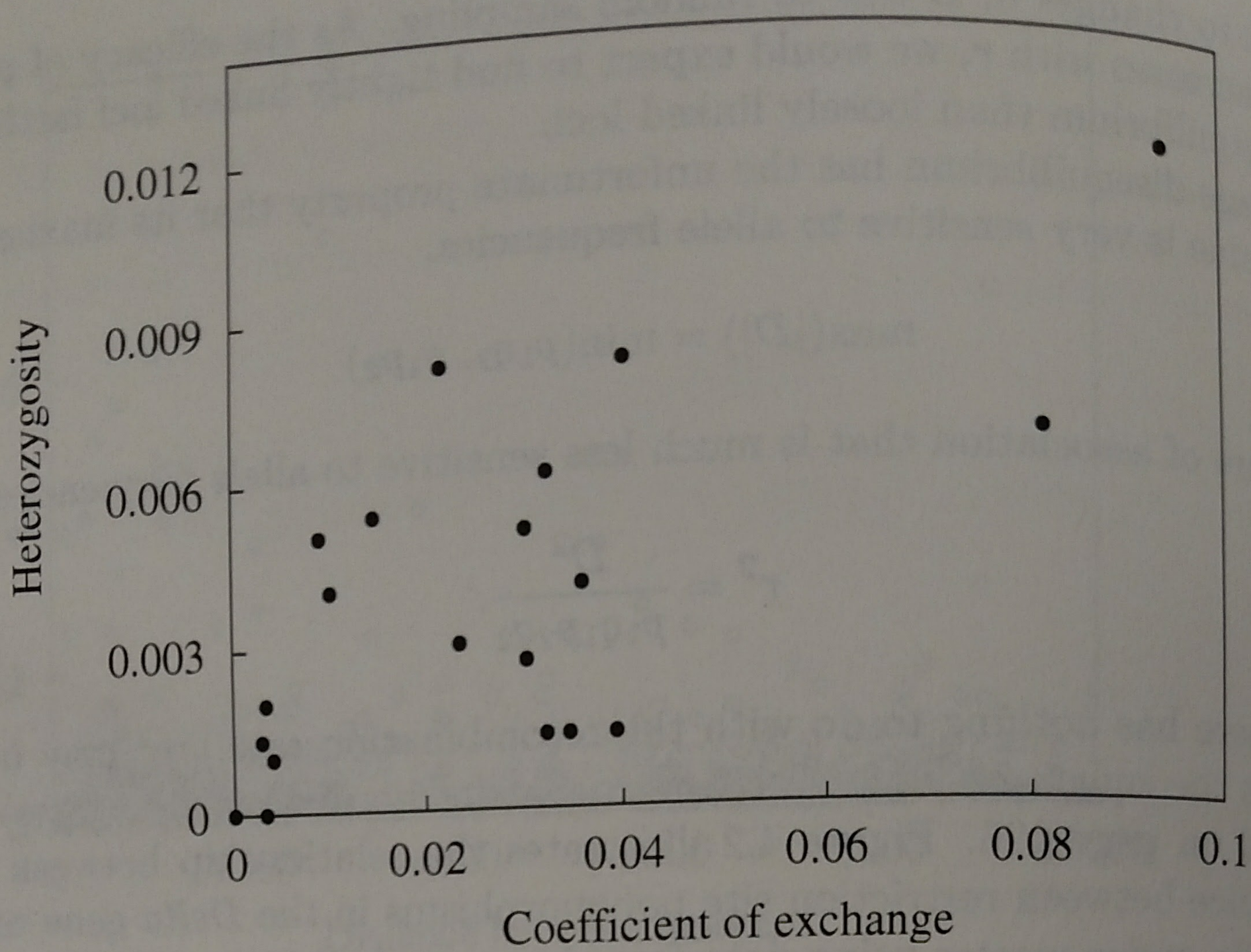

Figure 4.3: The observed silent heterozygosity on the X chromosome of *Drosophila melanogaster* as a function of the local rate of recombination. The data are from Begun and Aquadro (1992).

The rates of silent substitution in regions of low recombination are about the same as those in regions of high recombination.

The second explanation involves a process called hitchhiking. Consider the events associated with the appearance in the population of a new advantageous mutation that sweeps through the population to fixation. When the mutation first appears in the population, it sits on a single chromosome. As it increases in frequency, alleles at loci that are closely linked to it on that chromosome will increase as well. The increase in the frequency of the closely linked alleles is called hitchhiking. Alleles will continue to hitchhike until a sufficient amount of recombination occurs to, as it were, release them from the pull of the selected mutation. Very tightly linked alleles will be carried to fixation (or near fixation) by the selected allele. As a result, the genetic variation in the immediate vicinity of the selected mutation will be reduced or eliminated. In regions of low recombination, the reduction in variation will extend over a larger portion of the genome than in regions of high recombination. If selected sweeps are spread uniformly throughout the genome, regions of low recombination should have less genetic variation than regions of high recombination. This is the favored explanation for the pattern seen in Figure 4.3.

The reduction in genetic variation due to hitchhiking is determined by at least two parameters, the selection coefficient at the selected locus and the rate of recombination between the selected locus and the linked locus, which is usually assumed to be a neutral locus. The task before us is to find how these two parameters interact to reduce variation and to judge whether hitchhiking is a plausible explanation for the reduction in variation in regions of low re-

combination. Our approach will be a straightforward extension of that of the previous chapter: assign fitnesses to genotypes and calculate the changes in gamete frequencies in a single generation. As in the previous section, the model will focus on two di-allelic loci with gametes A_1B_1, A_1B_2, A_2B_1 and A_2B_2 at frequencies x_1, x_2, x_3 and x_4. The A locus will be the selected locus and the B locus will be the neutral locus, although these properties will not be invoked until we describe a more general model.

The following table gives the frequencies of all ten genotypes in the population after random mating and viability selection.

	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1	$x_1^2w_{11}$	$2x_1x_2w_{12}$	$2x_1x_3w_{13}$	$2x_1x_4w_{14}$
A_1B_2	·	$x_2^2w_{22}$	$2x_2x_3w_{23}$	$2x_2x_4w_{24}$
A_2B_1	·	·	$x_3^2w_{33}$	$2x_3x_4w_{34}$
A_2B_2	·	·	·	$x_4^2w_{44}$

Consider, for example, the second entry in the first row, which is for the A_1B_1/A_1B_2 genotype. If the gamete frequencies, x_i, are those just before random mating (just after recombination), then the frequency of unions of A_1B_1 and A_1B_2 gametes when random mating occurs is $2x_1x_2$. The viability of this genotype is called w_{12} and so the frequency of the genotype after mating is proportional to $2x_1x_2w_{12}$, as given in the table. Note that we have not yet normalized the frequencies by the mean fitness of the population.

Using the first row of the table, the frequency of the A_1B_1 gamete, after random mating, selection and recombination, is found to be

$$x_1' = \frac{x_1^2w_{11} + x_1x_2w_{12} + x_1x_3w_{13} + x_1x_4(1-r)w_{14} + rx_2x_3w_{23}}{\bar{w}}.$$

The first three terms in the numerator are analogous to the terms in the numerator for single locus selection as described on page 62. The new features are in the last two terms. For the fourth term, note that the probability that an A_1B_1/A_2B_2 genotype produces an A_1B_1 gamete is $(1-r)/2$. Thus, the frequency of A_1B_1 gametes from these genotypes is proportional to $x_1x_4(1-r)w_{14}$. Similarly, A_1B_2/A_2B_1 gametes can produce A_1B_1 gametes through recombination, which leads to the fifth term. A little rearranging lets us write

$$x_1' = \frac{x_1\bar{w}_1 - w_{14}rD}{\bar{w}},$$

where we have assumed that $w_{23} = w_{14}$ and have defined the marginal fitness of A_1B_1 to be

$$\bar{w}_1 = x_1w_{11} + x_2w_{12} + x_3w_{13} + x_4w_{14}.$$

The marginal fitness of a gamete is the average fitness of all genotypes containing that gamete just before selection occurs. It may be written as

$$\bar{w}_i = \sum_{j=1}^{4} x_jw_{ij}$$

if we agree that $w_{ij} = w_{ji}$, when $j > i$. Finally, the mean fitness of the population is

$$\bar{w} = \sum_{i=1}^{4} x_i \bar{w}_i.$$

The assumption that $w_{23} = w_{14}$ means that there are no cis-trans effects for these two loci. Fitness depends only on the alleles present in a genotype, not on their arrangement on the chromosome.

The change in the frequencies of the four gametes in a single generation are

$$\Delta x_1 = \frac{x_1(\bar{w}_1 - \bar{w}) - rw_{14}D}{\bar{w}}$$

$$\Delta x_2 = \frac{x_2(\bar{w}_2 - \bar{w}) + rw_{14}D}{\bar{w}}$$

$$\Delta x_3 = \frac{x_3(\bar{w}_3 - \bar{w}) + rw_{14}D}{\bar{w}}$$

$$\Delta x_4 = \frac{x_4(\bar{w}_4 - \bar{w}) - rw_{14}D}{\bar{w}}.$$

As

$$\sum_{i=1}^{4} x_i = 1,$$

only three of the four equations are required to study the dynamics. Usually the fourth equation is dropped. The equations show that the change in the frequency of a gamete is approximately

$$\Delta x_i \approx \Delta_s x_i + \Delta_r x_i,$$

where the conflict between the directions of selection and recombination are apparent.

In the 1970s population geneticists studied the dynamics of this model in great detail. Many strange features emerged, such as multiple stable equilibria, but little of this has been applicable to our current interest in molecular evolution and polymorphism. However, one application did turn out to be of considerable importance. In a classic paper written in 1974, John Maynard Smith and John Haigh used this model to describe the effects of direction selection at one-locus on the variation at a linked neutral locus. Although not the first paper on hitchhiking, this was the first paper to bring hitchhiking to bear on the Great Obsession. The fitnesses needed to study hitchhiking come from the following table:

	A_1B_1	A_1B_2	A_2B_1	A_2B_2
A_1B_1	1	1	$1-hs$	$1-hs$
A_1B_2	.	1	$1-hs$	$1-hs$
A_2B_1	.	.	$1-s$	$1-s$
A_2B_2	.	.	.	$1-s$

Note that only the A locus determines the fitness of a genotype. The B locus is neutral. The marginal fitnesses for the four gametes are

$$\bar{w}_1 = \bar{w}_2 = 1 - q_1 hs$$
$$\bar{w}_3 = \bar{w}_4 = 1 - p_1 hs - q_1 s.$$

The mean fitness,

$$\bar{w} = 1 - 2p_1 q_1 hs - q_1^2 s,$$

is the same as for the one-locus model. In fact, the change in frequency of the A_1 allele,

$$\Delta x_1 + \Delta x_2 = \frac{p_1 q_1 s[p_1 h + q_1(1-h)]}{1 - 2p_1 q_1 hs - q_1^2 s},$$

is also the same as for the one-locus model. It couldn't be otherwise as the B locus is neutral and cannot affect the dynamics. More interesting is the change in the frequency of the B_1 allele:

$$\Delta x_1 + \Delta x_3 = \frac{Ds[p_1 h + q_1(1-h)]}{1 - 2p_1 q_1 hs - q_1^2 s}. \tag{4.4}$$

(When verifying this equation, make the initial substitutions $x_1 = p_1 p_2 + D$ and $x_3 = q_1 p_2 - D$.) The coefficient D shows that selection will change the frequency of the B_1 allele only if there is linkage disequilibrium ($D \neq 0$). The change is not due to selection directly on the B locus, but rather because of the non-random association of B alleles with A alleles. If $D > 0$, then the B_1 allele is associated with the A_1 allele and, because of this, will increase in frequency. The converse holds if $D < 0$. Thus, hitchhiking depends on linkage disequilibrium.

The most compelling question to ask of this model involves the reduction in variation at the neutral locus caused by a selective sweep at the selected locus. For example, imagine a population in which there are two equally frequent gametes, A_2B_1 and A_2B_2. The heterozygosity at the B locus in this case is one-half. If a new A_1 allele were to appear in a population of N individuals and then were to sweep through the population, what would be the heterozygosity at the B locus when the sweep was complete? The initial frequency of the four gametes could be:

A_1B_1	A_1B_2	A_2B_1	A_2B_2
$\frac{1}{2N}$	0	$\frac{1}{2} - \frac{1}{2N}$	$\frac{1}{2}$

(4.5)

Here we have assumed that the A_1 mutation happened to fall on an A_2B_2 gamete instead. In fact, because these two gametes are equally frequent, it is equally likely that the A_1 mutation will fall on either one of them.

To obtain the final heterozygosity, we need only iterate the equations until $p_1 \approx 1$ and then use the final value of p_2 for the heterozygosity, $2p_2q_2$. This cannot be done with mathematics as the equations are too complicated. (Maynard Smith and Haigh were able to do this because they used a haploid rather than a diploid model.) We could use some approximations to obtain the answer,

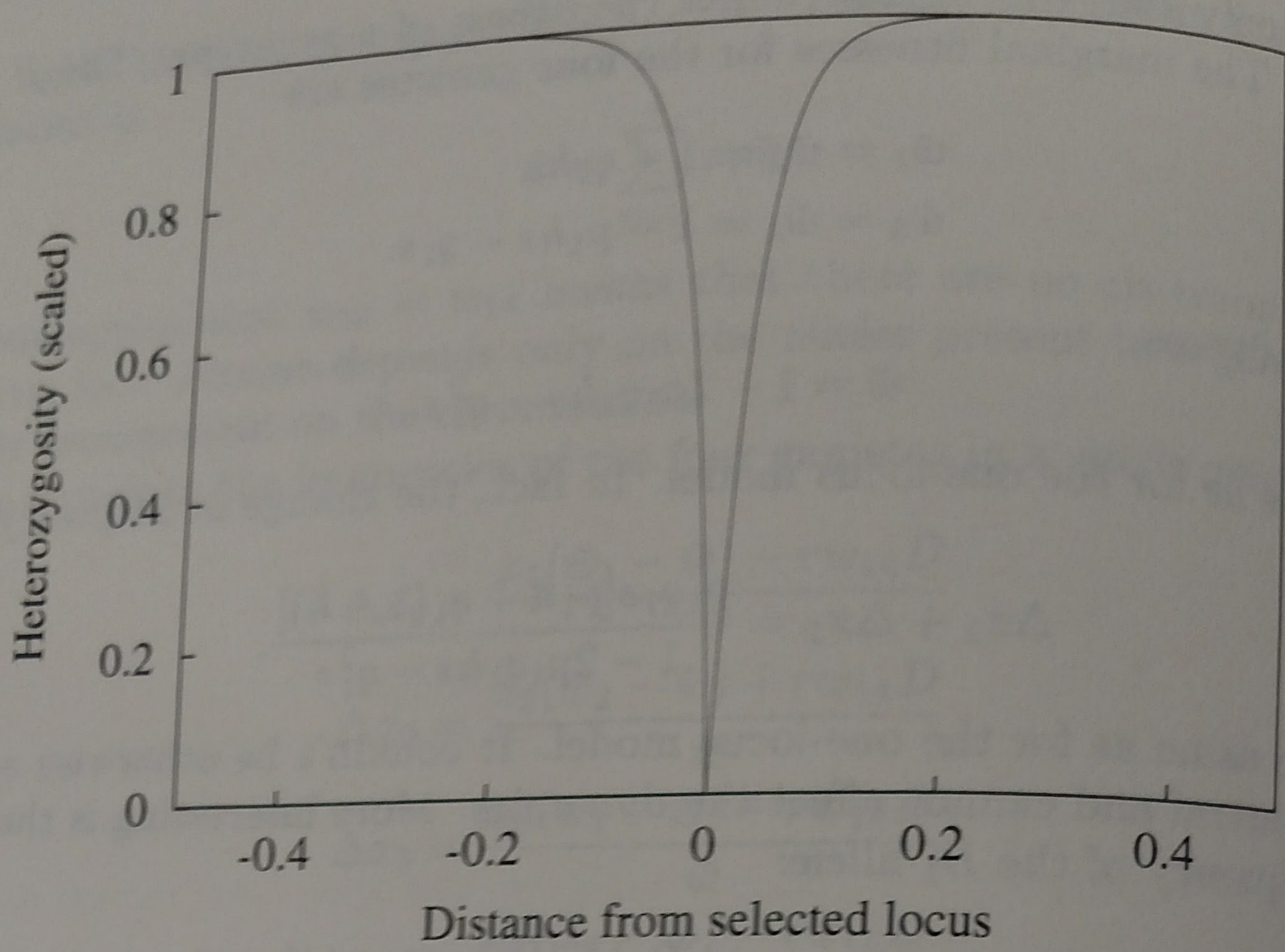

Figure 4.4: The ratio of the final to initial heterozygosity at a neutral locus as a function of the distance from the selected locus as measured by r/s. Negative values of r/s are left of the selected locus, positive values are to the right.

but the most expeditious route to the final answer is computer simulation. The data for Figure 4.4, which graphs the ratio of the final heterozygosity to the initial heterozygosity at the B locus for different values of r/s, were obtained by computer simulations. The simulations used $h = 1/2$, $s = 0.1$, and $N = 5000$ and were run until $p_1 > 0.9999$.

Problem 4.3 *Write a computer simulation to reproduce the data graphed in Figure 4.4. Try different values of s and N to see how much this changes the graph, always using r/s for the horizontal axis.*

The figure shows that hitchhiking will reduce the heterozygosity at the B locus if r/s is less than about 0.1. In other words, if $r/s \ll 0.1$, selection dominates recombination and hitchhiking removes much of the linked variation. If $r/s \gg 0.1$, recombination is the stronger force and there is little reduction in the levels of linked variation. If, as appears to be the case, the probability of a recombination between neighboring nucleotides is about 10^{-8}, and if, say, $s = 0.001$, then a selective sweep will lower the level of polymorphism in a stretch of about 10^4 nucleotides on either side of the selected nucleotide.

Problem 4.4 *The most storied two-locus model is called the symmetric viability*

model. The fitnesses are usually displayed as follows:

	B_1B_1	B_1B_2	B_2B_2
A_1A_1	a	b	a
A_1A_2	c	d	c
A_2A_2	a	b	a

If $a < b$, $c < d$ the fitness of a genotype increases with the number of heterozygous loci and there will be a stable polymorphism with both loci segregating. Use computer simulations to study this model. See if you can find parameters that give an equilibrium where $D \neq 0$. Increase r to see if the $D \neq 0$ equilibria disappear. Compare your findings to those in Lewontin and Kojima (1960).

4.3 Genetic draft

The description of hitchhiking in the previous section had a hidden random element that, when uncovered, leads to a new stochastic force in evolution. The new A_1 mutation that ultimately fixes in the population was initially placed on a chromosome with a B_1 allele. That choice was entirely arbitrary; it could just as well have been placed on a chromosome with a B_2 allele. In natural populations, this placement is a random event. If, at the generation when the A_1 mutation appears, there are p_2 A_2B_1 gametes and q_2 A_2B_2 gametes, then the probability that the A_1 allele first appears on a gamete with a B_1 allele is p_2. This is the case described in Equation 4.5. The following table gives the two possible placements, their probabilities, and the implied initial conditions:

Gamete	Probability	x_1	x_2	x_3	x_4	D
A_1B_1	p_2	$1/2N$	0	$p_2 - 1/2N$	q_2	$q_2/2N$
A_1B_2	q_2	0	$1/2N$	p_2	$q_2 - 1/2N$	$-p_2/2N$

Note that the initial value of D, when put into Equation 4.4, determines whether p_2 will increase or decrease. Figure 4.5 gives some example trajectories for the B_1 allele. In the upper two curves, the initial conditions match the first row in the table with the initial value of $p_2 = 0.3$; the lower two match the second row with the same initial value for p_2. The probability that the frequency of the B_1 allele evolves as in the upper curves is p_2. We could summarize this figure for the case $r = 0$ by saying that the B_1 allele fixes with probability 0.3 and is lost with probability 0.7. When $r/s > 0$, we would say that p_2 increases or decreases to some value with probabilities 0.3 and 0.7, respectively.

This is a different sort of stochastic force than we have encountered thus far. With both genetic drift and selection in a random environment, the allele frequencies bounce around at random each generation as the result of the stochastic force. With hitchhiking, only the initial conditions are random. Once the process begins, it is deterministic (unless some other stochastic force is added). Hitchhiking, when viewed as a stochastic force, is called genetic draft* because of some surprising similarities to genetic drift.

*Bill Gilliland is responsible for the name genetic draft.

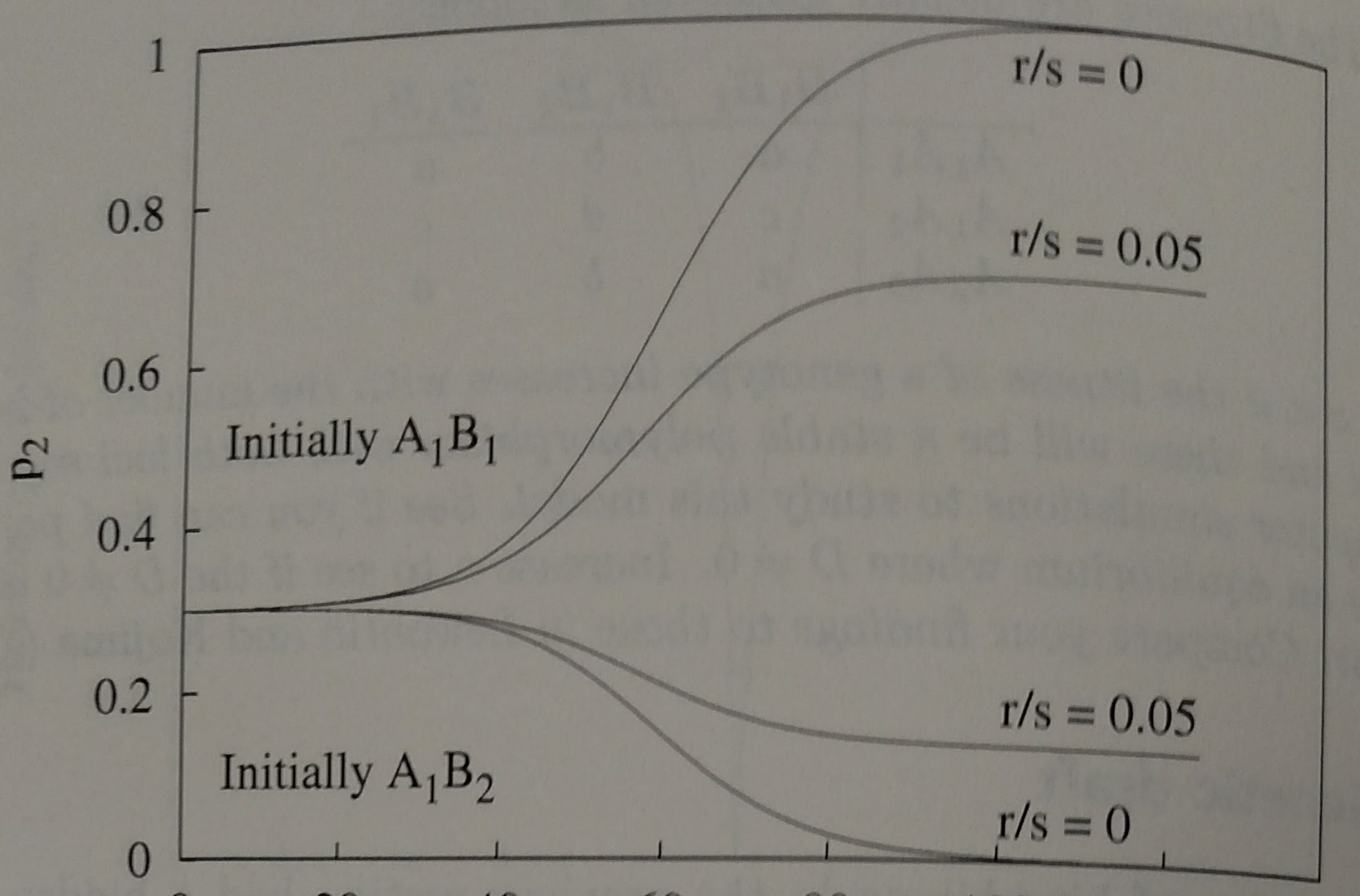

Figure 4.5: The frequency of the B_2 allele under different hitchhiking scenarios. For the upper two curves, the A_1 allele in initially linked to the B_1 allele; in the bottom two, it is linked to the B_2 allele. $s = 0.2$ for all trajectories.

The special case $r = 0$ is a good place to begin a study of genetic draft. The two consequences of hitchhiking on the B locus in this case may be summarized as follows:

$$p' = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } q, \end{cases}$$

where the subscript 2 for p and q have been suppressed. Examples of these two outcomes are given by the two outer curves in Figure 4.5. p' is a Bernoulli random variable as described on page 192. Its mean and variance are

$$E\{p'\} = p$$
$$\text{Var}\{p'\} = pq.$$

The mean allele frequency after a hitchhiking event is the same as the allele frequency before hitchhiking. The variance in p' should be familiar: It is the same as for genetic drift with $N_e = 1/2$. These two properties suggest that the stochastic dynamics of genetic draft might resemble, at least to some extent, those of genetic drift.

In any particular generation, three things can happen: No sweep is initiated and thus no change in the allele frequencies at the B locus; a sweep is initiated leading to the fixation of B_1; or a sweep is initiated leading to the loss of B_1:

$$p' = \begin{cases} p & \text{with probability } 1 - \rho \\ 1 & \text{with probability } \rho p \\ 0 & \text{with probability } \rho q, \end{cases}$$

where ρ is the probability that a hitchhiking event is initiated in a particular generation. The moments of p' are now

$$E\{p'\} = p$$
$$\mathrm{Var}\{p'\} = \rho pq.$$

These moments resemble those of genetic drift with $N_e = 1/2\rho$.

If hitchhiking events recur, the probability that a hitchhiking event is initiated in a particular generation, ρ, is also the rate of substitution. In order to model the effects of recurring hitchhiking, we need to make some simplifying assumptions. The first of these is that the time between substitutions at the A locus is much longer than the time required to complete one fixation. If the time required to complete a fixation is sufficiently small relative to the time between fixations, we can assume that fixations occur in a single generation without serious error. As a consequence, the change in the frequency of alleles at the B locus will also occur in a single generation. The model becomes particularly easy to study if the times of the hitchhiking events form a Poisson process, which was introduced on page 35. The Poisson process assumption makes the timing of successive hitchhiking events independent, which greatly simplifies certain calculations. All of these assumptions lead to a new model that is called the pseudohitchhiking model. The pseudo prefix is an acknowledgment that very complicated two-locus dynamics have been obliterated in the creation of a one-locus model for the B locus. Computer simulations have shown that the pseudohitchhiking model does approximate the dynamics of the full two-locus model rather well when the appropriate value of ρ is used.

When recombination is allowed, only a slight change need be made to the pseudohitchhiking model. The motivation comes from the observation that when a new advantageous mutation enters the population, it is initially linked to only a single copy of one of the neutral alleles. That one copy will increase in frequency until it is released by recombination. Meanwhile, the frequencies of all of the other alleles are reduced by the same factor because they are not hitchhiking. If the final frequency of the hitchhiking copy is y, then we can write the pseudohitchhiking model as

$$p' = \begin{cases} p & \text{with probability } 1-\rho \\ p(1-y)+y & \text{with probability } \rho p \\ p(1-y) & \text{with probability } \rho q. \end{cases}$$

The expression in the second line shows that the frequency of the B_1 allele will be y, the final frequency of the one copy that hitchhiked, plus $p(1-y)$, the frequency of all other copies of the B_1 alleles that did not hitchhike. If the B_2 allele hitchhiked, then the third line shows that the frequency of the B_1 allele will be reduced by the fraction $1-y$. In this case one copy of the B_2 allele had its frequency increase from $1/2N$ to y at the expense of the B_1 allele and other copies of the B_2 allele. It cannot be emphasized enough that these dynamics are only approximations that experience has shown to be fairly faithful to the

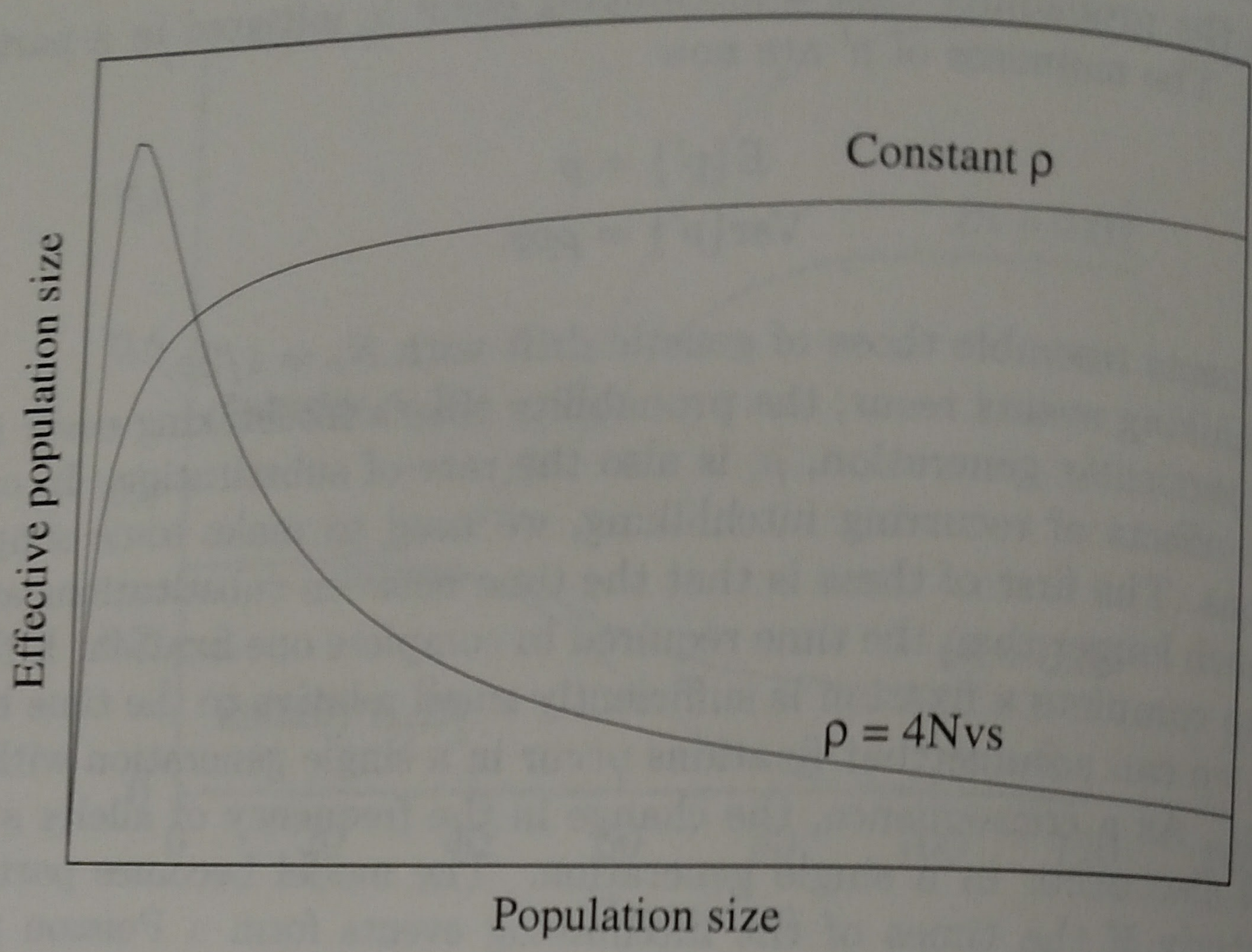

Figure 4.6: The relationship between the population size and the effective population size under genetic draft.

full two-locus dynamics. Once again, the mean of p' is p, but now the variance is

$$\mathrm{Var}\{p'\} = \rho y^2 pq.$$

The value of y is determined by the strength of selection at the selected locus, the rate of recombination between the neutral and the selected locus, and the initial conditions. In more realistic models that incorporate genetic drift, y will be a random variable reflecting the random fluctuations that occur in the early stages of the hitchhiking process. In this case the variance in p' is just the mean of the previous expression,

$$\mathrm{Var}\{p'\} = \rho E\{y^2\} pq.$$

In a population with both genetic drift and genetic draft, the variance in p' will be the sum of the variances due to genetic drift and genetic draft,

$$\mathrm{Var}\{p'\} = pq\left(\rho E\{y^2\} + \frac{1}{2N}\right).$$

The variance effective size of the population is

$$N_e = \frac{N}{1 + 2N\rho E\{y^2\}}. \tag{4.6}$$

Figure 4.6 illustrates the intriguing relationships between N_e and N for two models of the rate of substitution. If ρ is independent of population size, then

the effective size of the population increases with N, eventually reaching an asymptote at

$$\lim_{N\to\infty} N_e = \frac{1}{2\rho E\{y^2\}}. \tag{4.7}$$

This curve reflects the growing relative importance of genetic drift as genetic drift becomes weak with increasing N. Draft dominates drift when $\rho E\{y^2\} > 1/(2N)$.

On page 95 we saw that $\rho = 2Nus$ is commonly used for the rate of substitution of advantageous mutations. In this case, Figure 4.6 shows that the effective size of the population eventually decreases with increasing population size. Under this model, the rate of hitchhiking increases with N and, as a consequence, the effective size ultimately decreases with N.

Genetic draft is like genetic drift in that it removes genetic variation from the population. The rate of reduction, which was derived using only $\mathrm{Var}\{p'\}$ on page 52, is

$$\Delta H = \frac{1}{2N_e} H.$$

If mutation restores variation at rate

$$\Delta_u H = 2u(1-H),$$

(see Equation 2.11), then the equilibrium heterozygosity is

$$\hat{H} = \frac{4N_e u}{1+4N_e u}.$$

If we use Equation 4.6 for N_e, we get

$$\hat{H} = \frac{4Nu}{1+2N\rho E\{y^2\}+4Nu}.$$

The relationship between $\hat{H}$ and N mirrors that between N_e and N. If ρ is constant, the heterozygosity steadily increases to the asymptotic value

$$\frac{2u}{\rho E\{y^2\}+2u}.$$

Once N is sufficiently large, genetic draft dominates and genetic variation becomes insensitive to population size. This suggests a solution to the neutral theory's problem that levels of variation in natural populations are remarkably similar across species in spite of great differences in population sizes. In fact, the inability to find strong correlations between population size and molecular polymorphism and divergence argues that genetic drift is not an important force in natural populations. Genetic draft, as it does not suffer from this dependency, may well prove to be the most important stochastic force in natural populations.

If genetic draft is to be considered an important force, we need to show that it could account for observed levels of silent variation. We will do this for an

infinite population ($N = \infty$) in order to remove all effects of genetic drift. A measure of genetic variation that is particularly well suited for this task is π. On page 46 we saw that $\pi = 2u\bar{t}$, where $\bar{t}$ is the mean number of generations back to the common ancestor of a pair of alleles drawn at random from the population. In order for a pair of alleles to have a common ancestor in a particular generation in the past, a hitchhiking event must have occurred in that generation and both alleles must be descended from the single copy of the allele at the B locus that was initially linked to the A_1 allele. For a given value of y, this is just ρy^2. As y is a random variable, the mean probability is $\rho E\{y^2\}$. The mean time back to the common ancestor is the reciprocal of this coalescence probability, which gives

$$\pi = \frac{2u}{\rho E\{y^2\}}. \tag{4.8}$$

(You might have guessed this answer from the drift result, $\pi = 4N_e u$, and Equation 4.7.) Consider a typical block of 10,000 bases on an autosome in *Drosophila*. The recombination rate between any pair of nucleotides within this block is less than about $r = 10^4 \times 10^{-8} = 0.0001$. A sweep anywhere within the block with a selection coefficient of $s = 0.001$ ($r/s < 0.1$) will cause the entire block to become nearly homozygous, which means that $E\{y^2\} \approx 1$. If we assume that such substitutions occur at a rate $\rho = 10^{-7}$ within the block and that $u = 10^{-9}$, then, by Equation 4.8, π should be about 0.02, which is close to the observed value in *Drosophila simulans*. The total rate of silent substitution for this block of ten thousand nucleotides is $10^4 \times 10^{-9} = 10^{-5}$. Thus, we require that about one out of every 100 substitutions is strongly selected in order to account for observed levels of silent variation in this species. None of this seems too outrageous, making genetic draft a stochastic force that can rescue the neutral theory from its awkward dependency on population size.

4.4 Answers to problems

4.1 The three equations are:

$$\begin{aligned} x_2' &= (1-r)x_2 + rp_1q_2 \\ x_3' &= (1-r)x_3 + rq_1p_2 \\ x_4' &= (1-r)x_4 + rq_1q_2\,. \end{aligned}$$

4.2 The verification for A_1B_2 is

$$\begin{aligned} x_2 &= p_1q_2 - D \\ &= (x_1 + x_2)(x_2 + x_4) - (x_1x_4 - x_2x_3) \\ &= x_2. \end{aligned}$$

The other cases are done similarly. The verification of Equation 4.3 is easy (and fun).

4.3 The following program, written in Python, will print out the ratio of the final to starting heterozygosities at the B locus.

```
s, r, N = 0.1, 0.001, 5000
eps = 1.0 /( 2 * N)
x1, x2 = eps, 0.0
x3, x4 = 0.5 - x1, 0.5
while x1 + x2 < 1.0 - eps:
    p1 = x1 + x2
    q1 = 1.0 - p1
    wBar1 = 1.0 - q1 * s / 2.0
    wBar3 = 1.0 - p1 * s / 2.0 - q1 * s
    wBar = 1.0 - q1 * s
    rWD = r * (1 - s / 2.0) * (x1 * x4 - x2 * x3)
    x1 = (x1 * wBar1 - rWD) / wBar
    x2 = (x2 * wBar1 + rWD) / wBar
    x3 = (x3 * wBar3 + rWD) / wBar
    x4 = (x4 * wBar3 - rWD) / wBar
p2 = x1 + x3
q2 = 1.0 - p2
print 2.0 * p2 * q2 / 0.5
```