

LINKAGE DISEQUILIBRIUM (LD)

0.1. Preface	2
1. LD THEORY	2
1.1. Early theory	2
1.2. The effect of small population size	4
1.3. Measuring LD in finite populations	6
1.4. Measures of LD	6
2. LIBD and LD	8
2.1. Introduction	8
2.2. An aside on the effect of fixation	8
2.3. Argument #1: Probability=Correlation	9
2.4. Argument #2: LIBD and homozygosity	12
2.5. The length of identical segments	21
2.6. Argument #3 - Sampling into LIBD classes	21
2.7. LIBD computer simulation	26
2.8. Some concluding remarks.	33
3. LD UNDER A MUTATION MODEL	34
3.1. LD at first appearance of a mutation	34
3.2. Subsequent generations	37
3.3. Computer simulation	39
4. SELECTION AND LD	42
4.1. Two-locus models	42
4.2. Associative overdominance	43
4.3. Stabilising effect of a selected locus on a neutral locus	44
4.4. The apparent selective value	45
4.5. The combination of balancing and positive selection	46
4.6. Hitchhiking	49
4.7. The Hill-Robertson effect	49
4.8. Background selection	50
5. A MEASURE OF OVERALL LD	51
References	53

Contents

0.1. Preface.

LD is a huge topic that has expanded enormously in the last decade or two. My involvement dates from more than 40 years ago. I was involved in several publications on LD theory in the years 1968 - 1977, but despite working sporadically on the topic since then, not much has got to the print stage.

This chapter describes the theory I was involved in, including correlation of frequencies and its relationship with linked identity-by-descent and joint homozygosity. Despite the simplicity of this theory, it has had little impact on the LD field. This may be because of inadequacies in the theory or inadequacies in the way it was presented. I still believe that it has something to contribute, and this chapter is partly an attempt to rehabilitate the arguments. In fact I have recently managed to publish two papers, one applying the theory to calculate LD within and between populations [27] and one which includes a review of the theory [31].

Because a lot of this has not been published, there is more tedious detail in this chapter than in other parts of the PIFFLE. It starts from the beginning of LD theory because there is no other obvious place to start.

1. LD THEORY

1.1. Early theory.

Genes that are closely linked may or may not be associated in populations. Looking at parents and offspring, if genes at closely linked loci are together in the parent then they will usually be together in the offspring. But looking at individuals in a population with no known common ancestry, it is much more difficult to see any relationships.

Suppose that there is an allele A with frequency p_A in a particular population. At a closely linked locus, the frequency of the B allele is p_B . The question is, what is the expected frequency of the allele pair, or 'haplotype', AB ?

It has been known since 1918 that even for loci that are closely linked, alleles at the two loci are expected to be 'associated at random' in the population. In other words, the expected frequency of the AB genotype (haplotype), p_{AB} , is $p_A p_B$, just as if the A and B loci were unlinked.

It is reasonably easy to see why this should be true. We start by defining a new parameter, d , which goes under the slightly awkward name of the 'coefficient of linkage disequilibrium', and is defined as

$$d = p_{AB} - p_A p_B$$

which is the difference between the frequency of the AB haplotype, p_{AB} , and its expectation $p_A p_B$ if there is no LD. Note that this parameter is often denoted as D rather than d .

What Robbins showed in 1918 is that if the recombination frequency between the two loci is c , then

$$d' = (1 - c).d \tag{1}$$

where d' is the corresponding coefficient one generation later. Crow and Kimura in their 1970 textbook [2] have a two-line derivation of this relationship. With probability c the gamete is a recombinant. Assuming random mating, the A gene is therefore combined with a random B gene, giving the probability of AB in the next generation as $p_A p_B$. Amongst gametes with no recombination, the frequency of the AB haplotype stays the same. Overall the frequency of the AB haplotype in the next generation is

$$p'_{AB} = cp_{AB} + (1 - c)p_A p_B$$

and this rearranges to give equation (1). Note that all this assumes that the population size is infinite.

Since c , the recombination frequency, is some small positive number, the quantity $(1 - c)$ will be less than unity, and the coefficient d is expected to fall in each generation. Eventually it will reach zero, although this may take some time for very closely linked loci. It is for this reason that it is expected that even closely linked alleles are expected to be in 'linkage equilibrium', at least in populations that have been around for some time.

One exception to this expectation has been known since the 1950s. If there is selection, and the allele pair AB is favoured, then if the loci are sufficiently closely linked, natural selection may lead to a situation in which the A and B alleles are closely associated, so that d is some positive quantity. However this assumes that there is a substantial level of selective interaction between closely linked loci, which is only to be expected for a small minority of gene pairs.

1.2. The effect of small population size.

When we (Walter Bodmer, Ed Reed and I) started doing calculations on genetic loads [29], we assumed, like everyone else, that it didn't matter whether or not we were dealing with closely linked loci. Since these whole genome calculations had to do with thousands of loci, many of them would have to be closely linked. The 1918 theory meant, however, that we could disregard this complication.

One day, under circumstances that I don't recall, I suddenly realised that there was an enormous effect that we were ignoring. What would happen if you had lots of very closely linked loci, and the population was small? It seemed immediately clear that even if you started off with complete linkage equilibrium, this couldn't be maintained for long. Suppose, for example, that a population of 100 chromosome types was started in linkage equilibrium. Then the chromosomes might be something like:

Chromosome 1	A	b	c	d	E	f	G	H	...
Chromosome 2	A	b	C	D	E	F	G	h	...
Chromosome 3	a	B	c	d	e	f	G	H	...
Chromosome 4	A	B	C	D	e	F	G	H	...
.									
.									
Chromosome 99	a	b	c	d	E	f	G	h	...
Chromosome 100	A	b	c	d	E	F	g	h	...

There are 100 different types. Let's assume, as a first approximation, that there is no recombination. Then no new types will be created over generations - there will only be losses. After one generation it is expected that by chance about 37 (a fraction e^{-1}) types will be lost, so that about 60-70 types will remain. After two generations less than 50 types are expected to be still around. It doesn't take too many generations before the population is down to 4 types, then 3 and then 2. Let's arbitrarily assume that the two remaining chromosome types at this stage are Chromosomes 1 and 2. Then the population will look something like:

Chromosome 1	A	b	c	d	E	f	G	H
Chromosome 2	A	b	c	d	E	f	G	H
Chromosome 3	A	b	c	d	E	f	G	H
Chromosome 4	A	b	c	d	E	f	G	H
.									
.									
Chromosome 98	A	b	C	D	E	F	G	h
Chromosome 99	A	b	C	D	E	F	G	h
Chromosome 100	A	b	C	D	E	F	G	h

We can simplify by ignoring some of the loci, since they are 'fixed'. So the population can be re-written as

Chromosome 1	c	d	f	H
Chromosome 2	c	d	f	H
Chromosome 3	c	d	f	H
Chromosome 4	c	d	f	H
.					
.					
Chromosome 98	C	D	F	h
Chromosome 99	C	D	F	h
Chromosome 100	C	D	F	h

The amazing thing about the population at this stage is that no matter what pair you look at, the loci are in total linkage disequilibrium, as far removed as possible from the initial state of linkage equilibrium. I've obviously made things as extreme as possible by assuming such a small population size, and zero recombination. But there seemed no doubt from this simple simulation that there is a very striking tendency for closely linked genes to become associated if the population size is small. The expectation that closely linked genes will be in linkage equilibrium, coming from the 1918 infinite population calculations, totally misses this point.

Most people nowadays would be amazed that anyone could have believed in linkage equilibrium for closely linked genes. The whole field of LD mapping relies on this association of closely linked genes. The effect was, of course, equally obvious to others who were thinking about the same sorts of problems, and the paper by Hill and Robertson: "Linkage disequilibrium in finite populations" [12] appeared around the same time as mine, and slightly later, Ohta and Kimura: "Linkage disequilibrium due to random genetic drift" [18].

The title of my paper "The stability of linked systems of loci with a small population size" [24], sounds rather different to the Hill & Robertson and Ohta & Kimura papers. The reason was that I was rather hung up on the heterozygote advantage model at the time (see Chapter 1 on genetic loads). So I pushed on to see what is the expected effect of linkage disequilibrium on the heterozygote advantage model. Some results from this study are given in Section 4. For the moment I would like to concentrate on measuring the amount of LD due to finite size.

1.3. Measuring LD in finite populations.

As a starting point in [24], I tried to calculate the expected amount of LD for given values of c , the recombination rate, and N , the population size. I was thinking in terms of loci held at equilibrium by selection, but was able to cope with only the rather limited case where there are two alleles at each locus, held at a frequency of one-half at each. Since it is a symmetric model, positive and negative values of d are equally likely, so that the expected value of d is zero. The calculation was therefore of the expected value of d^2 , and I came up with the value of

$$E[d^2] = \frac{1}{16(1 + 4Nc)} \quad (2)$$

Computer simulation was pretty slow and expensive in those days. I did a couple of runs with $N = 50$ going for 2,000 generations and found that the average calculated wasn't too far off.

The formula with restriction to 50% frequencies obviously has very limited application. After my paper came out, I saw the paper of Hill & Robertson [12] which introduced the parameter r^2 , the square of the correlation of gene frequencies. This is a normalised version of d^2 , calculated as $d^2/p_A(1-p_A)p_B(1-p_B)$. This had come up in my paper [24] but I hadn't realised its significance..

The parameter r is an ordinary correlation coefficient defined in the usual way. The parameter r^2 is closely related to the χ^2 for a 2x2 table, the relationship being $r^2 = \chi^2/2N$. Partly for this reason it has considerable advantages of application over d^2 , and I started to try to work in terms of r^2 rather than d^2 .

1.4. Measures of LD.

The range of values that the parameter d can take is -0.25 to 0.25. However the range is dependent on allele frequencies, and the maximum

and minimum values can only be attained if the allele frequencies are 0.5. If, for example, the allele frequencies are $p_A = 0.3$, $p_B = 0.1$, then the possible range is restricted to -0.03 to 0.07.

For this reason Lewontin (1964) introduced the parameter d' , in which the value of d is divided by its minimum and maximum values for the particular observed allele frequencies, giving the parameter the range -1 to 1. The parameter d' has enjoyed a large amount of use, possibly following the recommendation of Hedrick [10].

The parameter r has, in one respect, a similar effect in removing some of the effects of allele frequency. For the case $p_A = 0.3$, $p_B = 0.1$, for example, the possible range of values is -0.22 to 0.51. This removes some of the range restrictions on d , but does not allow the full range of values allowable for d' .

While at first sight it seems reasonable to use a parameter that allows the full range of frequencies, there are strong reasons for not doing this. The fact that allele frequencies are unequal is not devoid of information. It implies that there is not a complete correlation of frequencies at the two loci. The parameter d' throws out this information and potentially gives the same value as the case of equal allele frequencies. The parameter r , on the other hand, properly takes this information into account.

There is nothing magic about the marginal allele frequencies, particularly for a neutral model, that requires that an LD parameter be made conditional on these allele frequencies. What happens, for example, if we consider the correlation between two variables, such as levels of education and income in a population. These are positively correlated, I believe. Would one then want to ask what is the level of correlation between all those individuals in the population with a particular mean income and a particular level of education? It seems to make little sense to calculate a correlation conditional on particular marginal values.

As an aside I'd also like to comment on the illogicality of the notation, in which r stands for the correlation and c stands for the recombination frequency. If one was starting from scratch, surely one would do it the other way around. Unfortunately r has been used for the correlation coefficient for the best part of 100 years, so it's not really possible to change that. It's easy occasionally to get confused between the two.

2. LIBD AND LD

2.1. Introduction.

Why should LD arise in a finite population? It is clear from the simple simulation given earlier that the reason LD arises is that after one or more generations there will be multiple copies of particular haplotypes in the population, essentially because of co-ancestry.

It is simplest to look at the case of just two loci. There are two ways of looking at the population. The first is via probability arguments, essentially the probability that two chosen gene pairs (gametes or haplotypes) will be identical copies from some previous generation. I have come to call this the Linked Identity-by-Descent (LIBD) probability. The second is via frequency arguments, that there will be a correlation of gene frequencies, or LD.

It seemed to me that LIBD and LD provide alternative descriptions of the same phenomenon [25] [28]. Furthermore, the LIBD argument leads to great simplification in deriving r^2 expectations. However the question of whether the probability (LIBD) and frequency (LD) approaches are totally equivalent is one that I'm still unsure of. Nobody else seems to have taken up the approach, which, considering the vast LD literature, can probably be taken to mean that there is a problem with it. Anyway I'll now attempt to summarise the whole of this sorry saga, and then attempt some further clarification.

2.2. An aside on the effect of fixation.

All measures of LD have the property of either being zero or undefined if allele frequencies at either of the two loci are zero. I will be dealing mainly with r^2 as a measure of LD, which becomes zero divided by zero or undefined if one of the two loci is 'fixed'.

In practice, there may seem no reason to want to calculate r^2 in such a case. However in calculating the expected value of r^2 , one wishes to give recurrence relationships for the value in one generation in terms of the value in the previous generation. The question then arises - if there is a certain probability that one of the loci becomes fixed in going from one generation to the next, how can this be taken into account in the calculation? Note that this problem does not seem to arise with backwards, or coalescence, simulation. However it is not clear to me that it is always possible to remove the problem in this way.

When considering LIBD, by contrast to LD, questions of fixation do not arise. The LIBD probability is independent of allele frequencies, being

dependent simply on population structure and recombination rates. It seems therefore, in trying to equate LIBD and LD, that fixation, or its probability, will create problems. This issue will arise at several places below.

2.3. Argument #1: Probability=Correlation.

I was involved in two papers putting forward the LIBD argument, (Sved, 1971 [25] and Sved & Feldman, 1973 [28]). It is convenient to deal with the later paper first here, since it is a much simpler argument.

The basic argument of [28] depends on the analogy with single locus calculations. The focus here is on inbreeding, specifically on the way in which the coefficient of inbreeding can be defined in terms of either frequencies or probabilities.

The definition of an inbreeding coefficient in terms of the correlation between uniting gametes is usually attributed to Sewall Wright (eg. [33]), following earlier work by Pearl and Jennings. Wright's original definition, in terms of path coefficients, seems a hybrid of probability and frequency coefficients. However the inbreeding coefficient can be defined purely in terms of a conventional correlation coefficient (Crow & Kimura [2], p67).

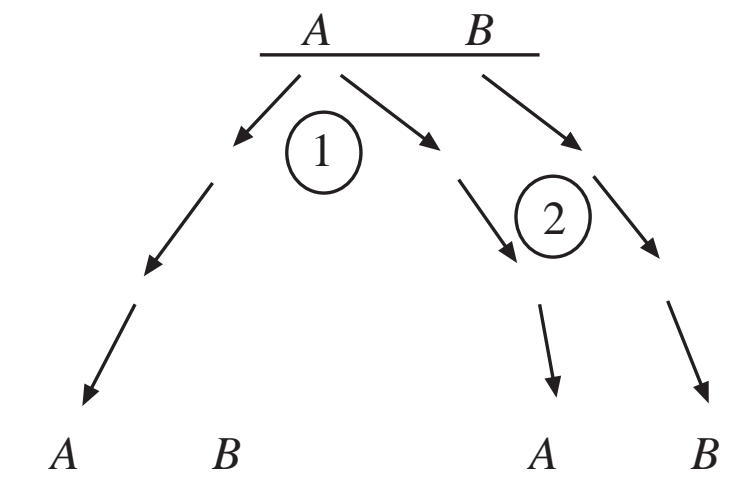
Somewhat later, the identity-by-descent definition of inbreeding was introduced by Malecot and others. By contrast to the correlation definition of the inbreeding coefficient, the IBD definition involves probabilities, not allele and genotype frequencies.

The relationship between the correlation and probability definitions may be seen in the following simple way, closely related to the argument from Crow & Kimura [2], p66. If two genes are identical by descent, then their correlation is 1. If they are not identical by descent, then their correlation is 0. Overall, therefore,

$$r_A = f_A \cdot 1 + (1 - f_A) \cdot 0 = f_A$$

This argument will only work if correlations are additive. The verity of the argument can be checked by writing out the full set of matings [2] Table 3.2.1.

The argument so far has looked only at genes at a single locus (see (1) in the above diagram). The equivalent two-locus argument can be seen by following the pathways labelled (2) in the diagram. The probability that the *A* and *B* alleles are transmitted intact without recombination



on the pathway from the common A locus ancestor can be defined as f_{AB} . In such an event, the correlation is equal to 1. On the other hand any recombination event will connect the A allele to a random B allele in the population, assuming random mating. (See later for a discussion of this point). The correlation between the A and B genes in such a case will thus be 0. The overall correlation is equal to

$$r_{AB} = f_{AB} \cdot 1 + (1 - f_{AB}) \cdot 0 = f_{AB} \quad (3)$$

The LIBD probability L defined previously is equal to f_{AB}^2 . This assumes that events in the two pathways leading to the present gametes are independent. So the result can be expressed in terms of the probability of LIBD, L , as

$$E[r_{AB}^2] = f_{AB}^2 = L. \quad (4)$$

Unfortunately unlike the single locus calculations given in [2] Table 3.2.1, there is no obvious direct demonstration that the two locus correlations are additive. For a single locus, one can easily write down the frequency of AA genotypes as $f_A \cdot p_A + (1 - f_A) \cdot p_A^2$. It is not obvious, to me at least, how one writes the frequency of AB gametes in terms of f_{AB} .

2.3.1. *The expectation for L and r^2 .* L , the probability of LIBD, refers to gametes, or haplotypes, sampled from a particular population. The first question to be settled is whether this sampling should be with or without replacement. The argument that follows assumes sampling *with* replacement. While it may seem artificial to sample the same gamete twice and refer to this as LIBD, such sampling is necessary to

equate probabilities with statistics calculated from gene frequencies. Any statistic, such as d^2 , is calculated by multiplying frequencies as if the population size was infinite. Note that it is sometimes possible to calculate statistics that do not make this assumption. For example the true frequency of homozygosity for an allele having n copies in a population of $2N$ alleles would be $n/2N \cdot (n-1)/(2N-1)$, rather than $(n/2N)^2$. Sampling without replacement would be the valid procedure if homozygosity was calculated in this way. Sampling without replacement was actually assumed in [25], but corrected to sampling with replacement in [28].

In a population of $2N$ haplotypes, the chance that the same haplotype is chosen twice is $1/2N$. Conversely, the chance that two different haplotypes are chosen is $1 - 1/2N$. Under the Wright-Fisher model in which sampling is at random, with replacement, from the previous generation, the probability that two such haplotypes are identical is simply the equivalent probability L in the previous generation, multiplied by $(1 - c)^2$, the probability that there has been no crossover on either pathway from the parent. Thus the recurrence relationship is

$$L' = \frac{1}{2N} + (1 - \frac{1}{2N})(1 - c)^2 L \quad (5)$$

This equation easily generalises to any number of generations. It gives an equilibrium value for L of

$$\frac{1}{1 + (2N - 1)(2c - c^2)}$$

so that for small values of c we have

$$E[\hat{L}] \approx \frac{1}{1 + 4Nc} \quad (6)$$

This agrees with (2) derived earlier under conditions where allele frequencies are held at a selective equilibrium of one-half.

The rate of approach is given by

$$(1 - \frac{1}{2N})(1 - c)^2.$$

So now, if one believes that $E[r^2] = L$, the expected value of r^2 in the offspring generation in terms of the parent generation is:

$$E[r^{2'}] = \frac{1}{2N} + (1 - \frac{1}{2N})(1 - c)^2 r^2 \quad (7)$$

and

$$E[\widehat{r^2}] \approx \frac{1}{1 + 4Nc} \quad (8)$$

2.4. Argument #2: LIBD and homozygosity.

The argument of the previous section focuses on LIBD and r^2 while ignoring homozygosity. Clearly LIBD will lead to an increased frequency of double homozygotes over what is expected in a population in which there is no LD.

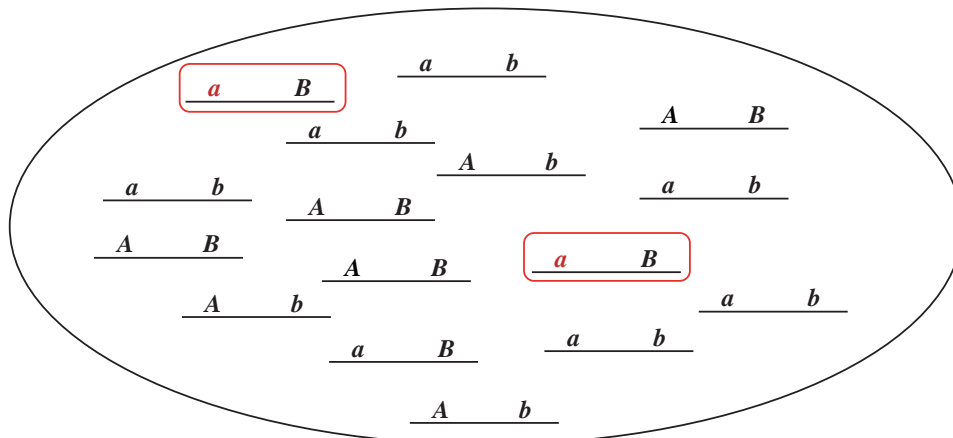
My original attempt [25] to derive a relationship between LIBD and r^2 used homozygosity as the basis for the argument. Joint homozygosity, it was argued, could be defined in terms of frequency parameters or in terms of probability parameters. Equating the two approaches led to (4). I have spent a lot of time over the years going back to try to validate this argument. In this section I will describe the derivation, highlight a couple of what now seem like mistakes, and try to put forward a valid version.

The obvious expectation for the frequency of joint homozygosity at two loci would appear to be based on the following argument. LIBD necessarily leads to joint homozygosity. The non-LIBD class, in which recombination occurs in one or other pathway, might be expected to contain double homozygotes at just the frequency in the overall population, the product of the homozygous frequency at the individual loci. Unfortunately this argument doesn't seem to work, and leads to no simple equation of probability and LD parameters.

The only way I was able to derive such a relationship was by considering not simply the probability of LIBD at two loci, but rather what was described as a 'conditional probability'. I need to elaborate on this here. What is being considered is a situation in which there is an A locus with A and a alleles segregating, and a linked B locus with B and b alleles. The way I looked at it was that at the A locus all A alleles are IBD from some previous ancestral gene, and similarly all a alleles are IBD. On the other hand A alleles are not IBD with a alleles. My analysis required disregarding alleles known not to be IBD, in other words conditioning on only alleles identical in state at one locus. It is rather a messy situation, trying to force a model with two alleles at each locus into a probability framework.

Coalescence theory would require a specific mutation parameter that is missing from this analysis. The analysis presented later in this section is more or less in such terms, assuming that mutation is much rarer

than recombination. Therefore haplotypes with the A allele coalesce to a different ancestral haplotype compared to those haplotypes with the a allele.



The argument in 1971 [25] was the following. Suppose that one chooses a haplotype, and then chooses another haplotype containing the same allele at the A locus. How does this affect homozygosity at the B locus? Note that the second haplotype could be the same haplotype selected twice.

Assuming that there is some LD, it seems clear that there will be increased 'homozygosity' at the B locus. In the above diagram we've randomly chosen haplotypes containing an a allele. The existence of LD makes it more likely that the B locus will be B/B if d is negative and b/b if d is positive.

The calculations below look at the amount of homozygosity at the B locus. It is convenient to introduce the symbol h to describe this frequency, remembering that this refers specifically to homozygosity at the B locus. Similarly the symbol h_c is introduced to describe the homozygosity at the B locus conditional on choosing the same allele at the A locus. The conditional probability of homozygosity is:

$$h_c = \frac{p_{AB}^2 + p_{Ab}^2}{p_A} + \frac{p_{aB}^2 + p_{ab}^2}{p_a}$$

Substituting for the haplotype frequencies using $p_{AB} = p_A p_B + d$, and similarly for the other three haplotypes, this simplifies to

$$h_c = p_B^2 + p_b^2 + \frac{2d^2}{p_A p_a} = h + \frac{2d^2}{p_A p_a} \quad (9)$$

So far, this has been a frequency argument. We now need to bring in a probability parameter to measure LIBD. In the 1971 paper I used the parameter Q . As mentioned above, this was defined conditioned on choosing alleles IBD at the A locus. This was all introduced in a very messy way, and was not understood by anyone, evidently including myself. Anyway I'll first repeat the basic argument here. I'll make one change by calling the LIBD parameter L rather than Q . And later I'll introduce an extra parameter that specifically measures LIBD conditional on choosing the same allele at the A locus.

What is the probability of homozygosity at the B locus, h_c , in terms of the parameter L ? If there is no crossingover on either pathway then the probability of homozygosity is 1. On the other hand, one or more crossovers will ensure that the alleles at the B locus are combined at random, giving the probability of homozygosity as $h = p_B^2 + p_b^2$. The random mating assumption is the same as one made in the calculation of the previous section, and will be considered further *here*. Under these circumstances, the overall probability of homozygosity is

$$h_c = L \cdot 1 + (1 - L) \cdot h$$

which simplifies to

$$h_c = h + 2Lp_Bp_b \tag{10}$$

Comparing the two approaches for predicting h_c , ie comparing (9) and (10):

$$\frac{2d^2}{p_Ap_a} = 2Lp_Bp_b$$

so that

$$\frac{d^2}{p_Ap_ap_Bp_b} = r^2 = L$$

This relationship of frequency with probability parameters is only an expectation over populations with the same probability history, so that we should write

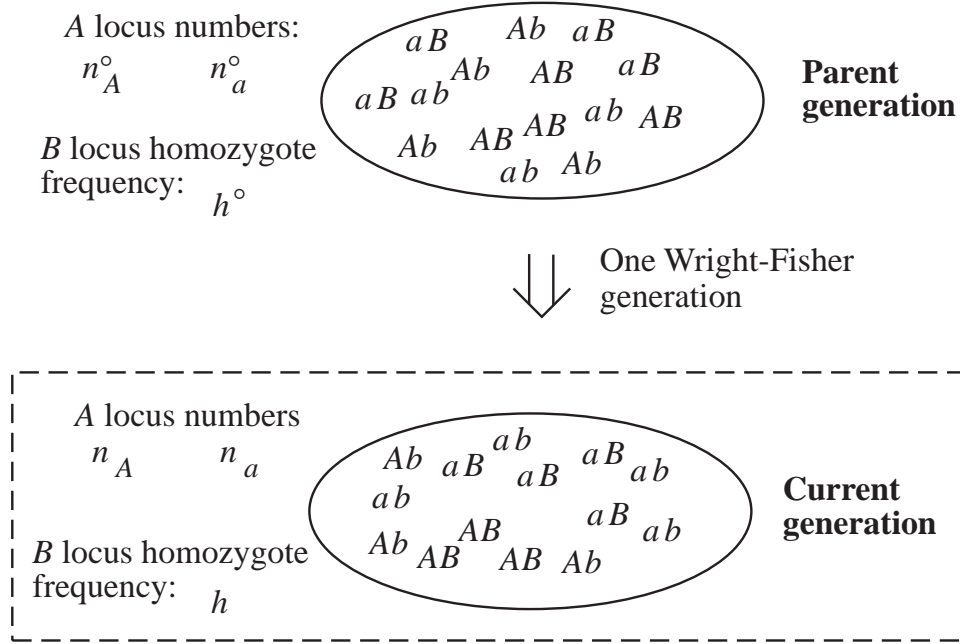
$$E[r^2] = L. \tag{11}$$

This is the same as equation (4) derived in the previous section. I'll have more to say below on this equation. However first comes an attempt to clear up the confusion regarding the treatment of LIBD and homozygosity.

Equation (9) is straightforward. It deals simply with frequencies in any current population. The problem is with (10). The probability approach appears to need a mixed population of A and a alleles. I assumed in 1971 [25] that the coalescence properties of such alleles were identical to those of a population where allele classes were not specified. My only defence is that coalescence didn't really exist as a method at the time, and I was basically working in the dark.

In hindsight it is clear that one can't write the probabilities of IBD as I assumed. If all alleles are equivalent, the probability of choosing the same allele twice from the population is $1/2N$. But that seems wrong in the case where one is specifically directing attention to A alleles which constitute only a portion of the population. I'll now try to take account of this complication.

It seems convenient to introduce a diagram such as that given below, and to concentrate on frequencies in the 'current generation', marked with the rectangle. In this generation there are n_A A alleles and n_a a alleles ($n_A + n_a = 2N$). We then ask: "what's the probability that randomly chosen pairs of haplotypes with the same A allele from the current generation are LIBD?". I'll call this probability L_c , the subscript indicating that this is a conditional LIBD parameter. The absence of a superscript indicates that it is a parameter of the present generation. The rather clumsy use of a 'o' superscript for all parameters of the parent generation (see figure) is needed to avoid changing the frequencies of the current generation. I could have used a prime symbol, ', instead of the 'o' superscript, except that this is usually used for the following generation rather than the previous one.



We now select two alleles from the current generation population. The probability of selecting the first allele as A is $n_A/2N$. In this case the probability that the exact same haplotype is selected twice is $1/n_A$. If the same haplotype is selected, then LIBD is assured. If a different A allele is selected, with probability $1 - 1/n_A$, the haplotypes could still be identical from the previous generation, provided there has been no recombination between the generations. The probability of these events is $(1 - c)^2 L_c^\circ$, where L_c° is the equivalent conditional probability in the parent generation.

Overall, the contribution to LIBD from selecting an A gene is

$$\frac{n_A}{2N} \left[\frac{1}{n_A} + \left(1 - \frac{1}{n_A}\right) (1 - c)^2 L_c^\circ \right]$$

which is equal to

$$\frac{1}{2N} + \left(\frac{n_A}{2N} - \frac{1}{2N} \right) (1 - c)^2 L_c^\circ$$

To this must be added the equivalent contribution from the other possibility, that the a allele is selected at the A locus. This contribution is equal to

$$\frac{1}{2N} + \left(\frac{n_a}{2N} - \frac{1}{2N} \right) (1 - c)^2 L_c^\circ$$

The sum of these two terms is the conditional probability of LIBD in the current generation L_c . This simplifies to

$$L_c = \frac{1}{N} + (1 - \frac{1}{N})(1 - c)^2 L_c^\circ \quad (12)$$

Note that the probability from new LIBD, $1/N$, is twice the regular IBD probability. In other words, the coalescence distance is only half the value of that for alleles not chosen as being of the same class. I have simulated this using a forward simulation and checking back over generations, and it does work. Presumably for three alleles the distance would be one third of the regular value.

Now we come to a second misunderstanding in [25] which seems to more or less cancel out the first one. The frequencies in (9) clearly ought to relate to the population in the current generation. But is this the case for (10)? It's like the case of inbreeding for a single locus, where, for example, the frequency of homozygotes can be given as $p_B^2 + p_b^2 + 2fp_Bp_b$, where f is the inbreeding coefficient. But if the allele frequencies p_B and p_b refer to frequencies within the current population, and there is random mating, then the frequency of homozygotes would be simply $p_B^2 + p_b^2$. The inbreeding equation only works if the frequencies refer not to those in the current population but to an overall, or reference, value. Clearly the frequencies in equation (10) should similarly refer to a reference value, not to the current generation value. The important point is that this value doesn't change between generations. I'll call this homozygosity at the B locus H .

Using the L_c parameter, the conditional homozygosity in the parent and current generation would then be:

$$h_c^\circ = H + (1 - H)L_c^\circ \quad (13)$$

$$h_c = H + (1 - H)L_c \quad (14)$$

The argument for these is equivalent to that used previously for the derivation of (10). It again assumes that amongst haplotype pairs where there has been one or more crossovers in the history, the probability of homozygosity is the same as in the current generation. As discussed later in connection with Figure 8, there is an unstated assumption here about the A and B frequencies in the history of the current population.

Equations (13) and (14) should be contrasted with the equivalent equations which I gave previously. The equations for successive generations are:

$$h_c^\circ = h^\circ + (1 - h^\circ)L^\circ \quad (15)$$

$$h_c = h + (1 - h)L \quad (16)$$

Equation (16) is the same here as equation (10), while (15) is the equivalent equation one generation earlier. Note that the latter equations are in terms of L rather than L_c , and that the frequency of homozygosity changes between generations. I am not still claiming that these are correct. The purpose of this calculation is to compare this set of equations with the correct set, (13) and (14).

It is now convenient to derive a recurrence relationship for h_c in terms of h_c° by eliminating the L parameters. Looking first at equations (13) and (14), this requires first substituting for L_c in terms of L_c° from equation (12), and then eliminating the L_c° parameter. Doing this and simplifying gives:

$$h_c = \frac{1}{N} + H\left(1 - \frac{1}{N}\right)(2c - c^2) + \left(1 - \frac{1}{N}\right)(1 - c)^2 h_c^\circ \quad (17)$$

We now follow through an analogous calculation for equations (15) and (16). This needs to be based on a relationship between L and L° . Equation (5) provides just such a relationship. However I will assume a slightly different relationship here, in which $2N - 1$ substitutes for $2N$:

$$L = \frac{1}{2N - 1} + \left(1 - \frac{1}{2N - 1}\right)(1 - c)^2 L^\circ \quad (18)$$

This can be used to substitute for L in terms of L° in equation (16). Then equation (15) can be used to express L° in terms of h° and h_c° . Substituting this into equation (16) gives a recurrence relationship for h_c in terms of h_c° :

$$h_c = \frac{1}{N} + h^\circ\left(1 - \frac{1}{N}\right)(2c - c^2) + \left(1 - \frac{1}{N}\right)(1 - c)^2 h_c^\circ \quad (19)$$

This is identical with equation (17), except for the constant homozygosity term, H in one case and h° in the other. These two homozygosities

are defined in rather different ways, H in connection with L_c and h° in connection with L . In order to compare the two, it seems necessary to consider a boundary condition, in which the population starts with no LIBD. In this case

$$L^\circ = L_c^\circ = 0$$

Then comparing equations (13) and (15) shows that H must equal h° to make the two approaches equivalent.

What this identity of (17) and (19) shows, I claim, is that the approach of [25] leads to the correct result, or that equation (11) is validated:

$$E[r^2] = L.$$

But one aspect of this result doesn't seem to make sense. Equation (11) looks identical to equation (4). But the definition of L in the latter was simply the probability of LIBD. In the derivation of equation (11) in Sved (1971) [25], and in the current section, L was defined as the probability of LIBD conditional on selecting identical genes at the A locus. So which is correct?

In trying to answer this question, it seems relevant to note that the treatment of this section specifies the parameter L_c objectively, but not the parameter L . The latter is defined just by equation (10) and by the recurrence relationship (18). Nowhere in its definition is there anything about conditioning. It seems clear that equation (4) is the correct one, or that the definition of L is simply the LIBD probability.

Why has it been so difficult to derive this result using the homozygosity argument of this section? Essentially I believe that it is because of the inherent problem introduced by considering a two-allele model. The usual coalescence arguments, eg. Hudson (1985) [13], are based on models incorporating both mutation and recombination. The two-allele model assumes that all A alleles coalesce to a single ancestral allele, and similarly all a alleles to a different ancestral allele. Such a model thus has the contradictory requirement of low probabilities of mutation at both loci while at the same time requiring segregation at both loci. The necessity for conditioning thus appears to come from attempting to coerce a two-allele model into a coalescence framework. I believe that the arguments using L_c sidestep this complication, at the cost of leaving the L parameter essentially undefined.

As an aside, the recurrence relationship (18) based on $1/(2N - 1)$ needs to be used rather than the recurrence relationship (5) based on

$1/2N$. There has always been some ambiguity here. The result after one generation starting from linkage equilibrium clearly gives an r^2 value of $1/(2N - 1)$ as shown long ago by Haldane [8]. But computer simulation seems to show that from the second generation the formula based on $1/2N$ is more accurate. There is very little difference between the two for realistic values of N .

Finally, is the equivalence of equations (17) and (19) just an 'algebraic fluke'? I believe not, but will leave the rationalisation as an exercise to the reader. Please tell me if you see a simple reason why the difference between the L_c and L parameters exactly cancels out the differences in homozygote frequencies between generations.

I don't seriously expect anyone to follow through all this stuff, but my congratulations, or condolences, to anyone has bothered. It may seem rather pointless trying to resurrect the argument of 1971 [25]. However I still believe that this argument is important, because it validates the key concept that homozygosity of linked genes studied using LD parameters (r) is exactly equivalent to homozygosity using probability (LIBD) parameters.

I should note here that a number of authors have given solutions to the problem of finite size disequilibrium, and there has also been discussion about homozygosity and LD, eg. Sabbati & Risch [21]. These latter authors have used observed two-locus homozygosity, which I don't believe can lead to a simple relationship like that derived above using conditional homozygosity. However their interest is mainly in multiple allele disequilibrium, and perhaps there are advantages to their approach. I should also mention the multiple parameter approach of Weir and Cockerham (1974). This considers all possibilities of IBD at two loci. However their parameter for joint IBD at the two loci combines the case of LIBD with that of IBD at two loci via separate pathways. While this may lead to a more comprehensive treatment, it obscures the simplicity of the LIBD approach.

There are many others whose contributions I have not acknowledged. The obvious one that has the simplicity of the LIBD approach is that of Hill and Robertson [12]. Their solution gave the equilibrium value of r^2 as $1/4Nc$, simply equating gain of r^2 through segregation, $1/2N$, with loss through recombination, $(1 - c)^2$. And Bill Hill (personal communication) has pointed out to me that the relationship can easily be modified to give $E(r^2) = 1/(1 + 4Nc)$, the same as equation (8). An equivalent demonstration of (7) has also recently been given by Tenesa

et al. [32], using just the properties of the correlation coefficient and not IBD arguments.

2.5. The length of identical segments.

Deviating from the main argument of two-locus LD, I'd like to briefly discuss the topic of chromosome segments. While LIBD can be thought of in terms of two loci, it can also be described in terms of the length of identical chromosome segments. Ultimately what matters in a population is the position of crossovers along a chromosome. Each time a crossover occurs, it creates a 'junction', in the terms of Fisher (1949) [5]. Some of these segments defined by junctions will overlap each other, leading to a sub-segment on which all pairs of genes are identical by descent (IBD). LIBD in a population can thus be described in terms of the distribution of length of such segments.

I calculated a statistic like this in [25]. However I considered a simpler statistic. Given that one has a locus at which the two alleles are IBD, what is the distribution of identical segment around such a locus? In a population of size N , the mean length of segments at equilibrium is not very dependent on chromosome length, and is approximately

$$\frac{1}{2N}(\log N - 1)$$

The mean and standard deviations of segment lengths for three population sizes are as follows:

Population size	100	1,000	10,000
Mean	1.8	0.3	0.04
Standard deviation	6.8	2.2	0.7

In all cases, particularly the highest population size, the standard deviations are high compared to the mean. It seems that it is occasional long homozygous segment that contributes the most to the mean.

The Hapmap SNP project, and other sequencing projects, have started to provide information on such segment lengths. The above theory and the junction theory of Stam (1980) [23] could, in theory, be used to estimate past effective population sizes. In practice, however, the method may not have much advantage over two-locus methods.

2.6. Argument #3 - Sampling into LIBD classes.

There is another, different, way of looking at the buildup of LD in a finite population. The figure below defines what I would like to call

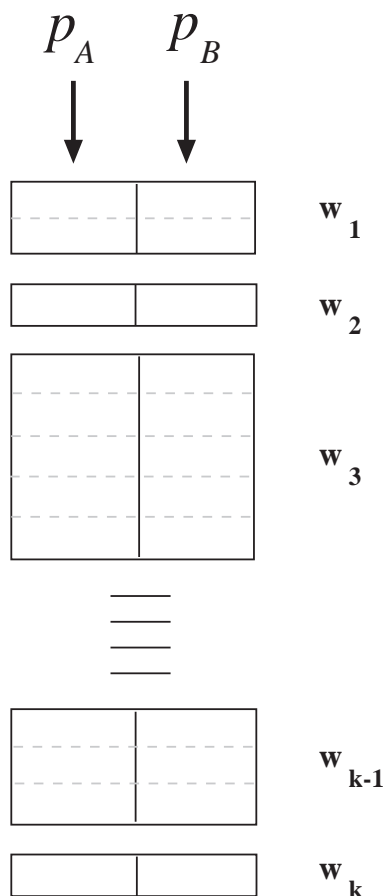


FIGURE 1. Sampling into LIBD boxes using the current allele frequencies

sampling of A and B alleles into the unequal-sized LIBD boxes as shown in the figure.

The distribution of class sizes is easily calculated in this model. It is exactly equivalent to the model of a single infinite allele locus, with mutation replacing recombination from the two-locus model. By arguments similar to those given above for the recurrence relationship of L (6), and analogous to the argument of Kimura and Crow [14], the expected equilibrium value of the sum of squares of the class frequencies, w_i , is

$$E\left[\sum_{i=1}^k w_i^2\right] = \frac{1}{1 + 4Nc}$$

I will first give some calculations that show that random sampling of alleles into the LIBD classes leads to the approximate relationship $E[r^2] = E[\sum w^2]$. Then computer simulation is given showing that there is an extra historical correlation not taken into account by the random sampling calculation.

2.6.1. *The sampling process.*

The first point to note is that the $\sum w^2$ term arises naturally from independent sampling into the unequal sized boxes. This is most easily seen from sampling of a single locus.

Consider a single allele with frequency p . The frequency from sampling into the classes can be formally written as \bar{p} , where

$$\bar{p} = \sum \delta_i W_i / 2N = \sum \delta_i w_i$$

where the summation is over all k classes and where $\delta_i = 1$ or 0 with probability p .

The variance of \bar{p} is equal to $p(1-p) \sum w^2$. Compared to sampling into boxes of size 1, the variance is increased by the factor $\sum w^2 / 2N$.

The same increase is true for any linear combination of allele frequencies. This is easily shown for any combination such as $c_1 p_1 + c_2 p_2$.

The correlation is not a linear combination of allele frequencies, but it turns out to be quite close to one. Accepting linearity, for the moment, then the expectation of r^2 can be written down. Sampling of A and B alleles into $2N$ boxes of size 1 gives the variance of r , or $E[r^2]$, as $1/(2N-1)$ or approximately $1/2N$. Sampling into unequal sized boxes increases the variance by a factor of $\sum w^2 / 2N$, provided that r is close to a linear combination of frequencies. The variance of r , or expected value of r^2 , is thus

$$E[r^2] = \sum w^2 \tag{20}$$

2.6.2. *Why r is close to a linear combination of frequencies.*

I follow the arguments of Fisher (1922) [4]. The context of Fisher's paper was a dispute regarding the number of degrees of freedom of a 2x2 contingency χ^2 . Fisher pointed out that the contingency χ^2 can be expressed in the form

$$\chi_c^2 = \frac{y^2}{V} \tag{21}$$

where

$$y = \frac{p_{AB}}{p_A} - \frac{p_{aB}}{p_a} = \frac{n_{AB}}{n_A} - \frac{n_{aB}}{n_a} \quad (22)$$

and

$$V = \frac{n_B}{2N} \cdot \frac{n_b}{2N} \cdot \left(\frac{1}{n_A} + \frac{1}{n_a} \right) \quad (23)$$

I haven't previously used quite this notation, but I hope it is clear that n_{AB} represents the number of AB haplotypes, n_A represents the number of A alleles, etc. If there is independence of the A and B alleles then y has expected value of zero. Assuming that p_B and p_b are estimated by $n_B/2N$ and $n_b/2N$ respectively, then the variance of y is equal to V . So the RHS of (21) is just an $(SND)^2$, or an ordinary one degree of freedom χ^2 . I believe that Fisher's opponent in this argument (ES Pearson?) was trying to claim that it should have 3df. I was brought up in a Fisherian department, partly by the great man RA Fisher himself, and so my memories on this are perhaps not to be trusted.

The point of this argument, as regards the expectation of r^2 , is that $r^2 = \chi^2/2N$, so that r^2 is also close to being a linear combination of frequencies. This is exactly the case for the numerator y , if n_A and n_a are fixed rather than random values. I believe that this is the way that Fisher thought about the problem. The two sections that immediately follow are an attempt to follow up on this question of whether it is legitimate to regard n_A and n_a as fixed values.

2.6.3. *fixed number vs. fixed probability sampling.* Fixed number sampling, as the name suggests, involves just a permutation exercise of assigning n_A A alleles into $2N$ boxes. Fixed number sampling makes little sense if only a single variable is being considered, since the order is of no consequence. However with a second independent sampling, of n_B B alleles into the $2N$ boxes, the number of AB haplotypes becomes a random variable.

For sampling of A and B alleles, there are therefore three possible scenarios:

- (i) fixed number sampling at both loci
- (ii) fixed probability sampling at both loci
- (iii) fixed number sampling at one (A locus) and fixed probability sampling at the other.

I believe that it is scenario (iii) that Fisher had in mind. Under this scenario, the numbers n_A and n_a are indeed constant.

Does the choice of sampling have any consequences? Computer simulation of these three scenarios was of course not possible in 1922, but it is easy now. For each of the three I did a quick simulation by sampling 10^8 replicates of populations of size $2N = 100$, with $n_A = 40$ or $p_A = 0.4$, and with $n_B = 20$ or $p_B = 0.2$. Each scenario led to average values of r^2 that were significantly different from $1/2N$ but indistinguishable from $1/(2N - 1)$. For the case of independent sampling of $2N$ allele pairs, therefore, fixed sampling at one locus and random at the other seems no less valid than the usual model of random sampling at both loci.

2.6.4. *fixed number number sampling into unequal boxes.* Fixed number sampling into unequal-sized boxes turns out to be a tricky proposition. We can, for example, consider the case where we wish to sample 43 A alleles into 50 boxes each of size 2. It can't be done.

This is, however, a very artificial case. I have simulated many cases with unequal-sized boxes, and in most cases where $2N$ is 100 or even less it turns out that it is possible to sample any number n_A of A alleles into the boxes in many different ways.

All of this gets rather messy, and I'm not sure whether it is worth pursuing in any more detail. I'll go into some detail in the following section, using computer simulation to test the validity of (20).

2.7. LIBD computer simulation.

The above theory predicts what happens when A and B alleles are sampled into LIBD classes, along with the buildup of LD. Any computer simulation to test the theory therefore needs to follow both classes and genes. I have written a haploid Monte-Carlo simulation program that specifically enumerates each class as it arises through recombination, and specifies the allele contained in each class at the A and B loci. In this way one can produce a complete picture of a population showing which individuals belong to the same LIBD class and what their genotype is.

The theory derived above has not involved any specific mutation parameters, and the first program considered also does not consider mutation. Each 'run' of the computer program therefore involves a starting population, typically something like a $2N = 512$ population with initial haplotype numbers 128 AB , 128 Ab , 128 aB and 128 ab . Each haplotype initially starts as a separate class. Classes increase in frequency or are lost by chance, and new classes are created each time

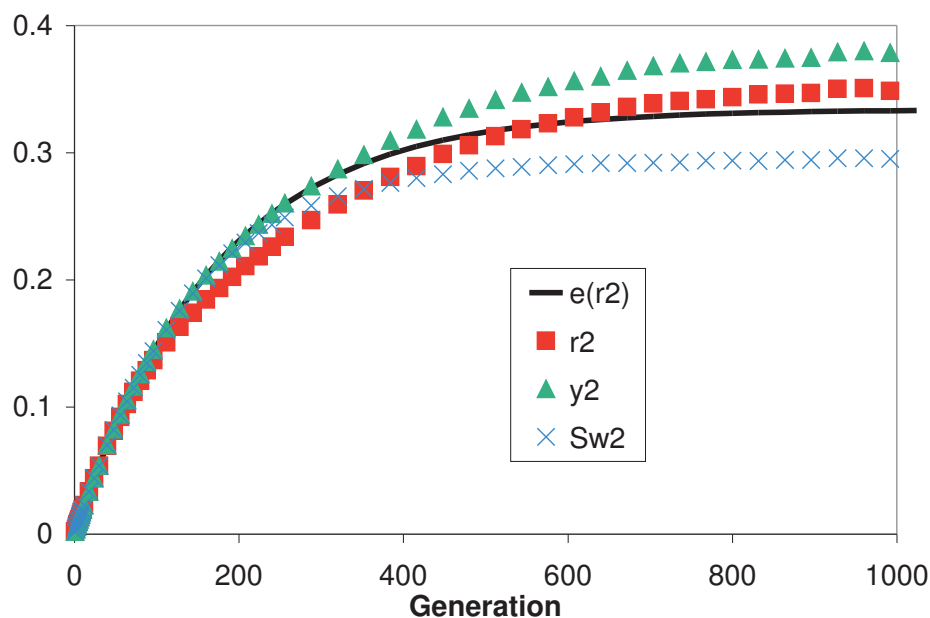


FIGURE 2. r^2 vs. expectation for build-up of LD

a recombination event occurs. Each 'run' ends with fixation at either the A or B locus. Usually, but not always, the run ends with some of the initial classes still present.

The Wright-Fisher process is simulated using sampling with replacement. For each new haplotype, with probability $(1 - c)$ a haplotype is sampled from the previous generation. With probability c a recombination event occurs, producing a new class containing A and B alleles sampled independently from the alleles of the previous generation. The independence follows from the random mating assumption that ensures that the non-allelic genes on homologous chromosomes in a diploid individual are combined at random.

For later use in the calculation, the allele and haplotype frequencies at the stage of production of each new class are recorded and stored. As a practical consideration, the continued gain and loss of classes requires renumbering of the classes at each generation of the simulation.

The first simulation shows the results from a series of runs with $Nc = 0.5$. Calculations of the correlation were made at every generation initially, then at successively longer generation intervals. They show the build-up of correlation over time (red squares) compared to its

expectation given by (7) (thick line). The agreement appears to be good in early generations, but less so in later generations.

The graph also shows two other statistics. The value of $\sum w^2$ is shown using crosses. The expected value of this statistic is the same as for $E[r^2]$. Surprisingly, since the derivation of $E[\sum w^2]$ involves no approximations, the observed value dips below its expected value in later generations.

The reason for this disagreement must be found in the premature termination of runs where fixation has occurred at either the A or B locus, at which time r^2 becomes indeterminate. From the point of view of calculating $\sum w^2$ there is no need to terminate the run. I have continued the simulation of fixed populations to show agreement between observed and expected $\sum w^2$. It seems clear that populations in which $\sum w^2$ is high by chance are more likely to be ones in which fixation occurs early. As mentioned previously, the topic of fixation, and its effect on the formulae for r^2 and $\sum w^2$, is a difficult one that is considered further below.

The other curve shown in Figure 2 is the statistic y^2 , the numerator of r^2 as defined in (22). This closely tracks the value of r^2 . The curves given in Figure 2 are based on a large number of replicate runs, more than 300,000. However the agreement between y and r can be seen from a much smaller number of replicates. Figure 3 shows a randomly chosen set of 100 replicate populations. The fluctuations are much wider, particularly towards the later generations where ultimately only 8 populations are still segregating in this particular simulation. However y and r still track each other closely. It is clear that the value $2NV$ from (23) plays a reasonably small role in the calculations.

2.7.1. *Aside on values of $2NV$.*

It may seem strange that the value $2NV = \frac{n_B n_b}{2N} \cdot (\frac{1}{n_A} + \frac{1}{n_a})$ should be close to one. There is one circumstance, however, in which it is easily shown that this is the case. In the extreme case of $r^2=1$, it must be the case that either $p_A = p_B$ or $p_A = p_a$. In either case, it is seen that $2NV$ is equal to unity.

Figure 4 shows the results from a joint plot of the values of r^2 and $2NV$ over 1,000 replicate populations at generation 1024 for the case $Nc = 1$. The graph plots $\ln(2NV)$ because of the high range of values that $2NV$ can take. However it is clear that such high and low values can occur only in the range of very low r^2 values. In the parts of the range

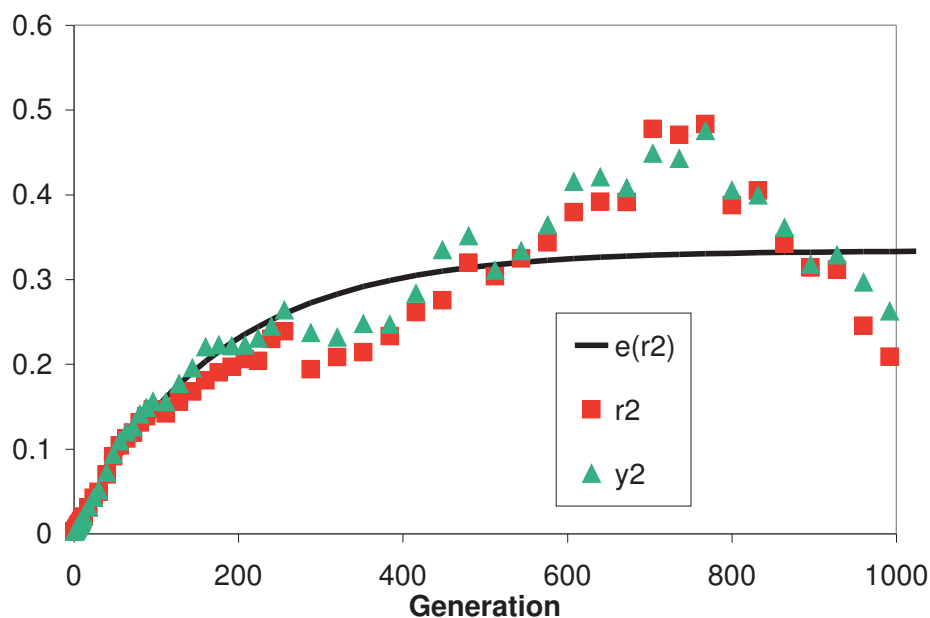


FIGURE 3. r^2 vs. expectation for build-up of LD with small number of replicates

where high values of r^2 occur, the value of V is constrained to close to $\ln(2NV) = 0$ or $2NV = 1$. It is precisely these high values of r^2 that contribute strongly to the mean value.

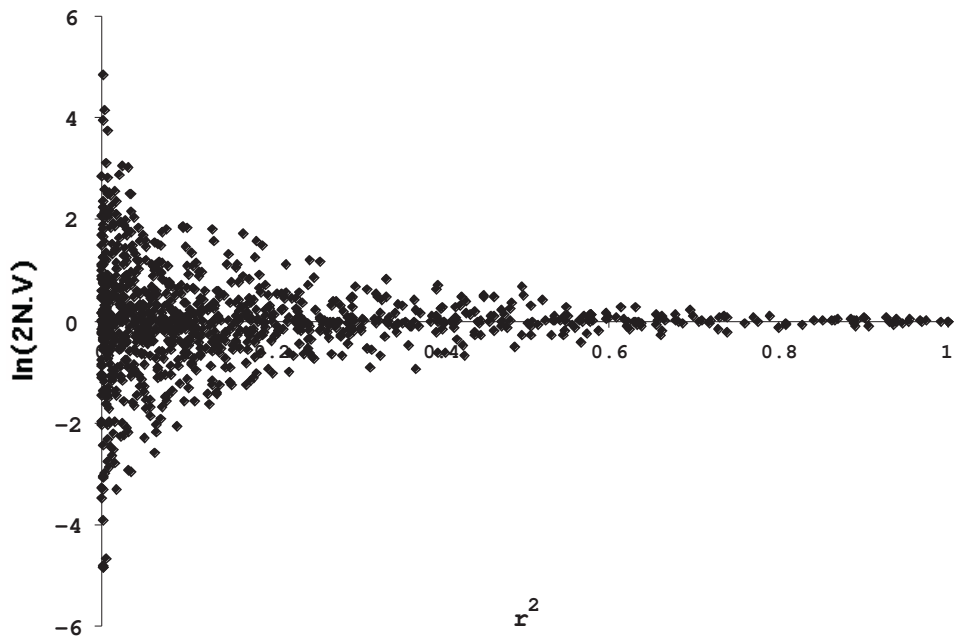
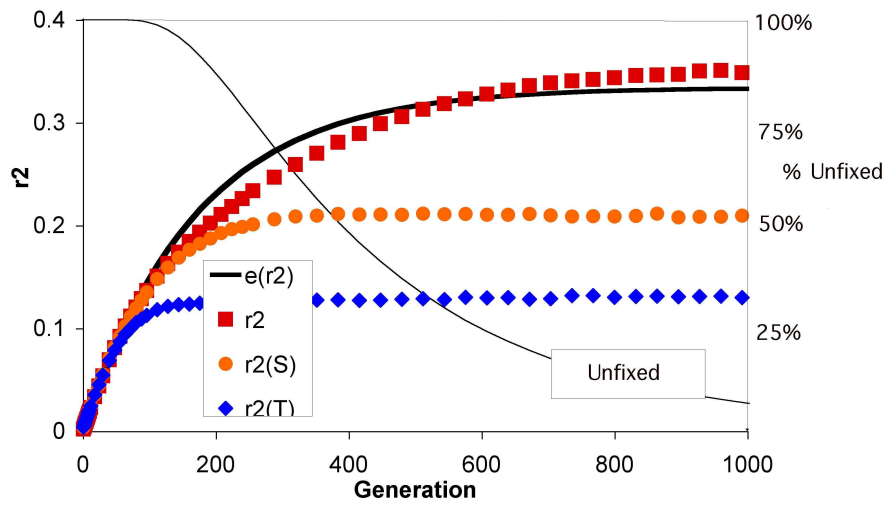
2.7.2. Sampling into LIBD classes.

The simulation of classes and frequencies gives us an opportunity to check on the theory derived above for sampling into LIBD classes. Each time r^2 and the w_i values were calculated in a population, an additional sampling operation was carried out in which populations were made up by sampling A and B alleles randomly into the existing LIBD classes. Furthermore this sampling was done in two ways.

(a) Alleles were sampled into each class with the same probability as the allele frequency when the class was created. The r^2 statistic was then calculated for each population, and designated as $r^2(S)$.

(b) Alleles were sampled into each class with the probability given by the current allele frequency. This statistic was designated $r^2(T)$.

Results are shown in Figure 5. The values of $r^2(S)$ (orange circles) and $r^2(T)$ (blue diamonds) differ substantially. What's somewhat distressing, however, is that the values of $r^2(S)$ depart so markedly from the values of r^2 . Here we are sampling into the exact same classes

FIGURE 4. Plot of r^2 against $2NV$ FIGURE 5. Comparison of sampling methods against r^2

as used to calculate r^2 , using the exact same allele frequencies from the population at the time that sampling for the new LIBD class took place. So what's going on?

One clue comes from the fact that by the time there is a noticeable discrepancy between $r^2(S)$ and r^2 , fixation has started to occur. A second graph (Figure 5) expands the early generations, showing the divergence coinciding with the start of fixation.

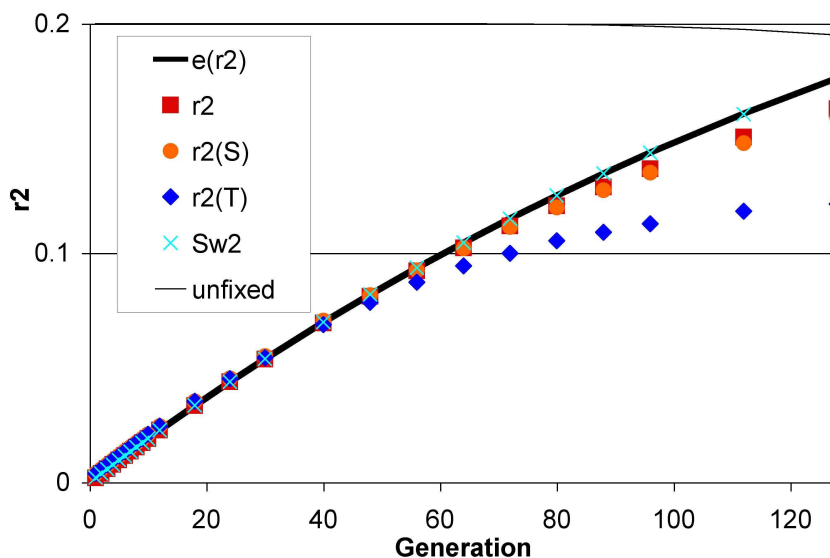


FIGURE 6. Early generations of Figure 5

The discrepancy can be rationalised in a similar manner to that argued above to explain the discrepancy between $\sum w^2$ and its expectation. Fixed populations tend to be the more extreme ones. Eliminating these brings in a bias.

It is possible to investigate the effects of bias by continuing the fixation of populations beyond the point of fixation. Values of r^2 cannot be calculated in such cases and must be omitted. However many of the fixed populations will give a determinate result for $r^2(S)$, based on older LIBD classes started before fixation occurred. Figure 7 shows the results from one such simulation, having the same starting parameters as Figure 5. The discrepancy between $r^2(S)$ and r^2 is reversed in this case. Extreme $r^2(S)$ values are found in later generations, where few LIBD classes contribute segregating alleles.

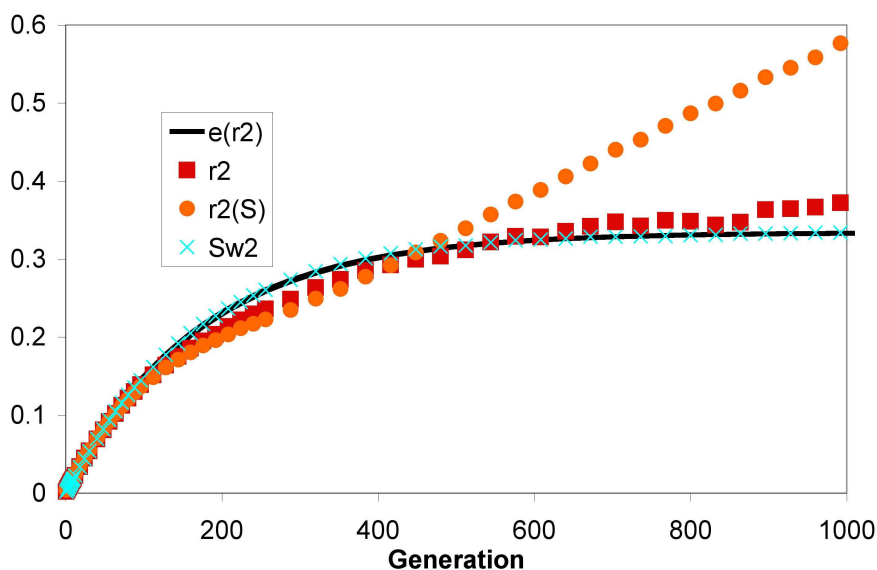
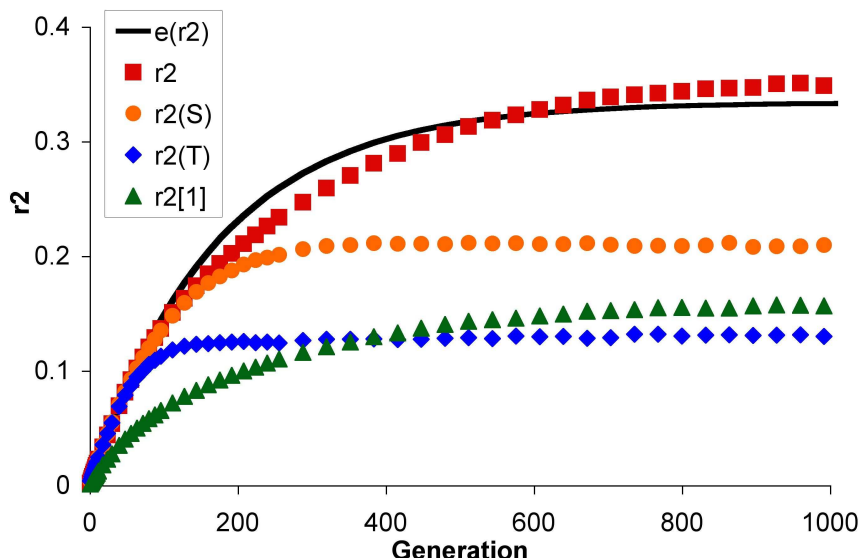


FIGURE 7. $r^2(S)$ plot including cases of fixation

The important comparison for present purposes is between $r^2(S)$ and $r^2(T)$ in Figure 5. This is repeated in Figure 8. Note that $r^2(T)$ is undefined for fixed populations, so that this comparison can only be made for the unfixed case. The discrepancy between these two statistics is highly significant. It shows that the assumption of sampling into LIBD classes with current frequencies, as assumed in the calculation leading to (20), introduces a substantial error.

Figure 8 also shows a second illustration of this effect. Plotted in green triangles is the calculation of r^2 but weighted so that each LIBD class contributes only a single observation, rather than a number weighted by the size of the class. This $r^2[1]$ statistic is a test of whether the frequencies are uncorrelated *between* classes, as assumed in the derivation of (20). Again this assumption is shown to be substantially inaccurate, as the value of $r^2[1]$ rises steadily over the course of the simulation.

It is convenient to introduce the term 'Historical Correlation' to describe this effect. Over the course of time, allele frequencies fluctuate. It doesn't matter whether the frequencies at the A and B loci go up or down together or in opposite directions, classes started over a particular time span will tend to be closer to each other than classes started at more distant times, and this will introduce a correlation. This $r^2[1]$ statistic exaggerates the effect, as each population will usually contain a few very recent LIBD classes that contribute strongly to $r^2[1]$ but

FIGURE 8. $r^2(S)$ versus $r^2(T)$ and $r^2[1]$

contribute little to r^2 because they are recent and therefore likely to have small weights.

I'm not sure that anyone has specifically written about this historical correlation, but Bill Hill in a letter to me many years ago described an effect that I think was essentially this, although I did not understand it at the time. And McVean [16] has written: "Also note that the identity coefficient approach of Sved (1971) is quite different from that presented here, because he implicitly assumes that allele frequencies remain constant over time." I suppose that he has the same idea, although I think that the key here is more that the allele frequencies are correlated rather than that they are constant.

2.8. Some concluding remarks.

It seems clear that the value of r^2 in a population has a contribution from two sources:

- (1) The size of the LIBD classes
- (2) The historical correlation

I can't presently see an easy way of seeing how these two combine to give the value of r^2 . But I also can't understand why the formula for $E(r^2)$, which ostensibly ignores the historical correlation, should give such a good agreement with the observed values. This despite the fact that there is a substantial disagreement between sampling into the LIBD

classes using the actual frequencies and the present allele frequencies ($r^2(S)$ and $r^2(T)$ of Figure 5). I'd be grateful for suggestions on these points.

3. LD UNDER A MUTATION MODEL

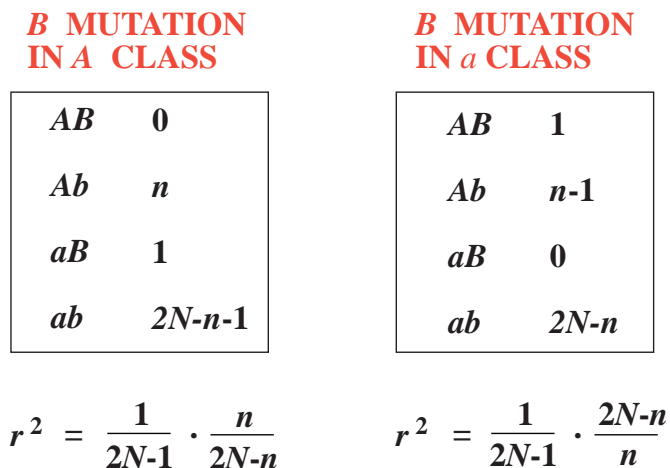
The simulations reported in Figure-2 and below are all started with high frequencies of both alleles, usually 0.5. This is of course an unrealistic starting condition, chosen partly to minimise the amount of fixation during the simulation.

3.1. LD at first appearance of a mutation.

The realistic starting condition in a neutral model is following the first appearance of an allele after a mutation event. I will assume that a new mutation occurs at the B locus. Any neutral linked segregating A locus can be assumed to be at a frequency determined by population size and mutation rate. If the mutation rate is constant, however, the frequency of mutant alleles at the A locus turns out to be very close to a reciprocal distribution, dependent on the population size and independent of the mutation rate (Fisher, 1930). In other words, the probability that an individual has n copies of the A allele is proportional to $1/n$, and can be expressed as T/n , where $T = 1/\sum(1/n)$, the summation going from $n=1$ to $n=2N-1$. The small exception to this is in the penultimate classes, those with 1 or 2 and $2N - 1$ A alleles, where the frequencies are slightly reduced below expectation.

I have simulated this situation for values of $2N=1024$ and $2N=16,384$ and $2Nu = 1$ and $1/4$, and the results have been precisely in accord with Fisher's expectation. I find Fisher's derivation entirely obscure, but it is remarkable that he should so long ago have given the solution to the distribution under what is now known as the 'infinite sites mutation model'. Fortunately more recent derivations, at least of the reciprocal distribution, are somewhat easier to follow (see eg. Ewens, 1979, Eqn 5.23).

The important conclusion from this result is that the new B mutation will usually occur in a population with few A mutant alleles and many a alleles. The overall probability that the mutant A allele will be less frequent than the a allele is

FIGURE 9. Population configuration after a single *B* mutation

$$P = \frac{\sum_{n=1}^N \frac{1}{n}}{\sum_{n=1}^{2N} \frac{1}{n}}$$

which is approximately

$$\frac{\ln(N)}{\ln(N) + \ln(2)} \quad (24)$$

which is reasonably close to 1 for large *N*.

It is now necessary to specify whether the new *B* mutation occurs in a gamete containing the ancestral *a* allele or in one containing *A*. The two possibilities are shown in Figure 9. Also shown in this figure are the values of r^2 . If *n* is small, as discussed above, the r^2 value for mutation alongside the *a* allele is small, approximately $n/(2N)^2$, while for *A* the r^2 value is relatively large, approximately $1/n$.

On average, therefore, the common *B* mutation will lead to a low value of r^2 and the rare mutation will lead to a high value. The remainder of this section is devoted to quantifying this effect. First, however, it is necessary to qualify Figure 9 in two ways. First, this formulation does not take into account the fact that *n* is sometimes high. Secondly, the question of interest is not whether the *B* mutation occurs in the *A* or *a* gamete, since in practice the mutant and ancestral alleles cannot be distinguished. Rather it should be directed at the relative contribution to r^2 of the *common* and *rare* *B* mutations.

	B MUTATION IN COMMON A LOCUS CLASS	B MUTATION IN RARE A LOCUS CLASS
MUTANT (A)	AB 0	AB 1
	Ab n	Ab n-1
	aB 1	aB 0
	ab 2N-n-1	ab 2N-n
	$\frac{T}{n} \cdot \frac{2N-n}{2N} = \frac{T}{n} - \frac{T}{2N}$	$\frac{T}{n} \cdot \frac{n}{2N} = \frac{T}{2N}$
ANCESTRAL (a)	AB 1	AB 0
	Ab 2N-n-1	Ab 2N-n
	aB 0	aB 1
	ab n	ab n-1
	$\frac{T}{2N-n} \cdot \frac{2N-n}{2N} = \frac{T}{2N}$	$\frac{T}{2N-n} \cdot \frac{n}{2N} \approx 0$
	SUM = $\frac{T}{n}$	SUM = $\frac{T}{2N-n}$

FIGURE 10. Probabilities of the various population configurations

Figure 10 extends Figure 9 to take these factors into account. First, it concentrates on 'common' and 'rare' alleles at the *A* locus, rather than specifically on *a* and *A* alleles. These headings imply that the value of *n* must lie in the range 1 to *N*. Secondly, it considers the *B* mutation in both *A* classes. The population genotypes in the two bottom populations are identical to those in the top with the *A* and *a* alleles permuted. However the frequencies of the two classes, shown in blue, are not the same - T/n and $T/(2N - n)$ respectively.

The probabilities of *B* mutation in the two *A* classes, $(2N - n)/2N$ and $n/2N$ respectively, are shown in red. These are multiplied by the respective probabilities of the *A* and *a* configurations, and then summed to give the overall probabilities T/n and $T/(2N - n)$ respectively, shown

in purple in Figure 10. These probabilities are identical to the mutant and ancestral class frequencies. It is hard to see intuitively why this should be the case.

The overall probability of mutation in the common class can be obtained by summing T/n over all classes where the B mutation occurs in the common class, ie. n in the range 1 to N . This leads to the same result as given previously (24):

$$\frac{\ln(N)}{\ln(N) + \ln(2)}$$

The more important calculation concerns the mean value of r^2 given by common and rare B mutations. Taking into account the r^2 values given in Figure 9, the mean value for common mutations is equal to

$$\begin{aligned} \frac{T}{n} &\cdot \frac{1}{2N-1} \cdot \frac{n}{2N-n} \\ &= \frac{1}{2N-1} \cdot \frac{T}{2N-n} \end{aligned}$$

while the mean value of r^2 for rare mutations is equal to

$$\begin{aligned} \frac{T}{2N-n} &\cdot \frac{1}{2N-1} \cdot \frac{2N-n}{n} \\ &= \frac{1}{2N-1} \cdot \frac{T}{n} \end{aligned}$$

Thus the values of r^2 are identical to, but reversed from, the probabilities of the two mutation classes, in each case multiplied by the factor $1/(2N-1)$. The rare mutations contribute more to r^2 . Overall the contribution from rare mutations is again equal to

$$\frac{\ln(N)}{\ln(N) + \ln(2)}$$

multiplied by the factor $1/(2N-1)$. The overall sum of r^2 from both common and rare mutations is equal to $1/(2N-1)$ (Ohta and Kimura, 1969) [18].

3.2. Subsequent generations.

The following calculation deals with just the simplest possible case, the first generation of buildup of LD when there is no recombination. The simplification here is that there are only three possible genotypes. It is then convenient to describe the possible offspring populations in terms of n_A and n_B , the numbers of A and B alleles respectively. The left

	<i>B</i>	<i>b</i>	
<i>A</i>	.	n_A	n_A
<i>a</i>	n_B	$2N - n_A - n_B$	$2N - n_A$
	n_B	$2N - n_B$	$2N$

	<i>B</i>	<i>b</i>	
<i>A</i>	n_B	$n_A - n_B$	n_A
<i>a</i>	.	$2N - n_A$	$2N - n_A$
	n_B	$2N - n_B$	$2N$

$$r^2 = \frac{n_A}{2N - n_A} \cdot \frac{n_B}{2N - n_B}$$

$$r^2 = \frac{2N - n_A}{n_A} \cdot \frac{n_B}{2N - n_B}$$

FIGURE 11. Genotypes following one Wright-Fisher generation

side of Figure 11 shows the common mutation, where the *AB* class is absent. Since the initial frequencies of the *Ab*, *aB* and *ab* classes in the parent population are $n/2N$, $1/2N$ and $1 - n/2N - 1/2N$ respectively, the probability of obtaining the genotype configuration of the offspring population is

$$\frac{2N!}{n_A!n_B!(2N - n_A - n_B)!} \left(\frac{n}{2N}\right)^{n_A} \cdot \left(\frac{1}{2N}\right)^{n_B} \cdot \left(1 - \frac{n}{2N} - \frac{1}{2N}\right)^{2N - n_A - n_B}$$

This expression needs to be multiplied by the associated value of r^2 , $\frac{n_A}{2N - n_A} \cdot \frac{n_B}{2N - n_B}$ (Figure 11). The expected value of r^2 is then given by summing this quantity over all possible values of n_A and n_B .

There is no exact simplification of this expression containing terms in $2N - n_A$ and $2N - n_B$ in the denominator. However if each of these terms is replaced by $2N$, which leads to only a small underestimation, the expression simplifies to $n/(2N)^2$, approximately the same value as found in the initial generation in Figure 9.

This result suggests that there is no increase in the value of r^2 . However this is misleading, because the result is averaged over all populations, including the case of $n_A = 0$ and $n_B = 0$. The values of r^2 shown in Figure 11 for this case are equal to zero. In reality, the values are undefined, since the derivation of r^2 for $n_A = 0$ or $n_B = 0$ involves a division of zero by zero. The true mean value of r^2 needs to be divided by the probability of obtaining unfixed populations. The probabilities of non-zero values of n_A and n_B are dominated by the *B* probability, since there is initially only one *B* mutation. The probability of fixation after one generation at the *B* locus is approximately e^{-1} , so that

the value of r^2 amongst unfixed populations is increased by the factor $1/(1 - e^{-1})$.

The equivalent result for the case of a B mutation in the rarer A genotype can be calculated in the same way. The r^2 value in this case is equal to

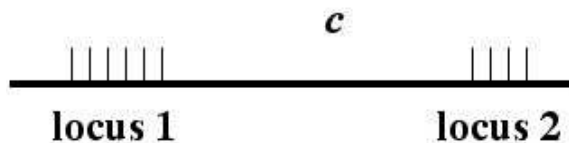
$$\frac{2N - n_A}{n_A} \cdot \frac{n_B}{2N - n_B}$$

As previously, $2N - n_A$ and $2N - n_B$ are each replaced by $2N$. In addition, n_A in the denominator needs to be replaced by its expected value n , which may involve a somewhat higher degree of approximation. With this substitution, the sum simplifies to $1/n$, which is again approximately the value as found in the initial generation in Figure 9. As previously, the sum needs to be corrected for unfixed populations by dividing by the factor $1 - e^{-1}$.

In summary, the r^2 values for both classes of population are expected to increase simply by the factor representing the probability that fixation has not occurred. The following section presents computer simulation to test this expectation.

3.3. Computer simulation.

I have written a Monte-Carlo simulation program to generate new mutations at an infinite number of sites. Essentially this means an infinite number of sites at each of two loci as shown below (there are actually three loci, in order to check on some 3-locus statistics but this is not relevant here):



New mutations are generated randomly, generally one per generation. Each new mutation starts a new site, randomly chosen from one of the two loci. Because of the finite size of the population, sites are regularly lost, thereby preventing the number of sites from increasing beyond limit. This does involve renumbering of sites each generation. Sites are lost when the new mutation is either lost or fixed in the population. The rate of fixation per generation is monitored to make sure that it agrees with the rate of mutation per individual (Kimura,

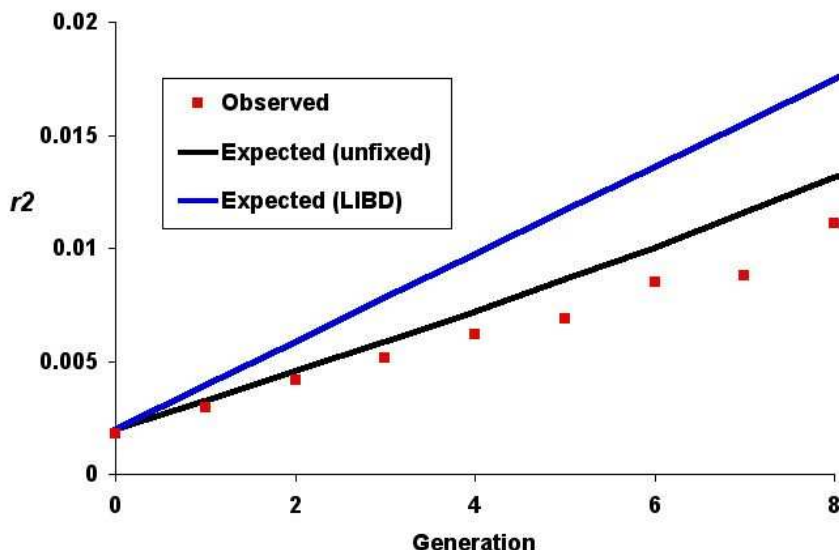


FIGURE 12. Observed and expected values of r^2 over 8 generations following production of a new mutant. Population size (N) is 256.

1970). At certain intervals, the value of r^2 is calculated and the results are cumulated.

Two graphs showing observed and expected are in Figure 12 and Figure 13. Note that the scale of r^2 values is very different in the two cases.

The expectations labeled Expected (unfixed) are calculated as follows. The expected value in generation 1 is taken as $1/(2N-1)$. Thereafter the expected values are calculated by dividing this figure by the proportion of populations left unfixed. The values shown as Expected (LIBD) are calculated using the LIBD-derived recurrence relationship (7).

The calculated (observed) values from the simulation lie slightly below their expectation in each of the two graphs. This is evidently because of the approximations in the trinomial calculation. However the expectations calculated by correcting for unfixed populations are clearly much closer to the observed values than the expectations from the LIBD calculation.

Figure 14 shows the results after many generations. An intermediate value of N (1024) is used here, and the simulation is extended until near complete fixation. The expected value in this case comes just

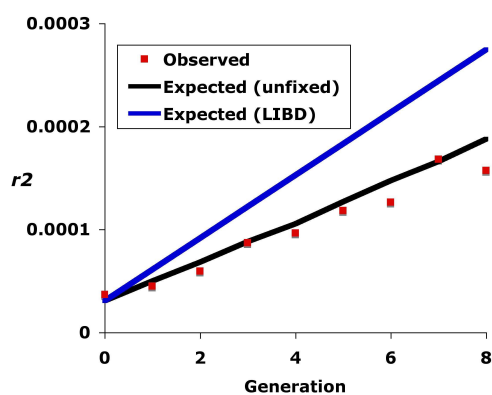


FIGURE 13. Observed and expected values of r^2 over 8 generations following production of a new mutant. Population size (N) is 16,384.

from the LIBD equation. Expectations given by correcting for unfixed populations become increasingly unreliable after the first few generations.

The general shapes of the observed and expected curves are similar, although by no means identical. As argued above in connection with figures 12 and 13, agreement is not even expected in the early stages, where fixation dominates the process. Somewhere after the mutation reaches a sufficiently high frequency, the LIBD-derived expectation becomes more accurate than the fixation expectation. It appears, though, that there is a second fixation-based discrepancy between observed and expected values at the high end of the range where fixation is almost complete, which the LIBD expectation does not take into account.

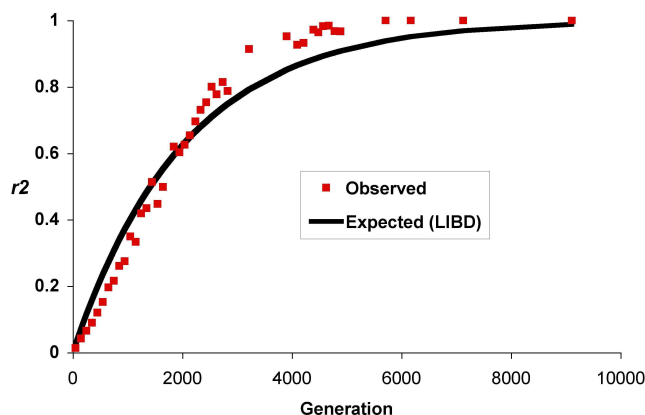


FIGURE 14. Observed and expected values of r^2 over 10,000 generations following production of a new mutant. Population size (N) is 1024. Observed values averaged over 100 data points to reduce chance fluctuations.

4. SELECTION AND LD

Hundreds of papers have been written on this topic, reflecting the range of possible selection models. I have attempted a brief overview of the various models.

4.1. Two-locus models.

Four selection models are considered. With two linked loci there are 10 possible combinations, as shown below. As written, the positive selection model arbitrarily assumes no dominance and the mutation-selection balance model assumes dominance, although these are not necessary features. Asterisks denote the combinations discussed below, red ones being where my work fits in.

Lines (3) and (4), the heterozygote advantage and mutation-selection models, basically summarise the 'balanced' and 'classical' world views. Under the balanced view, selection basically maintains diversity. Under the classical view, selection acts to purify the genome, opposing the deleterious effects of mutation. Each of these models postulates the existence of hundreds or thousands of loci, many necessarily closely linked to each other. These opposing views were highlighted in Lewontin's influential book [15], although I don't see much current interest in the argument.

	$\underline{A_1 A_1}$	$\underline{A_1 A_2}$	$\underline{A_2 A_2}$
(1) Neutral	1	1	1
(2) Positive selection	1	$1 + \frac{s}{2}$	$1 + s$
(3) Heterozygote advantage	$1 - s_1$	1	$1 - s_2$
(4) Mutation-selection balance	1	1	$1 - s$
	and $A_1 \xrightarrow{u} A_2$		

Second Locus

	Neutral	Positive	Het. adv.	Mut-selec.
(1) Neutral	Finite-size LD			
(2) Positive	Hitch-hiking *	Hill-Robertson *		
(3) Het. adv.	Associative overdom. *	Selection opposition *	Associative overdom.	
(4) Mut-selec.	Background selection *	Background selection? *	?	Background selection

The discussion here is mainly restricted to the interaction of finite-size LD and selection, although much of the literature is not couched in these terms. There is a large literature pre-dating the recognition of finite-size LD, much to do with the evolutionary advantages and disadvantages of recombination. Felsenstein's review paper [3] gives a succinct summary of this work.

4.2. Associative overdominance.

My papers were based on the assumption of widespread heterozygote advantage in the genome. The 'balance school' believed that some sort of balancing selection was needed to explain the large amount of variation at the molecular level that was being uncovered. The model of heterozygote advantage, particularly the model of symmetrical heterozygote advantage in which heterozygotes were at an equal advantage to both homozygotes, was a convenient one for calculations on

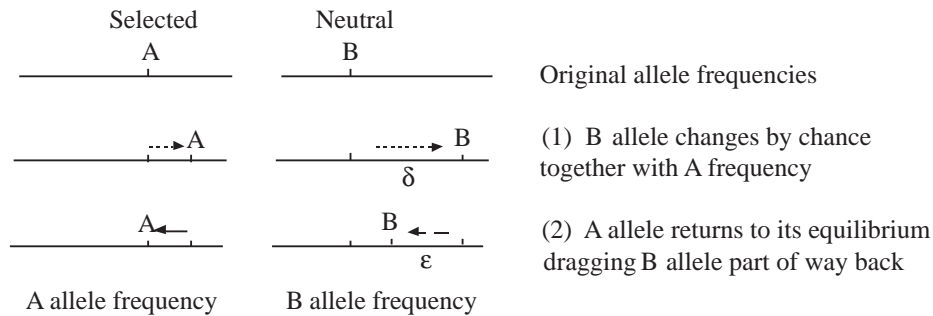


FIGURE 15. Two stages in the interaction of LD and drift

this model. There was a plausible rationale for this model in terms of potential superiority of hybrid enzymes. I'm not sure nowadays that I would be so enthusiastic about the model, although the experiments summarised in the chapter of *Drosophila* fitness of this PIFFLE aren't exactly supportive of the classical model either.

Referring to the 100 chromosome simulation given in Section 1.2 of this chapter, it is clear there are potentially large selective consequences of LD. In the final population of that simulation, every individual will be either totally homozygous or totally heterozygous. Thus there is a very strong reinforcing effect. Essentially the heterozygote advantages at different loci cumulate, so that at each locus the advantage is many times what it would have been with linkage equilibrium.

The situation with many linked loci is, of course, complex. The simplest model showing the effects of linkage and selection is the two locus model, with one selected locus and a closely linked neutral locus.

4.3. Stabilising effect of a selected locus on a neutral locus.

Consider alleles A and B with some level of LD between them. The A locus is assumed to be at a selective equilibrium because of heterozygote advantage while the B locus is selectively neutral, while. Figure 15 shows the expected consequences of selection and LD:

(1) Suppose that the frequency of the allele B at the neutral locus changes by chance by some fraction δ . Because of the change in frequency at the B locus, and the fact that there is non-zero LD between the two loci, the frequency at the A locus must also change. If d is positive then the frequencies of the two genes will change in the same direction, as pictured in the diagram above, but the effect is exactly equivalent if d is negative and the A and B alleles change frequencies in opposite directions.

(2) Selection will drive the A locus back to its equilibrium. It seems clear that will lead to a directed change at the B locus back towards its original frequency. The magnitude of this change is ϵ . In the diagram, the dotted arrows represent chance fluctuation, the solid arrow represents direct selection, and the dashed arrow represents the effects of LD and selection.

The theory derived in [24] says that the expected value of ϵ/δ is approximately $d^2/p_A(1-p_A)p_B(1-p_B)$. In other words selection tends to reduce the magnitude of fluctuations at the neutral B locus by this fraction, although this simplifies by pretending that it all happens in one generation. I didn't realise at the time that this fraction is equal to r^2 , the square of the correlation of frequencies between the two loci.

4.4. The apparent selective value.

A second way of looking at the situation is as follows. The values below show the selective values at the A locus and at the B locus. The values at the A locus are written this way to illustrate heterozygote advantage, implying $s > 0$ and $t > 0$.

AA	Aa	aa	BB	Bb	bb
$1 - s$	1	$1 - t$	S_{BB}	S_{Bb}	S_{bb}

The B locus is selectively neutral. However because of the association with the A locus, this is not how it appears. Homozygous genotypes at the B locus will tend to be associated with homozygous genotypes at the A locus. Therefore they will appear to be at a disadvantage to the heterozygote.

The values are as follows. The A locus frequency does not enter into the formulae since the A locus is assumed to be at equilibrium:

$$S_{Bb} - S_{BB} = \frac{d^2(s+t)}{p_B^2 p_b} \quad (25)$$

$$S_{Bb} - S_{bb} = \frac{d^2(s+t)}{p_B p_b^2} \quad (26)$$

Thus the heterozygote at the B locus will appear to have a selective advantage over both homozygotes. Furthermore, $(S_{Bb} - S_{BB})/(S_{Bb} - S_{bb}) = p_b/p_B$, which is exactly the condition required for equilibrium at the B locus. Therefore it looks as though the alleles at the B locus

are at a selective equilibrium with the heterozygote at an advantage, even though in reality the locus is neutral.

The same conclusions can be drawn from each of the above two results. The B locus is at what may be termed a 'pseudo-equilibrium'. Selection will tend to oppose any change from that frequency, and alleles at the locus will look as if they are at equilibrium. Over a period of time, however, frequencies may change to a new value. Selection will then oppose the change from this new frequency, and the locus will still appear to be at selective equilibrium.

I didn't think of giving a name to this phenomenon. Ohta and Kimura [19] independently (but a little later) derived similar results. They termed the phenomenon 'Associative Overdominance', which was actually a term coined previously by Frydenberg (1963) [6]. I had seen the term previously but had not thought it quite appropriate to this case. Anyway it taught me a lesson that the first thing one should do when finding a result is to find a name for it.

4.5. The combination of balancing and positive selection.

Two rather different scenarios can be envisaged for this situation. One refers to Darwinian selection, when a new favourable mutation arises. If a closely linked locus is held at equilibrium by balancing selection, substitution of the favoured gene may be retarded.

The second scenario refers to artificial selection for a quantitative trait. This is the situation that I considered in [30]. There are differences between the scenarios. First, the Darwinian natural selection would usually involve a single gene, whereas the artificial selection may involve many genes. In addition, the fact that there is a single initial occurrence of the favoured gene means that there must be some degree of initial LD, which influences the process. Selection for a quantitative trait influenced by many genes implies that some or all of these genes are polymorphic in the population. If there is no initial LD, balancing selection will have no effect (see below). So the process relies on finite-size LD.

By arguments similar to those in Section 4.1 for associative overdominance, it seems that LD should lead to natural selection opposing the effects of artificial selection. The only difference is that the change in gene frequency in Figure 15 is due to chance in the unselected case and due to positive selection in the current case.

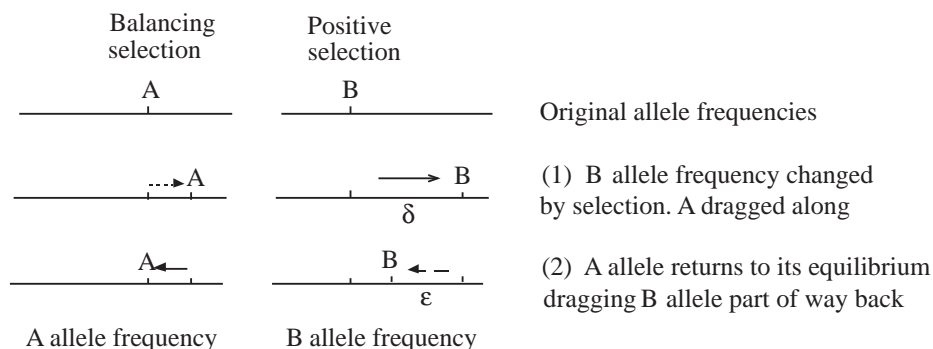


FIGURE 16. Balancing selection opposes positive selection

There is additional factor in the case of positive selection. Selective values at each locus of the model are as follows. It is convenient to start by assuming that the selection at the A (balanced selection) is symmetrical against the two homozygotes. Selective values of AB genotypes are obtained by multiplying the A and B selective values. Note that the loci have been reversed compared to [30] for consistency with the model of Section 4.1.

$$\frac{A_1A_1}{1-s} \quad \frac{A_1A_2}{1} \quad \frac{A_2A_2}{1-s} \quad \frac{B_1B_1}{1} \quad \frac{B_1B_2}{1+t/2} \quad \frac{B_2B_2}{1+t}$$

The opposition of selection can be studied by looking at the covariance of selective values. Doing the equivalent calculation to that given in [30], this comes to

$$std(p_1 - p_2)$$

where d is the coefficient of linkage disequilibrium and p_1 and p_2 are the A locus allele frequencies.

Looking at Figure 16, it can be seen that d and $p_1 - p_2$ will generally be of opposite sign. If d is positive, then an increase in the frequency of B_2 will lead to an increase in the frequency of A_2 , making $p_1 - p_2$ negative. Similarly if d is negative, the frequency of A_1 will tend to increase. Overall, therefore, there is a negative covariance. A more precise algebraic argument can be given for to show that the quantity $d(p_1 - p_2)$ is decreased by selection, and the argument can be extended to asymmetric selection at the B locus.

The opposition of natural selection to artificial selection therefore depends on the existence of LD. If $d = 0$, no effect is expected. Only if sufficient LD is generated by chance will there be an effect. So it is

not clear how significant this effect will be. Computer simulation of multiple locus models was used to approach the problem.

4.5.1. Computer simulation.

Simulations were done with a chromosome of 50 map units, with 12 quantitative loci interspersed with 96 loci with heterozygote advantage, either randomly or equally distributed along the chromosome. All runs were started with heterotic loci at 50% frequency, expected to give the maximum retarding effect, and a range of initial frequencies at the quantitative loci. Maximum and minimum levels of initial LD were simulated. A population size was 50 ($2N = 100$).

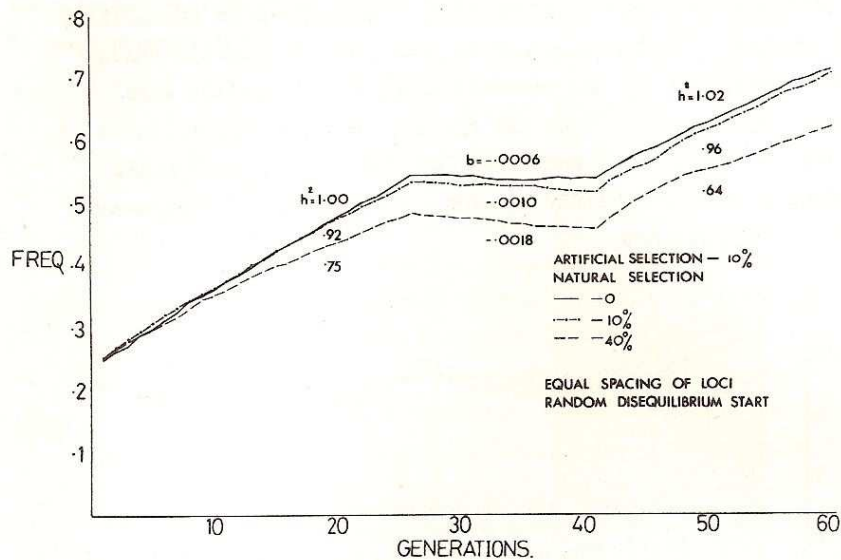


FIGURE 17. Computer simulation of the opposition of natural selection

Results in all cases showed low levels of retardation of the heterotic loci. Despite the larger numbers of stabilising loci held at intermediate frequencies, the quantitative loci seemed able to thread their way through. Figure 17 shows some results, averaged over 25 replicate runs. The simulations were based on 25 generations of selection, 15 generations without artificial selection to see if there was regression of the quantitative character, followed by 20 generations of selection. Two levels of natural selection were used, but even if there was a four-fold greater level of selective intensity devoted to balancing selection compared to positive selection, the losses in selective gain were modest.

All runs showed some regression to the initial value after artificial selection was relaxed, including the case of no natural selection. I believe that this apparently counter-intuitive result can be attributed to the fact that truncation selection induces an additive x additive interaction, whose gain is lost when recombination reassorts the genes during the period of no selection, a result predicted by Griffing (1960) [7].

I presented the above results at a quantitative genetics conference in 1977. The paper was published in the proceedings, although in retrospect I should have tried to publish it in a refereed journal. At the same conference, Alan Robertson tackled more or less the same problem. Fortunately our conclusions were similar regarding the low interference due to natural selection.

Alan carried out simulations in a novel way. Rather than by just simulating lots of linked loci on a chromosome, he kept track of specific segments, subdividing them each time a new crossover occurred. Although segments are sometimes lost, the number of segments keeps climbing to a rather high number over time. However if the computing power available at the time could cope with this number, I have often felt that with the currently available memory sizes I should also try to write forward simulation programs in this way.

4.6. Hitchhiking.

During Darwinian substitution of a gene, the regions surrounding the favoured gene show reduced heterozygosity, the opposite of what is expected with heterozygote advantage. The theory behind this is due originally to Maynard Smith and Haigh (1974). Although the theory was put forward without specific reliance on LD, as pointed out above there is a necessary generation of LD when a single favourable mutation arises.

Hitch-hiking has turned out to be of much more interest than associative overdominance, as it has become the main way of detecting Darwinian selection in evolution, particularly in human evolution. I wish I hadn't been so hung up on the heterozygote advantage model. The current theory is somewhat more sophisticated than the original hitchhiking theory, depending on multiple locus haplotypes rather than two locus haplotypes [22].

4.7. The Hill-Robertson effect.

As I understand it, the Hill-Robertson effect refers to the tendency of selection to lead to sub-optimal genotypes owing to interference caused

by LD. The name, originally popularised by Felsenstein in his review [3], is sometimes used more broadly to describe all finite-size LD effects. The theory was developed by Hill and Robertson (1966) [11], perhaps before they had clarified their ideas on what effects were due to selection and what to finite size [12]. The Felsenstein paper includes the following acknowledgment "I wish to thank Drs. A. Robertson and W. G. Hill for having the patience to re-explain their work to me as often as I asked them", which I think emphasises the complexity of the problem, and of teasing out what aspects are due to drift and what to selection.

4.8. Background selection.

This refers to the effect of linkage to deleterious genes maintained in the population by mutation. The theory was originally put forward by Charlesworth and colleagues [1], again without specific mention of LD.

One of my publications [26] relates to this topic. It is a paper of which I am not particularly proud. Like the model of Charlesworth et al, it assumes deleterious selection at one locus and selective neutrality at the second:

AA	Aa	aa	BB	Bb	bb
1	$1 - hs$	$1 - s$	S_{BB}	S_{Bb}	S_{bb}

This leads to 'apparent selective values' as follows:

$$S_{Bb} - S_{BB} = \frac{sd}{p_B^2 p_b} [-x_1 h - x_3 (1 - h)]$$

$$S_{Bb} - S_{bb} = \frac{sd}{p_B p_b^2} [x_2 h + x_4 (1 - h)]$$

where d is the usual linkage disequilibrium coefficient and x_1, x_2, x_3 and x_4 are the frequencies of the haplotypes AB, Ab, aB and ab respectively.

For values of h in between 0 and 1, one of these will be positive and one negative depending on the sign of d . Thus there will be directional selection at the B locus depending on which allele is associated with the favoured A allele.

However when one adds up the two equations:

$$2S_{Bb} - S_{BB} - S_{bb} = \frac{d^2 s}{p_B^2 p_b^2} (1 - 2h).$$

So if there is *any* degree of dominance, ie $h < 0.5$, the heterozygote will be have an advantage on average. This won't matter if there is only one selected locus. However if a neutral locus is linked to many such loci, with some positive and some negative d values, the heterozygote will be at a selective advantage, tending to stabilise the frequency.

I found this result quite early on during LD-selection calculations but forgot about it in favour of the much more readily interpretable equations 25 and 26. Following Ohta's 1971 paper on associative overdominance due to linked detrimental genes [17] I somehow felt obliged to come back to this result. What's worse, is that I then did a whole lot of computing to show that there was a stabilising effect, by setting up a model of many linked deleterious recessive genes all at 50% frequency. I seem not to have been worried about the question of how all these deleterious genes were supposed to get to 50% frequency.

Fortunately this paper [26] was generally ignored. However it and Ohta's paper were picked up by Palsson and Pamilo (1999) [20] who pointed out the contradiction between the stabilising effect of this model, and the effect of background selection in increasing the fixation rate of neutral genes. These authors claim that the value of Nhs is critical in determining whether selection will be stabilising (low Nhs) or destabilising (high Nhs).

One area where I feel that deleterious genes ought to be equivalent to heterozygote advantage is in slowing the rate of Darwinian substitution. Although this is a small effect (see Section 4.5), and I haven't specifically looked at the deleterious gene model, it seems intuitively that in a model with many deleterious genes, any new mutation will inevitably be linked to some, thereby slowing the rate of substitution.

5. A MEASURE OF OVERALL LD

I found one more result in my 1968 paper [24] that has had some application. If there are many linked loci, as is the situation in real life, there will be so many pairs of loci that summarising the overall level of LD is difficult. Looking at the simple simulation earlier in this chapter, even with a relatively small number of loci it is not easy to summarise the overall LD.

One quantity which seemed promising is the variance of heterozygosity in a random mating population. Looking at the initial population of the simulation at the beginning of this chapter, most individuals (pairs of chromosomes) will have a similar heterozygosity. In this situation, the variance between individuals of heterozygosity will be close to zero.

Looking at the final population, each individual is either totally homozygous or totally heterozygous. The variance of heterozygosity V_H is therefore at a maximum, corresponding to the increase in LD. I wondered whether there was a simple relationship between V_H and the sum of the values of d^2 . I did some computer calculations that showed that there must be a simple relationship, and then some long-winded algebra, the details of which I have long forgotten, that confirmed this.

I mentioned the result around the place, and a day later Sam Karlin, who loved to show that other people's results were trivial, showed that my long-winded calculations could indeed be replaced by a trivial calculation. What Sam noted was that the variance of the sum of heterozygosities could be written in the form:

$$V_H = V(H_1 + H_2 + \dots + H_k)$$

where H_1, H_2 etc are the individual heterozygosity. This could then be replaced by

$$V_H = \sum_i V(H_i) + \sum_i \sum_j Cov(H_i, H_j)$$

Substituting for the variance and covariance terms led easily to the relationship:

$$V_H = 8 \sum_{i,j} d_{ij}^2 + 16 \sum_{i,j} d_{ij}(0.5 - p_i)(0.5 - q_j) + 2 \sum_i p_i q_i (1 - 2p_i q_i)$$

Summation is over all pairs of loci for the first two terms and over all loci for the third. The important term is the first term, to which all pairs of loci contribute. The second term is usually less important, because positive and negative D values tend to cancel out. Finally the last term is independent of the amount of LD, essentially a correction for the amount of heterozygosity.

The V_H calculation was quite advantageous when I was doing multiple locus simulations in 1966 and 1967. With the computers available at the time, it was quite expensive to calculate all the d^2 terms in each

generation, since with a few hundred loci, the number of d^2 terms to be calculated was many thousands.

Even though calculation of many d^2 values is now trivial with modern computers, it is still sometimes advantageous to have a single multi-locus measure. The use of V_H as such a measure has been picked up by others, particularly since the calculations have been implemented in the computer program LIAN by Haubold and Hudson (2000) [9]. The fact that I first suggested this measure has, somewhat to my chagrin, been lost somewhere along the line.

REFERENCES

- [1] B Charlesworth, M T Morgan, and D Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303, 1993.
- [2] J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Harper & Row, New York, 1970.
- [3] J Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974 Oct.
- [4] RA Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p . *J. Roy. Statist. Soc.*, 85, 1922.
- [5] RA Fisher. *The theory of inbreeding*. Oliver and Boyd, Edinburgh, 1949.
- [6] O Frydenberg. Population studies of a lethal mutant in *Drosophila melanogaster*. i. behaviour in populations with discrete generations. *Hereditas*, 48, 1963.
- [7] B Griffing. Theoretical consequences of truncation selection based on the individual phenotype. *Aust J Biol Sci*, pages 307–343, 1960.
- [8] JBS Haldane. The mean and variance of χ^2 , when used as a test of homogeneity, when expectations are small. *Biometrika*, 31:346–360, 1940.
- [9] B Haubold and RR Hudson. Lian 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics*, 16:847–849, 2000.
- [10] P. W. Hedrick. Gametic disequilibrium measures: proceed with caution. *Genetics*, 117:331–341, 1987.

- [11] W G Hill and A Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8:269–294, 1966.
- [12] WG Hill and A Robertson. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, 38:226–231, 1968.
- [13] R R Hudson. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics*, 109:611–631, 1985.
- [14] M. Kimura and J. F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49, 1964.
- [15] R. C. Lewontin. *The genetic basis of evolutionary change*. Columbia U.P., N.Y., 1974.
- [16] GAT McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162:987–991, 2002.
- [17] T Ohta. Associative overdominance caused by linked detrimental mutations. *Genetical Research*, 19:277–286, 1971.
- [18] T. Ohta and M. Kimura. Linkage disequilibrium due to random genetic drift. *Genet.Res.*, 13:47–55, 1969.
- [19] T. Ohta and M. Kimura. Development of associative overdominance through linkage disequilibrium in finite populations. *Genet Res*, 16(2):165–177, 1970.
- [20] S Palsson and P Pamilo. The effects of deleterious mutations on linked, neutral variation in small populations. *Genetics*, 153(1):475–483, 1999 Sep.
- [21] C. Sabatti and N. Risch. Homozygosity and linkage disequilibrium. *Genetics*, 160:1707–1719, 2002.
- [22] PC Sabeti, P Varilly, B Fry, and etal. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918, 2007 Oct 18.
- [23] P Stam. The distribution of the genome identical by descent in finite random mating populations. *Genet Res*, 35:131–135, 1980.
- [24] J. A. Sved. The stability of linked systems of loci with a small population size. *Genetics*, 59:543–563, 1968.
- [25] J. A. Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol*, 2:125–141, 1971.

- [26] J. A. Sved. Heterosis at the level of the chromosome and at the level of the gene. *Theor Popul Biol*, 3(4), 1972.
- [27] J A Sved. Correlation measures for linkage disequilibrium within and between populations. *Genet Res*, 91:183–192, 2009.
- [28] J. A. Sved and M. W. Feldman. Correlation and probability methods for one and two loci. *Theor Popul Biol*, 4:129–132, 1973.
- [29] J. A. Sved, T. E. Reed, and W. F. Bodmer. The number of balanced polymorphisms that can be maintained in a natural population. *Genetics*, 55:469–481, 1967.
- [30] JA Sved. Opposition to artificial selection caused by natural selection at linked loci. In O Kempthorne, editor, *Proceedings of the International Conference on Quantitative Genetics*, pages 435–456, Ames, Iowa, 1977. Iowa State University Press.
- [31] John A Sved. Linkage disequilibrium and its expectation in human populations. *Twin Res Hum Genet*, 12(1):35–43, 2009 Feb.
- [32] A. Tenesa, P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and P. M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Res*, 17:520–526, 2007.
- [33] S Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.