LINKAGE DISEQUILIBRIUM (LD)

Author′s note to anybody reading this document for information about Linkage Disequilibrium.

You will find this document full of statements such as
"I did this .. I did that .. blah blah".
I apologise for this, but the reason is that this document is one chapter of my PIFFLE (www.handsongenetics.com/PIFFLE) written to summarise genetics projects that I have been involved in. It′s supposed to be a personal account, and so is inappropriately pretentious. When writing, I did not anticipate that this discussion of LD might be found from a Google search, independently from the introduction and overall web document.

The last two sections contain largely unpublished material. I would be very pleased if anybody is interested in following up on any of this. It is a consequence that I did not anticipate, that putting such unreviewed material online might deter the publishing of related work. I will not be trying to publish any of this material, and would be disappointed if its presentation in this way deters anybody from working and publishing on these topics. You could maybe add an acknowledgement (J. Sved, pers. comm.) should you use anything. Please write if in doubt.
Updated Dec 14, 2018

0.1. Preface 2
1. LD THEORY 3
1.1. Early theory 3
1.2. Fifty years of LE and fifty years of LD 4
1.3. Measuring LD in finite populations 8
1.4. Measures of LD 8
2. SOME RECENT PUBLICATIONS 11
2.1. LD and the estimation of N_e 11
2.2. LD in subdivided populations 12
2.3. When did European and African populations split? 14
2.4. LD between blocks of loci 15
3. SELECTION AND LD 18
3.1. Two-locus models 18

3.2. Associative overdominance 19
3.3. Stabilising effect of a selected locus on a neutral locus 20
3.4. The apparent selective value 21
3.5. The combination of balancing and positive selection 22
3.6. Hitchhiking 25
3.7. The Hill-Robertson effect 25
3.8. Background selection 26
4. LINKAGE DISEQUILIBRIUM (LD) AND LINKED IDENTITY BY DESCENT (LIBD) 28
4.1. Introduction 28
4.2. An aside on the effect of fixation 29
4.3. Argument #1, the basic justification for the LIBD method 30
4.4. LIBD with loose linkage 34
5. MORE ON LIBD 37
5.1. Argument #2 - LIBD and homozygosity 38
5.2. The length of identical segments 43
5.3. Argument #3 - Sampling into LIBD classes 44
5.4. LIBD computer simulation 49
6. LD UNDER A MUTATION MODEL 57
6.1. LD at first appearance of a mutation 57
6.2. Subsequent generations 60
6.3. Computer simulation 62
References 64

Contents

0.1. **Preface.**

LD is a huge topic that has expended enormously in the last decade or two. What started out as an obscure topic of interest to only a few theoretical population geneticists has turned out to have important applications, particularly in gene mapping. A 2018 review by Bill Hill and myself in Genetics Perspectives [39] summarises some of this history. The genesis of this review was the recognition that the initial calculations on LD came 100 years earlier (Robbins, 1918), while 50 years later Bill and I independently published papers (Hill and Robertson, 1968; Sved, 1968) noting that LD is expected to be ubiquitous due simply to population structure. The important applications in gene mapping, which Bill knows much more about than I do, depend on this finding, although perhaps only in a rather trivial way.

This document concentrates on those parts of LD theory that I have been involved in. It's rather long, because later sections go into considerable detail on unpublished, and probably unpublishable, work.

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 multiplied by 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 [3] 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

$$p'_{AB} = cp_A p_B + (1-c)p_{AB}$$

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' (LE), 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.

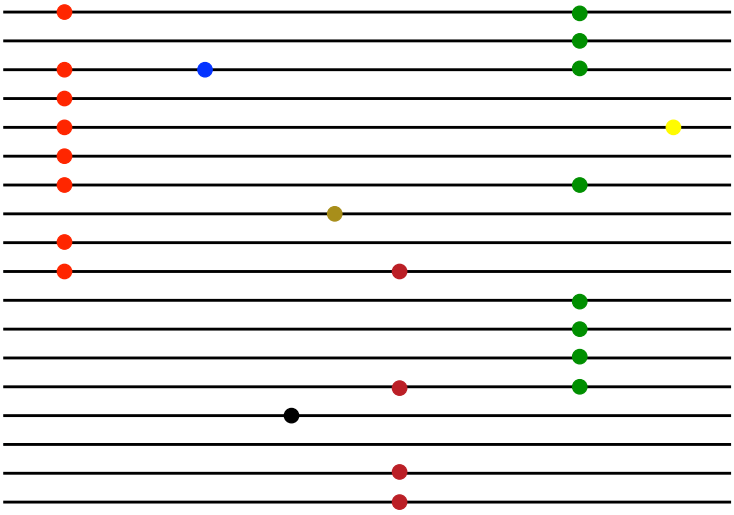
1.2. **Fifty years of LE and fifty years of LD.**

I'm using this title to describe what was formalised in the papers of Hill and Robertson (1968), Sved (1968) and Ohta and Kimura (1969). Suddenly it became clear that LD, rather than a rare event, had to be everywhere. One has to go back a year or two, however, to the papers of Lewontin and Hubby (1966) and Harris (1966) to understand the background to this paradigm shift.

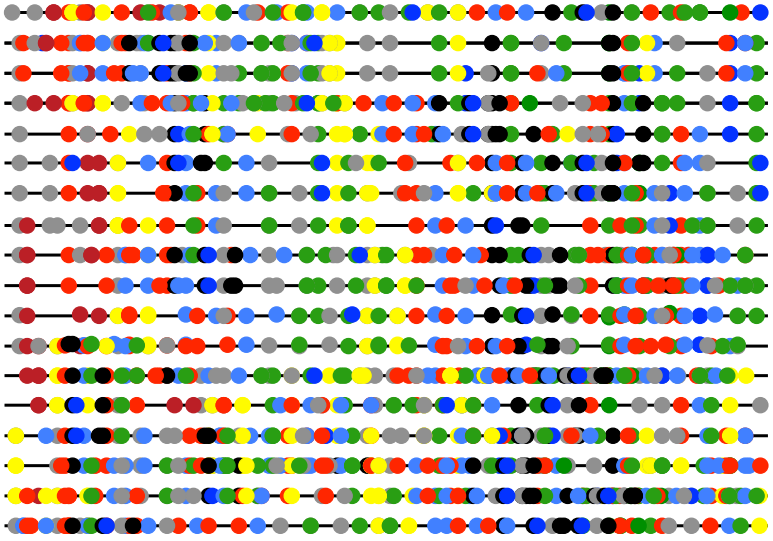
Prior to Lewontin, Hubby and Harris, I don't believe that much thought had been given to quantifying how much variability there is in natural populations. The underlying DNA structure of chromosomes, with their millions of bases, had been known for some time. But little attention had been paid to how many of these bases were polymorphic in actual populations. I suspect that if you had asked somebody to quantify what a population of chromosomes looks like, they would have come up with something like the first of the two diagrams below.

What Lewontin and Hubby did was to follow a bunch of enzymes that could be visualised on a gel, and to quantify how many had detectable variation in a Drosophila population. They found that around one third were polymorphic. Harris found a similar figure in human enzymes. So,

What a population looks like (pre-molecular era)
(single chromosome)



What a population looks like (post 1966)



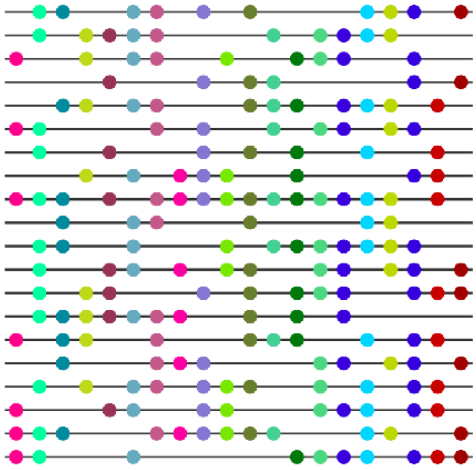
suddenly, it was clear that a more realistic picture of what a population looked like was the second figure. The reality at the DNA level of millions of polymorphic sites on a chromosome makes it impossible to picture a whole chromosome. Clearly the opportunity for LD between closely linked sites is vastly higher in the second figure.

I must have started to think about populations of this type sometime after the Lewontin and Hubby paper, although I don't recall the circumstances. But I suddenly realised that there was an enormous effect that was being ignored in the 1918 argument predicting that LD would be systematically decreased over generations. 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.

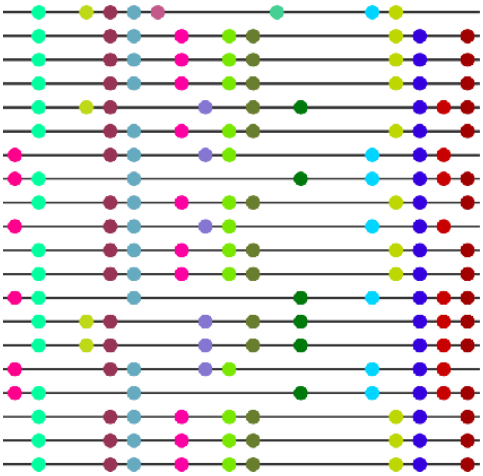
I'll show a small simulation that illustrates the effect. Suppose, for example, that a population of 20 chromosome types was started in linkage equilibrium. Then the chromosomes might be something like the first of the three diagrams below. Generations are then simulated under the usual Wright-Fisher model assuming no recombination. After 10 generations, many of the initial chromosome types have been lost, and after 20 generations the population is down to two types. The diagram is simplified by omitting colours of the sites that are fixed and thus add nothing to the LD statistics.

The amazing thing about the population at this last pictured 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 simple thought experiments like this that there is a very striking tendency for closely linked genes to become associated if the population size is finite. The expectation that closely linked genes will be in linkage equilibrium, coming from the 1918 infinite population calculations, totally misses this point.

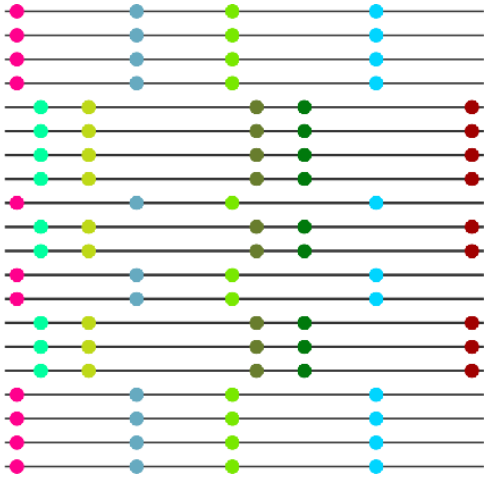
A short section of a random population – no LD



After 10 gens. of random mating – intermediate LD



After 20 gens. of random mating – complete LD



1.3. **Measuring LD in finite populations.**
As a starting point in [30], I tried to calculate the expected amount of LD for given values of c, the recombination rate, and N, the population size. In this calculation, I was thinking in terms of loci held at a fixed frequency 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

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

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 [16] 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 [30] 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.

I managed to come up with a reasonably simple expectation for the equilibrium value of r^2, essentially

$$E[r^2] \approx \frac{1}{1+4Nc} \tag{3}$$

This was derived using a probability method (LIBD) that is simple in application but hard to justify. Two long sections are included on this topic, later in this chapter.

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 [12].

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 logical to use a statistic for LD that allows the full range from -1 to +1, 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 more then 100 years, so it's not really possible to change that. It's easy occasionally to get confused between the two.

While on this semantic diversion, I'd also like to comment on the more fundamental issue of the terms 'gene' and 'allele'. I probably use them uncritically, sometimes exchangeably. There seems to be a trend to

avoid the use of the term 'gene' altogether, given the difficulty of defining exactly what is and is not a gene, e.g. [26]. But I find it difficult to take the next step to avoid the use in population genetics of the term 'gene frequency'. Everyone knows what this means. Perhaps 'allele frequency' is a suitable alternative. But then what about 'linked genes'? I really dislike the use of 'linked alleles'. When I learned genetics, 'alleles' were 'alternatives', by implication at a single locus. Wikipedia still seems to support this definition, which makes 'linked alleles' a contradiction in terminology. So what other than 'linked genes' are we to call these linked things, especially in a document that is devoted to them?

2. SOME RECENT PUBLICATIONS

I published nothing on LD for a period of more than 30 years following my little flutter in the 1970s. In the years 2008 - 2013 there has, however, been a flood of publications, four of them, that I'll describe here.

2.1. **LD and the estimation of N_e.**

The equilibrium value of r^2 depends on the quantity $N_e c$ (3). This means that if one knows the recombination frequency c, a method is available for estimating effective population size from measurement of LD at a single point in time [31]. This seems obvious now following the publication of many papers on the topic and publication of the computer program $LDNe$ [44], but I recall that it seemed a eureka moment at the time.

I never followed up on this, since there seemed little chance at the time that data would ever be available to really apply the method. Obviously I failed to foresee microsatellites and Hapmap. I also did not give any thought to the difference between estimates derived from closely linked loci and loosely linked loci, although I must have realised that the former were relevant to times in the past and the latter to recent times. Hayes et al [11] gave a clever argument to show that the critical time is given approximately by $1/2c$.

In fact I really only thought in any detail about very closely linked loci. As detailed below, I actually had two tries at an equilibrium formula for r^2, the first being approximately $(1-c)^2/(1+4Nc)$ and the second just $1/(1+4Nc)$. I thought that there was very little difference between the two, but Bill Hill once pointed out to me that for unlinked loci there is a factor of 4 difference between the two. Furthermore, he and Bruce Weir showed that the correct approximate formula for loosely linked loci is $[(1-c)^2+c^2]/(1+4Nc)$, which for unlinked loci, $c = 1/2$, lies midway between the two. I come back to this briefly in Section 4.4.

For loosely linked loci, up to unlinked loci, an extra factor comes into play, in that the estimate of r^2 contains a factor attributable to sample size, approximately $1/2S$ where S is the sample size (see paragraph below). This is potentially a much higher value than $1/N_e$. Bill Hill [14] first took this sampling factor into account. I was aware of this complication following work on LD with Newton Morton, but hadn't tried to

introduce this. In fact my first impression was that the size of the sampling correction would make it impossible to get a useful N_e estimate from unlinked loci. However following the work of Waples [44] it became clear that if one had enough microsatellite markers with enough heterozygosity then at least a rough estimate could be obtained.

There are, in fact, three reasons why unlinked loci are useful. First, in large data sets, most pairs are of this type. Secondly, one knows the c value for unlinked loci, whereas extensive family data are needed to estimate c for linked loci. Finally they give an estimate for very recent population size. So I realised that my prejudice against these was unwarranted, and started to try to understand the theory behind the correction for sample size. This depends on the work of Hill and Weir e.g. [47], which has several important results that I'll come back to in Section 4.4.. It became clear from the work of Waples e.g. [43] that there are subtle problems in these formulae, or at least in the way they were applied, that led to biases in the N_e estimates.

My efforts at this theory, and those of Emilie Cameron and Stuart Gilchrist's, and the application to their microsatellite data, have now been published [38]. I won't go into any detail here, because it's all published and rather complicated. However I will mention one aspect that remains up in the air. The expected value of r^2, assuming no expected LD, for the case where gametes can be recognised is, from the work of Haldane and others many years ago, $1/(2S-1) = 1/2S.[1 + 1/(2S-1)]$, S being the sample size. In the usual case where gametes can't be recognised, it is necessary to use the 'composite LD coefficient' [45]. Originally I thought the $[1 + 1/(2S-1)]$ bias factor would apply here, but simulation showed that this was not the case. The factor in this case seems to be $[1 - 1/(2S-1)^2]$. However my efforts to prove this depend on the 'ratio of expectations', rather than 'expectation of the ratio'. The closeness of simulations to this value suggest that it might be an exact result, analogous to Haldane's, so if anyone feels like trying this...

2.2. **LD in subdivided populations.**

There have been several papers calculating the expectation for LD within and between populations, the best known being those of Ohta [22]. Before considering the expectation, it is necessary to see what parameter is being estimated, and here I have to admit that I have never been able to understand the rationale behind the range of parameters that Ohta introduces to measure between population LD, D^2_{IT}, D'^2_{IS} and D^2_{IT}. Tachida & Cockerham [41] introduced a different and more

comprehensive set involving genes on the same gamete, genes on different gametes within an individual and genes on different individuals within a deme, and between demes.

It seemed to me that there is a key parameter not considered in either of these sets. The main point of interest when comparing LD values in different populations is how similar they are. An obvious way of measuring this is the covariance of LD values. In cases where populations have been separated for long periods of time, a zero covariance is expected. If there is a lot of migration between populations then similar LD values, or a high covariance, are expected.

The r parameter seems most useful here. The LD measure within populations is the usual r^2. The corresponding parameter to measure LD between populations is $r_i r_j$, for populations i and j. If populations i and j are unconnected, the expectation for $r_i r_j$ is zero. For high rates of migration, the expectation should be close to r^2.

One value of the theory of Linked Identity by Descent (LIBD), elaborated in the following section, is that it can easily be expanded to sub-divided or multiple populations. My paper [33] considers the usual island model, where there are k populations each exchanging migrants at the same overall rate m with other populations. Two parameters are considered, L_W, the LIBD probability for two haplotypes chosen from one population, and L_B, the LIBD probability for two haplotypes chosen from different populations.

Recurrence equations for L_W and L_B can easily be written down. At equilibrium, the expected value for L_W is, approximately,:

$$\hat{L}_W = \frac{1}{1 + 4Nc[1 + (k-1)\rho]}$$

where ρ is a measure of the ratio of recombination to migration:

$$\rho = \frac{m}{m + (k-1)c}$$

With a low ratio of migration compared to recombination, ρ goes to zero, and

$$\hat{L}_W = \frac{1}{1 + 4Nc}$$

With high migration, ρ becomes 1 in the limit, and

$$\hat{L}_W = \frac{1}{1 + 4Nkc}$$

According to the theory of LIBD, the value of r^2 within population equates to L_W. So, as expected, with low migration the value of r^2 is determined by the local population size, while with high migration the entire population determines the value of r^2.

Referring to L_B, the equilibrium value turns out to be

$$\hat{L}_B = \rho \hat{L}_W$$

L_B equates to the parameter for LD between populations, $r_i r_j$. Again, there is a simple equilibrium expectation, with low migration leading to zero expected LD between populations, and high migration giving a value of $r_i r_j$ which in the limit is indistinguishable from r^2.

I did a lot of computer simulation, mostly with just two populations. Agreement with expectation was by no means perfect, but the trends were all in the right direction. It was necessary to do a large number of simulations to achieve the required accuracy. It emphasised just how variable the r values are, and how limited conclusions can be drawn from observation of just a single pair of populations.

2.3. **When did European and African populations split?**

What's that got to do with LD? Maybe the following will explain it.

I haven't written down the recurrence relationships for between-population LD in the previous section. It can be written as:

$$E(r_i r'_j) = (1-c)^2[\beta r_W^2 + (1-\beta) r_i r_j]$$

where $\beta = (2m - m^2)/(k-1)$.

So there is a contribution from the within-population LD, and a contribution from the previous generation between-population LD. Population size doesn't come into it, although it does for the within-population LD relationship. Furthermore for the particular case of zero migration,

$$E(r_i r'_j) = (1-c)^2 r_i r_j$$

So the LD correlation goes down by a fraction $(1-c)^2$ in each generation. Actually this result is fairly obvious for the case of infinite-size populations. And this is the basic idea needed to measure the number of generations that two populations have been separated. It assumes that one can measure the LD correlation, and at least estimate the level of LD when the populations split.

Peter Visscher and I somehow got together to work on applying this theory to the Hapmap data, together with Allan McRae who did a lot of the calculations. It resulted in a paper published in American Journal of Human Genetics [40]. Although this is a high profile journal, the paper has had almost zero citations, partly I suspect because the editors made us publish it as a note rather than as a full paper, making it almost unreadable. It's a pity, because I think that there are a number of aspects to the paper that deserve more attention.

Our overall conclusion was that the split occurred less than 1,000 generations ago, which in itself is quite controversial since mitochondrial data seem to suggest 40-50,000 years. There are assumptions involved in the calculations that I thought might be attacked by others - better to be attacked than ignored. And we found one really weird bias for the most closely linked loci, which led to a negative estimate of split time. This turned out to be due to fixation, which we were able to document. And there were plenty of other potential biases related to estimating recombination frequencies, estimating past r^2 estimates, etc. The introgression of Neanderthal and Denisovan genes into European populations hadn't been published at that time, although my entirely untested assumption is that the low frequencies reported wouldn't be enough to influence overall LD levels.

The calculations all assumed no migration. I was able to put migration into the picture in the manuscript referred to above [33]. The main conclusion from this calculation was that migration between Europe and Africa could easily account for the discrepancies in estimated split times. However to account for the shape of the curve connecting recombination frequencies and estimated split times, it was necessary to assume that this migration was ancient rather than recent.

2.4. **LD between blocks of loci.**

I first looked at the question of batches of loci in my 1968 paper [30]. 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 (this was in my Stanford days) 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 heterozygosites could be written in the form:

$$V_H = V(H_1 + H_2 + ... + H_k)$$

where H_1, H_2 etc are the individual heterozygosities. 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) [10]. The fact that I first suggested this measure has, somewhat to my chagrin, been lost somewhere along the line.

Recently I realised that this method can be extended to looking at the covariance of heterozygosity of different blocks of loci rather than the variance of heterozygosity [37]. Again, the covariance of heterozygosity relates to the sum of D_{ij}^2 terms, this time to all pairs of loci in the different blocks. The statistic has the advantage that it can be applied immediately to diploid data, and is not biased by sample size.

I looked at Hapmap data to see if there were any signs of covariance between blocks on different chromosomes, perhaps as an indicator of the phenomenon of 'affinity' found many years ago in mice. These results were negative. Looking at blocks within chromosomes, it was possible to see LD at distances of up to 10cMs, a surprisingly large distance.

3. SELECTION AND LD

Three papers initially reported on the production of LD through finite size: Hill and Robertson (1968) "Linkage disequilibrium in finite populations" [16], Ohta and Kimura (1969) "Linkage disequilibrium due to random genetic drift" [23], and my paper, Sved (1968) "The stability of linked systems of loci with a small population size" [30]. The title of mine sounds rather different to the other two. 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. Although there were results of interest (see below), I think it was a mistake on my part to concentrate on this one model. In retrospect, other selection models have turned out to be of more interest.

3.1. **Two-locus models.**

Although realistic models of selection need to take into account selection at multiple loci, 2-locus models can give considerable information. This is particularly the case for studying how selection on one locus affects a linked neutral locus.

I haven't seen an attempt previously to classify all possible 2-locus models. The three most commonly discussed single-locus selection models are considered here, positive selection, heterozygote advantage and mutation-selection balance, in addition to neutrality. With two linked loci there are 10 possible combinations, as shown below, where the first column shows the effect of selection on a linked neutral locus. I've attempted to attach names to the different combinations. As written, the positive selection model arbitrarily assumes no dominance and the mutation-selection balance model assumes dominance, although these are not necessary features. I should also mention models of selective interaction, sometimes described as 'equilibrium models', [1] [7] [17], athough these are not considered here.

Lines (3) and (4), the heterozygote advantage and mutation-selection models, basically summarise the 'balanced' and 'classical' world views respectively. Under the balanced view, selection 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

	A_1A_1	A_1A_2	A_2A_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

First 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

linked to each other. These opposing views were highlighted in Lewontin's influential book [19], although I don't see as much current interest in the argument.

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 [5] gives a succinct summary of this work.

3.2. **Associative overdominance.**

My paper [30] was based on the assumption of widespread heterozygote advantage in the genome. Nowadays I would be so enthusiastic about this model. But referring to the 20 chromosome simulation given in Section 1.2, it is clear there are potentially large selective consequences

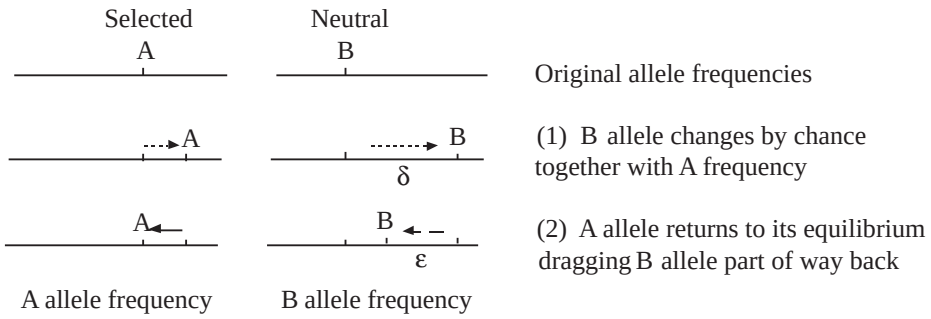


FIGURE 1. Two stages in the interaction of LD and drift

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. This is an extreme situation, but a two-locus model shows some of the properties.

3.3. **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. Figure 1 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 [30] 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.

3.4. **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:

$$S_{Bb} - S_{BB} = \frac{d^2(s+t)}{p_B^2 p_b} \tag{4}$$

$$S_{Bb} - S_{bb} = \frac{d^2(s+t)}{p_B p_b^2} \tag{5}$$

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 [24] independently (but a little later) derived similar results. They termed the phenomenon 'Associative Overdominance', which was actually a term coined previously by Frydenberg (1963) [8]. 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.

3.5. **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 [36]. 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 3.2 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 1 is due to chance in the unselected case and due to positive selection in the current case.

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 [36] for consistency with Figure 1.

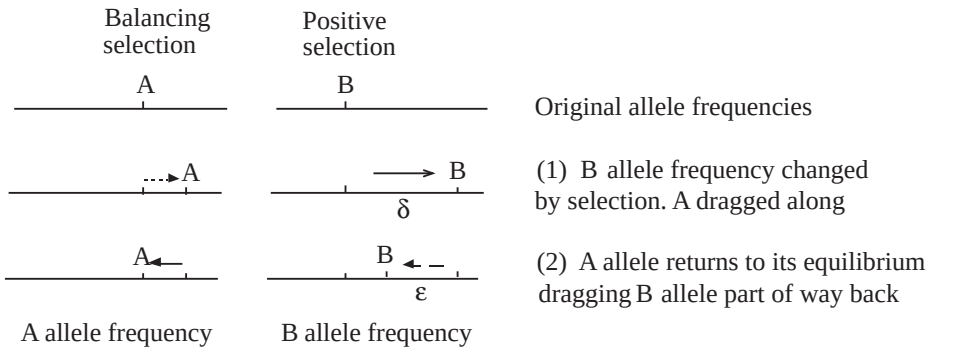


FIGURE 2. Balancing selection opposes positive selection

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

The opposition of selection can be studied by looking at the covariance of selective values. Doing the equivalent calculation to that given in [36], 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 2, 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.

3.5.1. *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).

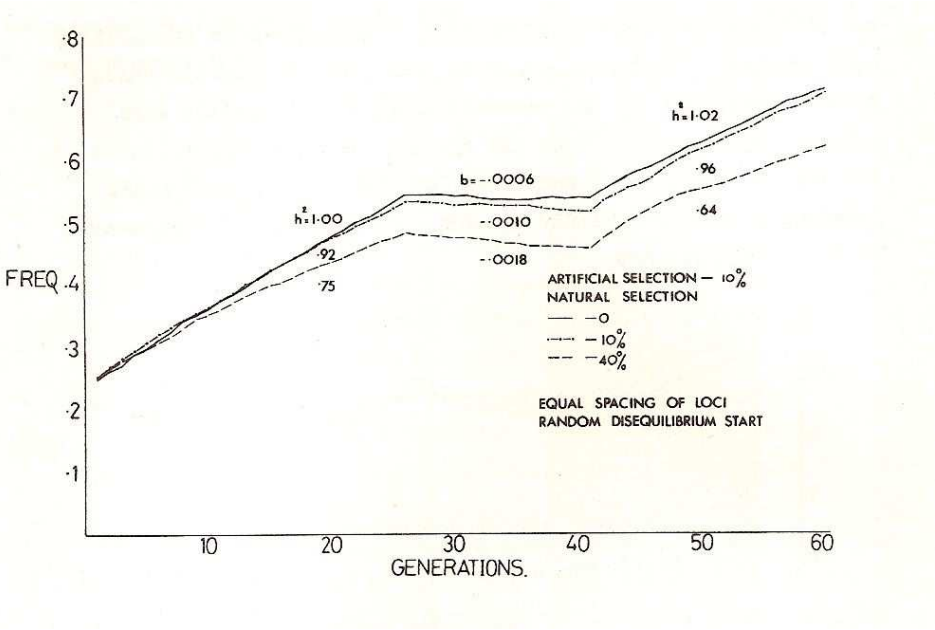


FIGURE 3. Computer simulation of how natural selection might oppose artificial selection

Results in all cases showed low levels of retardation due to 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 3 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 fourfold 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 re-assorts the genes during the period of no selection, a result predicted by Griffing (1960) [9].

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 the computing power available at the time coped with this number, and I have had a bit of success more recently simulating selection in this way thanks to Mr Moore's Law.

3.6. **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. As previously stated, I wish I hadn't been so hung up on the heterozygote advantage model. The current theory is somewhat more sophisticated than the original hitch-hiking theory, depending on multiple locus haplotypes rather than two locus haplotypes [28].

3.7. **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 [5], is sometimes used more broadly to describe all finite-size LD effects. The theory was developed by Hill and Robertson (1966) [15], perhaps before they had clarified their ideas on what effects were due to selection

and what to finite size [16]. 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 compexity of the problem, and of teasing out what aspects are due to drift and what to selection.

3.8. **Background selection.**

This refers to the effect that deleterious genes maintained in the population by mutation have on linked neutral polymorphisms. The theory was put forward by Charlesworth and colleagues [2], again without specific mention of LD. From my point of view, it seemed scarcely surprising that there could not be strict neutrality under these circumstances. What was perhaps surprising to me at the time was the direction of the effect, that polymorphism would actively be reduced compared to the neutral case. In retrospect, it seems clear that the rise in frequency of any new favourable mutation will be opposed, given that it is likely to be in LD with some deleterious gene(s). A similar argument might also be made for the chance rise in frequency of any new mutation.

To see, however, why the direction of response seemed surprising at the time, I need to refer back to a paper of Ohta (1971) [21] and a paper of mine [32]. Ohta's paper showed that, under some circumstances at least, there could be associative overdominance due to linked detrimental genes, i.e. a stabilising effect.

My original paper in 1968 [30] was concerned not with detrimental genes but with heterozygote advantage. However when Ohta's paper came out, it reminded me of an early calculation on deleterious genes I had made. The model is as follows:

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 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.

As mentioned, I found this result quite early during LD-selection calculations but forgot about it in favour of the much more readily interpretable equations (4) and (5). Following Ohta's 1971 paper I somehow felt obliged to come back to this result. It's not something I'm particularly proud of, given that Ohta had already published something similar. 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 [32] was generally ignored. However it and Ohta's paper were picked up by Palsson and Pamilo (1999) [25] 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). More recently Zhao and Charlesworth [49] studied the same problem in more detail, although their conclusions don't seem to contradict those of Palsson and Pamilo. Whether the effect goes one way or the other seems complicated.

4. LINKAGE DISEQUILIBRIUM (LD) AND LINKED IDENTITY BY DESCENT (LIBD)

4.1. **Introduction.**

Why should LD arise in a finite population? The first, conventional, way of looking at this is via frequency arguments, that there will be a correlation of gene frequencies, or LD. But it is also clear, e.g. via the simple simulation in Section 1, that LD is due to the inheritance of multiple copies of particular haplotypes from a common ancestor. So one might also ask - "what is the probability of identity by descent of such linked alleles?" I have come to call this the probability of linked identity-by-descent (LIBD). It refers specifically to identity by descent of two loci via the same pathway from a common haplotype in previous generations, i.e. descent in the absence of crossingover between the two loci on either pathway.

The main purpose of this section is to try to justify the assertion that the LIBD probability directly estimates the LD measure r^2. However I should make a slight diversion here to compare what I mean by the LIBD probability with parameters considered by other authors, specifically Sabeti et al (2002) [27], Hayes et al (2003) [11] and Weir and Cockerham (1974) [46].

Sabeti et al (2002) and Hayes et al (2003) introduced measures labelled respectively extended haplotype homozygosity (EHH) and chromosome segment homozygosity (CSH). From what I understand these are the same thing, directly measuring identity over a chromosome region by observation of identical SNPs. However Hayes et al make allowance for extra chance homozygosity beyond identity-by-descent through the same pathways. CSH and EHH are observable measures, but their probability is essentially the same as the LIBD probability. The measures are intended as direct estimates of LD, without relating specifically to frequency parameters such as r^2. In fact Hayes et al [11] show that CSH is a better measure of LD than r^2 because of its lower variance.

By contrast, Weir and Cockerham (1974), consider all possibilities of IBD at two loci. 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.

Anyway it seemed to me that LIBD and LD provide alternative descriptions of the same phenomenon [31] [34]. 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 still needs justification. Nobody else seems to have taken up the approach, which 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.

4.2. **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'. 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.

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, it would seem desirable to give recurrence relationships for the value in one generation in terms of the value in the previous generation. If, however, there is a certain probability that one of the loci becomes fixed in going from one generation to the next, it's not clear that any exact recurrence relationship can be given.

Much of the calculation on expected r^2 values avoids this problem by calculating not

$$E[\frac{d^2}{p_A(1-p_A)p_B(1-p_B)}]$$

but rather

$$\frac{E[d^2]}{E[p_A(1-p_A)p_B(1-p_B)]}$$

These are, of course, not the same thing. Hill (1977) [13] has shown how to approximate the difference between the two in terms of higher moments. In practice, computer simulation has shown that the two are reasonably similar, and the latter expression has usually been used, thereby avoiding the fixation problem.

4.3. **Argument #1, the basic justification for the LIBD method.**

I was involved in two papers putting forward the LIBD argument, (Sved, 1971 [31] and Sved & Feldman, 1973 [34]). It is convenient to deal with the later paper here, since it is a much simpler argument and, I think, the basic reason why the method works. The first paper has problems which I will go into in detail in section 5.

The argument of [34] depends on the analogy with single locus calculations. The focus in single locus calculations 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. [48]), 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 [3], 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 [3], p66. If two genes are identical by descent, with probability f_A, then their correlation is 1. If they are not identical by descent, then their correlation is 0. Overall, therefore,

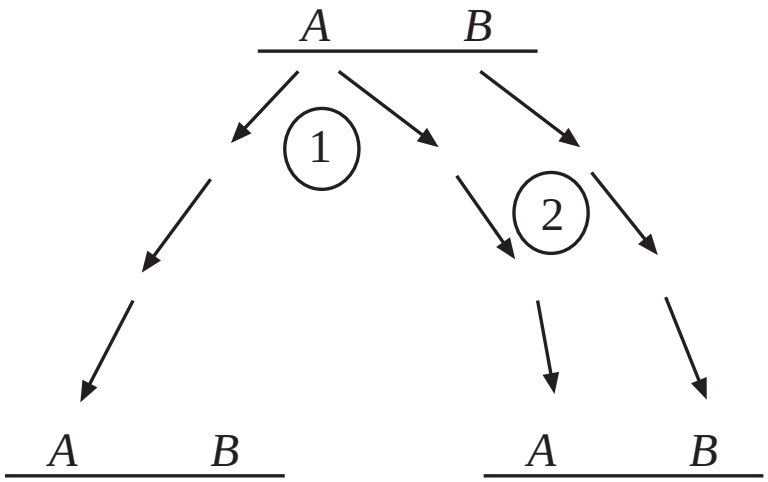
$$r_A = f_A.1 + (1 - f_A).0 = f_A$$

This argument will only work if correlations are additive. The verity of the argument can be checked here by writing out the full set of genotypes cf. Crow and Kimura, 1970, Table 3.2.1. This is given below, where p_A is the frequency of the A allele:

	A	a	
A	$(1-f_A)p_A^2 + f_A p_A$	$(1-f_A)p_A(1-p_A)$	p_A
a	$(1-f_A)p_A(1-p_A)$	$(1-f_A)(1-p_A)^2 + f_A(1-p_A)$	$1-p_A$
	p_A	$1-p_A$	

Then the correlation may be calculated most simply by assigning allele A the value '1' and a the value '0', giving

$$r_A = [(1 - f_A)p_A^2 + f_A p_A - p_A^2]/\sqrt{(p_A - p_A^2)(p_A - p_A^2)} = f_A$$



The argument so far has looked only at genes at a single locus (see (1) in the 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. The correlation between the A and B genes in such a case will thus be 0. The overall correlation is equal to

$$r_{AB} = f_{AB}.1 + (1 - f_{AB}).0 = f_{AB} \tag{6}$$

It has been pointed out to me by people who are much more rigorous in their approach that one can't assume that correlations are additive, as I have blithely assumed here. But I'm fairly sure that the argument works OK here, provided that the variances, the denominator in the correlation calculation, are the same for each of the A and B loci, and are also unaffected by crossingover. These variances must be determined just by population structure, which affects both loci in the same way. I've done some simulating just to confirm that assigning a random value in the range(0, 1), the same value for A and B with probability f, and different random values with probability $1 - f$, does give a correlation coefficient equal to f.

Unfortunately I can't see a direct comparison with the single locus table above. For a single locus, one can easily write down the frequency of AA genotypes as $f_A.p_A + (1 - f_A).p_A^2$. It is not obvious, to me at least, how one writes the frequency of AB gametes in terms of f_{AB} and the haplotype frequencies.

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

$$E[r_{AB}^2] = f_{AB}^2 = L. \tag{7}$$

4.3.1. *Calculating the LIBD probability L.* This probability requires a recurrence relationship between generations, specifically between a parent generation and an offspring generation. Assuming the simplest Wright-Fisher haploid model, offspring are produced by choosing from an infinite pool of gametes produced by the parent generation, equivalent to choosing two gametes with replacement from the parent gametes. LIBD in the offspring generation requires that there be no recombination in either gamete coming from the parent generation, since any recombination event will randomise the connection between the two loci. So the LIBD probability in the offspring generation is obtained by choosing two parent generation haplotypes, multiplied by the probability of no recombination in either, $(1 - c)^2$. In a parent population of $2N$ haplotypes, the chance that the same haplotype is chosen twice is $1/2N$. The chance that two different haplotypes are chosen is $1 - 1/2N$, in which case the probability that two such gametes are identical is, by definition, L, the LIBD probability of the parent generation. Overall, therefore,

$$L' = \frac{(1-c)^2}{2N} + (1 - \frac{1}{2N})(1-c)^2 L \tag{8}$$

L, the probability of LIBD, refers to gametes, or haplotypes, chosen from a particular population. This formula was given in [31]. Unfortunately when Marc Feldman and I [34] made this argument, we introduced a novelty that seemed appropriate at the time, and insisted that this process needed to be sampling with replacement not from the parent generation but from the offspring generation.

The reason why we (actually it was my fault rather than Marc's) did this, relates to the attempt to equate the LIBD parameter to the quantity r^2 which is calculated from quantities such as d^2 by multiplying

frequencies as if the sample size was infinite. While it may seem artificial to sample the same gamete twice and refer to this as LIBD, such sampling seemed necessary to equate probabilities with statistics calculated from gene frequencies. It is sometimes possible to calculate statistics that do not make this assumption; for example the true expected frequency of homozygosity for an allele having n copies in a population of $2N$ alleles would be $n/2N.(n-1)/(2N-1)$, rather than $(n/2N)^2$. Sampling without replacement would be the valid procedure if homozygosity was calculated in this way. However this is not the way that statistics such as r^2 are calculated. Single locus calculation with Barrie Latter [35] emphasised how sampling with replacement from the population is necessary to equate probability and frequency statistics.

What we [34] failed to take into account was clarified by Weir and Hill (1980) [47], although perhaps there are earlier similar arguments. They pointed out that there are two distinct processes - (1) producing an offspring population from the parent population, usually according to the Wright Fisher model, and then, if necessary, (2) taking a sample from the offspring. In constructing a between-generation recurrence relationship it does not make sense to take the second sampling process into account. The recurrence relationship derived in [34]

$$L' = \frac{1}{2N} + (1 - \frac{1}{2N})(1-c)^2 L \tag{9}$$

seems to give the correct answer if the whole population is sampled. However it is simpler, as well as being much easier to justify, if one ignores the second sampling process and concentrates just on the recurrence relationship between the parent and offspring populations.

Equation (8) easily generalises to any number of generations. It gives an equilibrium value for L of

$$\frac{(1-c)^2}{1 + (2N-1)(2c-c^2)} \tag{10}$$

so that for small values of c we have

$$E[\hat{L}] \approx \frac{1}{1+4Nc}$$

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, using $E[r^2] = L$, the expected value of r^2 in the offspring generation in terms of the parent generation is:

$$E[r^{2\prime}] = \frac{(1-c)^2}{2N} + (1 - \frac{1}{2N})(1-c)^2\, r^2 \tag{11}$$

and

$$E[\widehat{r^2}] \approx \frac{1}{1 + 4Nc} \tag{12}$$

The expected value of r^2 from sampling can then be calculated approximately from $\rho^2 + (1 - \rho^2)/(S - 1)$, where ρ is the correlation in the parent population and S is the number of gametes sampled [42].

4.4. **LIBD with loose linkage.**

One aspect of the LIBD argument has worried me for a long time; it gives the wrong answer for loose linkage. This has been most obvious in the case of unlinked genes, which has been discussed in connection with estimating population size (Section 2.1). As shown above, equation (10), the equilibrium value of the LIBD probability L contains the term $(1 - c)^2$ in the numerator. Weir and Hill (1980) [47] give an equilibrium expectation for r^2, with $(1 - c)^2 + c^2$ in the numerator. Actually the expectation is for the ratio of expectations rather than the expected value of the ratio (Section 4.2), but computing shows that the formula works well for most values of r^2. In particular, it gives a value for unlinked loci that is double the value given by (10). In our paper analysing microsatellites [38], we just accepted the Weir and Hill expectation, leaving unanswered the question of why the LIBD calculation appears to fail.

In early 2016 I received a letter from Igor Chybicki from Kazimierz Wielki University in Poland, suggesting a possible solution to this dilemma. He pointed out that my derivation of equation (8) misses an important possibility, chiefly because the derivation assumed a haploid rather than a diploid model. I had not appreciated this deficiency of the haploid model.

The extra term contributed by the diploid model comes from the fact that if two gametes are produced, each with crossingover, identity by

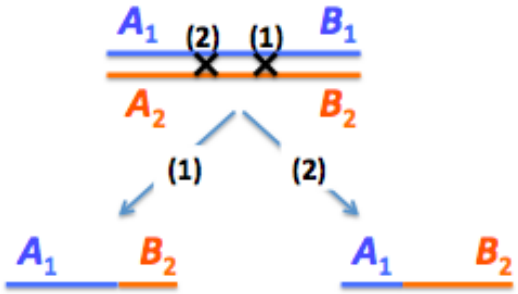


FIGURE 4. The effect of compensating crossovers

descent of the two gametes is assured at both loci. This event is described as 'compensating crossovers' in Figure 4 (to be distinguished from double crossingover). The probability of this event is c^2. The model differs from the haploid model where two gametes produced by recombination and IBD at the A locus would not be IBD at the B locus. The event with compensating crossovers cannot strictly be described as LIBD. However its consequences are the same, an increased correlation between alleles at the two loci equivalent to what would be expected with no crossingover.

Calculation of the relationship between generations, Figure 5, is a little more complex than for the haploid model. It focuses on individuals rather than haplotypes, with three possibilities rather than two:

[1] Two different offspring haplotypes come from the same parent individual. Two cases need to be considered here:

[1a] The same gene is selected at the first of the two loci. This is the situation described above, where compensating crossovers need to be taken into account as well as no crossovers, as shown in the first box of Figure 5.

[1b] If different genes are selected at at the first of the two loci, LIBD in the offspring is only possible in the case of LIBD in the parent. With LIBD of the two gametes in the parent, crossingover will have no effect on LIBD in the offspring.

[2] If the offspring haplotypes come from two different parents, with probability $1 - 1/N$, LIBD in the offspring is only possible if there is LIBD of the two chosen haplotypes, multiplied by the probability of no crossingover. Note that 'compensating crossovers' do not lead to LIBD where different parents are involved, except in the random, and

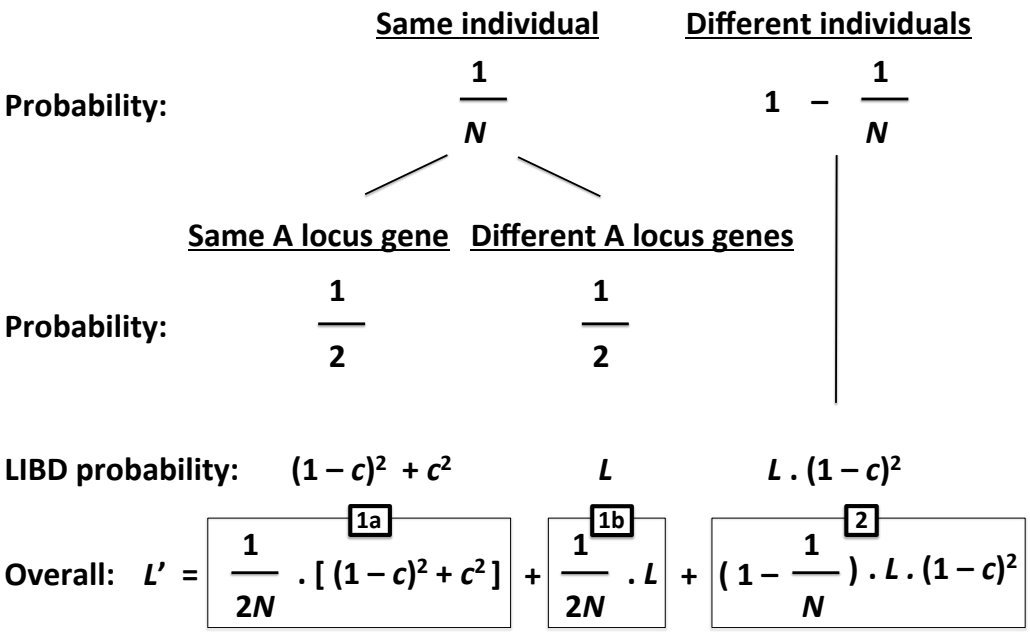


FIGURE 5. Calculation of the LIBD recurrence relationship for a diploid model

unlikely, event of LIBD of the second haplotypes possessed by the two parents.

The overall offspring LIBD probability is the sum of the three probability boxes of Figure 5. The steady state solution can again be found by putting $L' = L$, giving, in essential agreement with [47]

$$\hat{L} = \frac{(1-c)^2 + c^2}{1 + (2N-2)(2c-c^2)} \tag{13}$$

The calculation of Figure 5 assumes a diploid model with no sexes. Separate sexes can also be taken into account, although the calculation is more complex. The result in this case is the same as with no sexes, except that N in equation (13) is replaced by the effective population size N_e, defined by

$$\frac{4}{N_e} = \frac{1}{N_f} + \frac{1}{N_m}$$

Weir and Hill [47] also reported the surprising finding that monogamy, where parents mate for life rather than random choice of partners for each offspring as assumed in the calculation above, has an effect on the expected r^2 value. Igor and I were able to take this case into account, where an extra factor of compensating crossovers in sibling parents increases the LIBD probability. The numerator of equation (13) in this case becomes $(1-c)^2 + \frac{3}{2}r^2$, as found in [47].

Weir and Hill [47] further calculated the expected 'composite' r^2, previously mentioned in Section 2.1. In this case, monogamy has a larger effect on equilibrium r^2. The numerator of the equilibrium r^2 becomes $(1-c)^2+c+c^2$, which is double the non-monogamous value for unlinked loci (c = 0.5). We were again able to derive this result for the LIBD probability.

We submitted our calculations for publication, but were unable to get past the reviewers, for reasons that I found difficult to understand (naturally). Anyway it is available HERE in case you are interested. Perhaps it is not surprising that a paper that mostly just re-derives forty year old results is only of interest to somebody (me) trying to vindicate the probability method. But should there still be any interest in calculating LD or its expectation, the LIBD method still seems of value. Largely this has already been demonstrated by the calculations of LD within and between populations (Section 2.2), where pages of algebra can be replaced by a formulation that identifies and calculates very simple LD measures.

5. MORE ON LIBD

The first part of this section is an attempt to explain the thinking behind my paper [31] that introduced a predecessor to the LIBD argument given in the previous section. I'm embarrassed about the derivation in this paper, which I now see has major errors. Then follows a third attempt at justifying the LIBD argument in terms of 'LIBD classes'. This section could have been omitted, and can safely be skipped. It is included for two reasons. First, [31] is the only LD paper I have written that is cited nowadays, presumably because it has got into the literature as being the first mention of the equation $E[r^2] = 1/(1+4Nc)$, and nobody actually reads it. Nevertheless I feel some obligation to try to explain it. Secondly, although many of the arguments are highly circuitous, I do feel that they raise some points of interest.

5.1. **Argument #2 - LIBD and homozygosity.**
Argument #1 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 [31] 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 (7).

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 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 [31] was the following. Suppose that one chooses one haplotype, and then chooses another haplotype containing the same allele at the A locus. How does this affect homozygosity at the B locus?

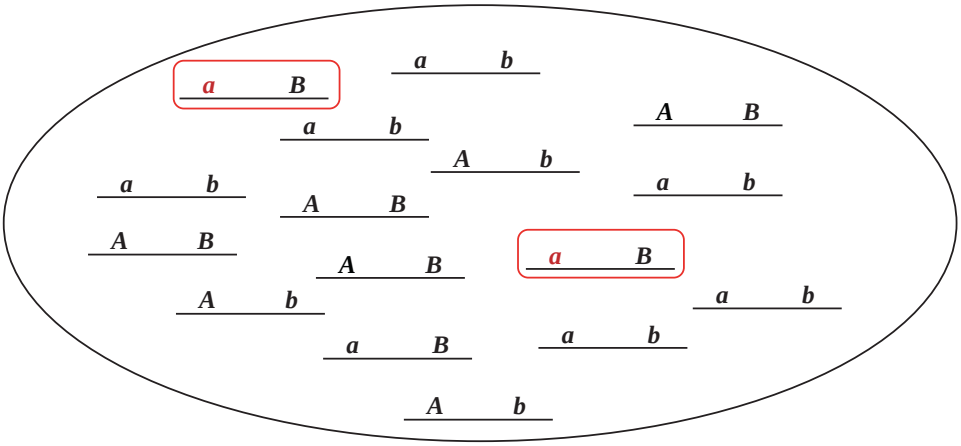


FIGURE 6. Calculation of the LIBD recurrence relationship for a diploid model

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 diagram here 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 now 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

$$h_c = p_B^2 + p_b^2 + \frac{2d^2}{p_A p_a} = h + \frac{2d^2}{p_A p_a} \tag{14}$$

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

$$h_c = L.1 + (1 - L).h$$

which simplifies to

$$h_c = h + 2Lp_Bp_b \tag{15}$$

Comparing the two approaches for predicting h_c, ie comparing (14) and (15):

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

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

$$E[r^2] = L. \tag{16}$$

5.1.1. *Where the argument goes wrong.*

Equation (9) derived in section 4.3.1:

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

purports to show the relationship between generations for two haplotypes selected, with replacement, from the population. Here L' is the LIBD probability from the offspring generation and L the probability

for the parent generation. The argument in [31] assumes that this relationship will work for the particular sampling in which two haplotypes are selected with the same allele at the A locus.

In hindsight it is clear that one can't write the probabilities 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.

Figure 7 shows haplotype numbers in two generations. In the Offspring generation, marked with the rectangle, 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 offspring generation are LIBD?″. I'll call this probability L'_c, the c subscript indicating that this is a conditional LIBD parameter.

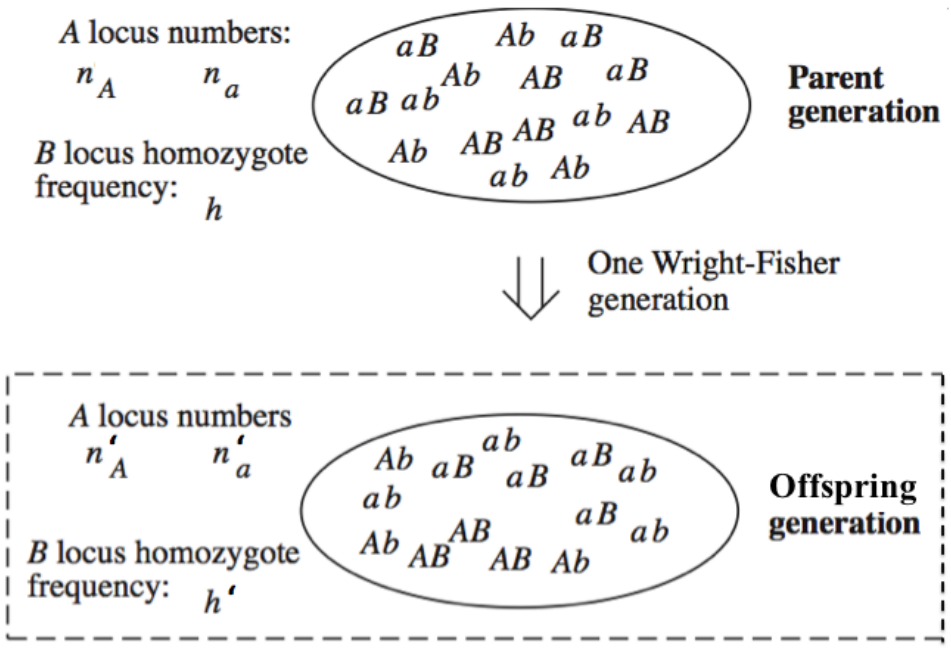


FIGURE 7. Parent and offspring generations

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 $(1-c)^2 L_c$, where L_c is the equivalent conditional probability in the parent generation.

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

$$\frac{n'_A}{2N}[\frac{1}{n'_A} + (1 - \frac{1}{n'_A})(1-c)^2 L_c]$$

which is equal to

$$\frac{1}{2N} + (\frac{n'_A}{2N} - \frac{1}{2N})(1-c)^2 L_c$$

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

$$\frac{1}{2N} + (\frac{n'_a}{2N} - \frac{1}{2N})(1-c)^2 L_c$$

The sum of these two terms is the conditional probability of LIBD in the offspring generation L'_c. This simplifies to

$$L'_c = \frac{1}{N} + (1 - \frac{1}{N})(1-c)^2 L_c \tag{17}$$

Comparing equations (17) and (8) shows that the probability from new LIBD, $1/N$, is twice the regular IBD probability. In other works, 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. By the same argument, for three alleles the distance would be one third of the regular value.

5.1.2. *Can the argument be resurrected?*

I've been through some tortuous calculations that I won't include to show that, maybe, it can. But I doubt there is much point in trying to resurrect this particular argument. As stated previously, the attempt to coerce a two allele model into a coalescence framework is a frustrating one.

I still believe, however, that homozygosity of linked genes studied using LD parameters (r) ought to be equivalent to homozygosity using probability (LIBD) parameters. But it will need somebody else to show this.

5.2. **The length of identical 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 determines haplotypes in a population is the position of crossovers along a chromosome.

I calculated a simple statistic along these lines in [31]. 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 turns out to be not very dependent on chromosome length, and is approximately

$$\frac{1}{2N}(logN - 1) \tag{18}$$

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

Population size	100	1,000	10,000
Mean (cMs)	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 the occasional long homozygous segment that contributes the most to the mean.

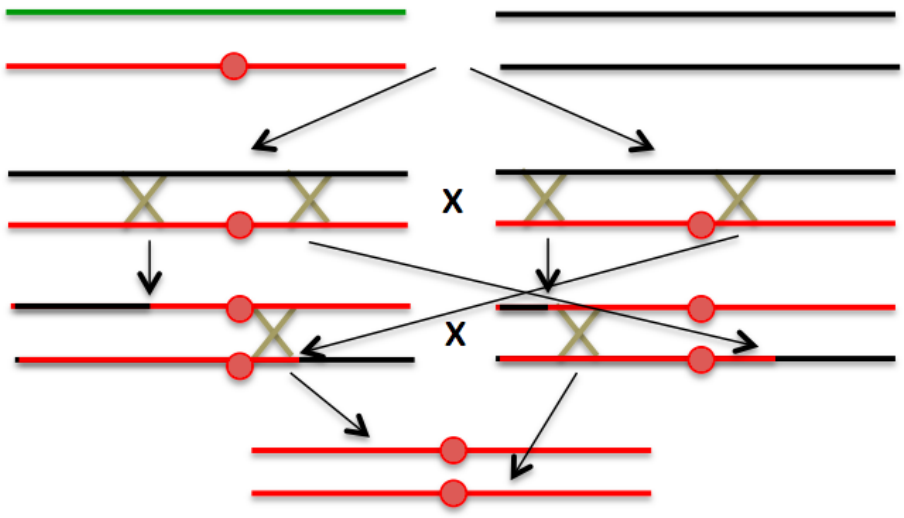


FIGURE 8. Total IBD with crossovers

Many authors have written on similar topics (see the references to EHH and CSH in Section 4.1). An important contribution came from Stam (1980) [29] who raised objections to (18), because my formulation ignores cases of IBD of a segment in which there has been some crossingover in past generations. Figure 8 shows an extreme example of this kind, where the entire chromosome surrounding an IBD locus is IBD despite multiple crossovers. I'm not sure whether such 'secondary' events with crossingover should be included in such a statistic or not. I was thinking of this statistic as giving a value for the likely 'apparent' selective value at a locus, in which case secondary events should probably be ignored. But it is true that when chromosome segments are studied directly, which is now possible due to cheap sequencing and SNP calling, all such events should be included. Stam also points out that my formula used the equilibrium LD value. I'm not sure that this assumption makes much difference, owing to the likely large recombination distances involved.

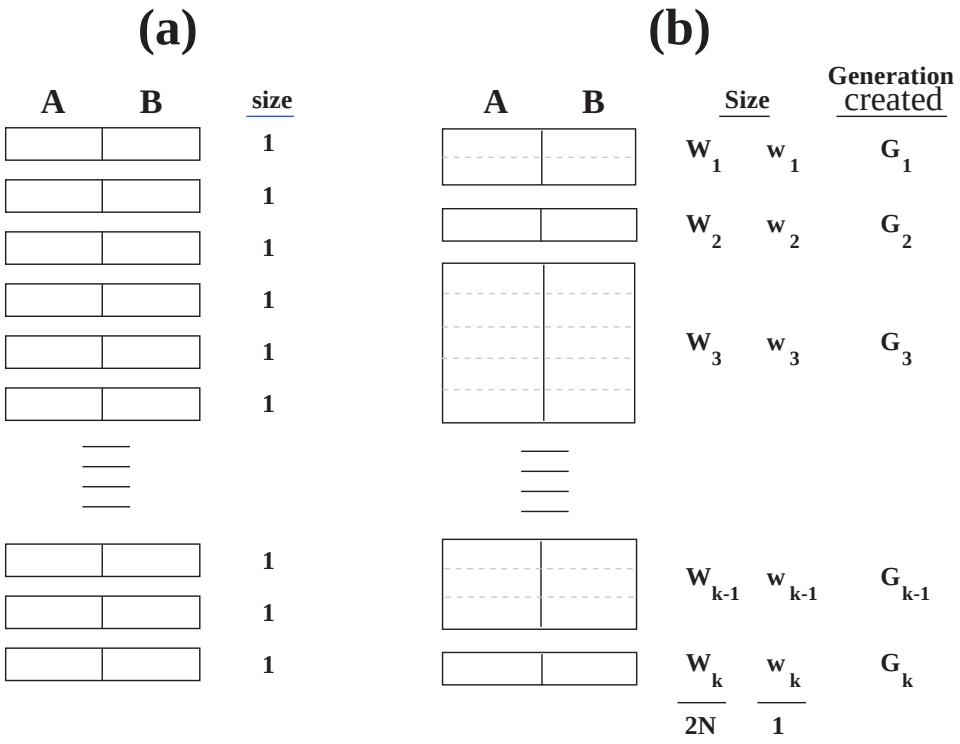
While on the topic of lengths of identical segments, I'd like to comment on one aspect that I'm not sure is generally appreciated. There is a substantial literature on how to infer haplotypes from diploid data (e.g. Excoffier and Slatkin, 1995 [4]). When sending off one's own DNA for sequencing to one of the commercial companies offering this service, the analysis cannot infer haplotypes, i.e. cannot distinguish which SNPs are in 'coupling' or 'repulsion'. Nevertheless when genomes of related individuals, e.g. cousins, are compared, these companies can infer which haplotype regions are identical in the two genomes.

I haven't seen how this analysis is done, but I assume that it must be at the diploid level. For example, if one individual at a particular site has SNPs (1, 1) and the second individual has (1, 2), then there is the potential for identity. On the other hand, if the SNPs are (1, 1) and (2, 2), then there cannot be identity. If the SNP density is sufficiently high, it seems that a fairly accurate picture emerges of the length of shared haplotypes, with only a small margin of error at the ends. The complexity of analysis of programs such as in [4] seem unnecessary if all that is required is a pattern of shared haplotypes.

5.3. **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 the 'LIBD classes' in a population. The population in (b) consists of $2N$ gametes. In each of the k classes, the gametes are identical copies of an ancestral gamete. New classes are created by recombination - each

recombination event starts a new class. For comparison, the population in (a) is depicted as being newly set up, with no LIBD between gametes. After some period of time, the population is as specified in (b). The size of class i can either be specified as a number, W_i, or as a frequency w_i. Each class is created by a recombination event at some generation labelled G_i in the figure below.



We now introduce alleles and allele frequencies into the model, assuming that there is segregation at the A and B loci in population (b). We call the frequencies of alleles A at the A locus and B at the B locus p_A and p_B respectively. The main point of the LIBD argument is to see to what extent sampling into the LIBD boxes using these current frequencies can account for the current value of r^2.

The initial assumption is for independence of A and B alleles in the LIBD classes. This is because each class is initiated by a recombination event that randomises the connection between A and B alleles. So we ask initially, what is the expected value of r^2 given by independent sampling of A and B alleles into the unequal-sized LIBD boxes as shown in the figure.

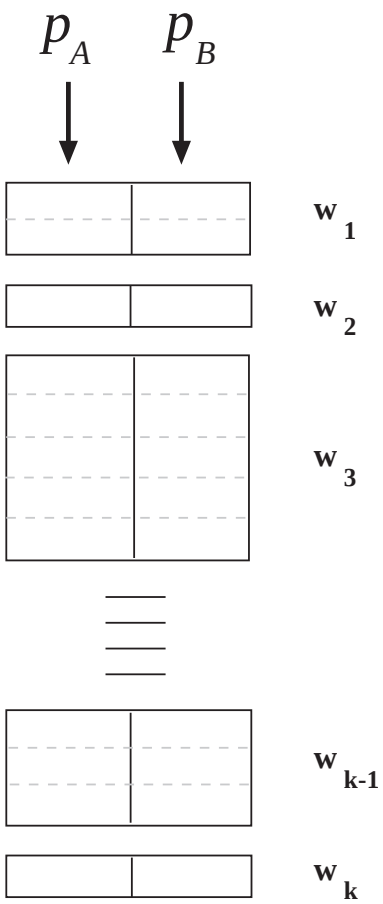


FIGURE 9. Sampling into LIBD boxes using the current allele frequencies

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 (10), and analogous to the argument of Kimura and Crow [18], the expected equilibrium value of the sum of squares of the class frequencies, w_i, is

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

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.

5.3.1. *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 $c_1p_1 + c_2p_2$.

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

$$E[r^2] = \sum w^2 \tag{19}$$

5.3.2. *Why r is close to a linear combination of frequencies.*
I follow the argumenbts of Fisher (1922) [6]. 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

$$\chi_c^2 \quad = \quad \frac{y^2}{V} \tag{20}$$

where

$$y = \frac{p_{AB}}{p_A} - \frac{p_{aB}}{p_a} = \frac{n_{AB}}{n_A} - \frac{n_{aB}}{n_a} \quad (21)$$

and

$$V = \frac{n_B}{2N}.\frac{n_b}{2N}.(\frac{1}{n_A} + \frac{1}{n_a}) \quad (22)$$

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 (20) 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.

5.3.3. *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 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.

5.3.4. *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 (19).

5.4. **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

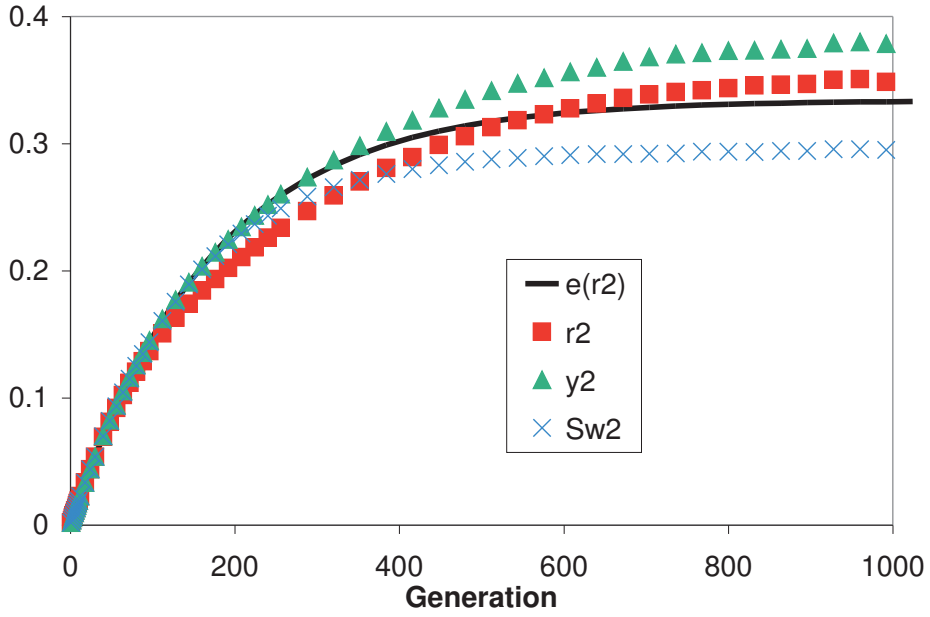


FIGURE 10. r^2 vs. expectation for build-up of LD

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 (Figure 10) 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 (11) (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 10 is the statistic y^2, the numerator of r^2 as defined in (21). This closely tracks the value of r^2. The curves given in Figure 10 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 11 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 (22) plays a reasonably small role in the calculations.

5.4.1. *Aside on values of* $2NV$.
It may seem strange that the value $2NV = \frac{n_B n_b}{2N}.(\frac{1}{n_A} + \frac{1}{n_a})$ 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 12 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

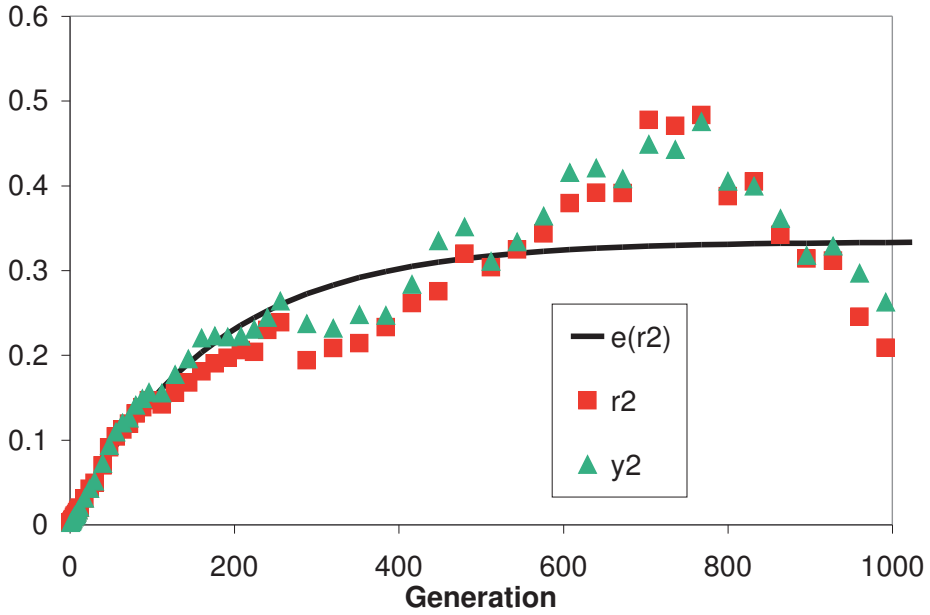


FIGURE 11. r^2 vs. expectation for build-up of LD with small number of replicates

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.

5.4.2. *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 13 . 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

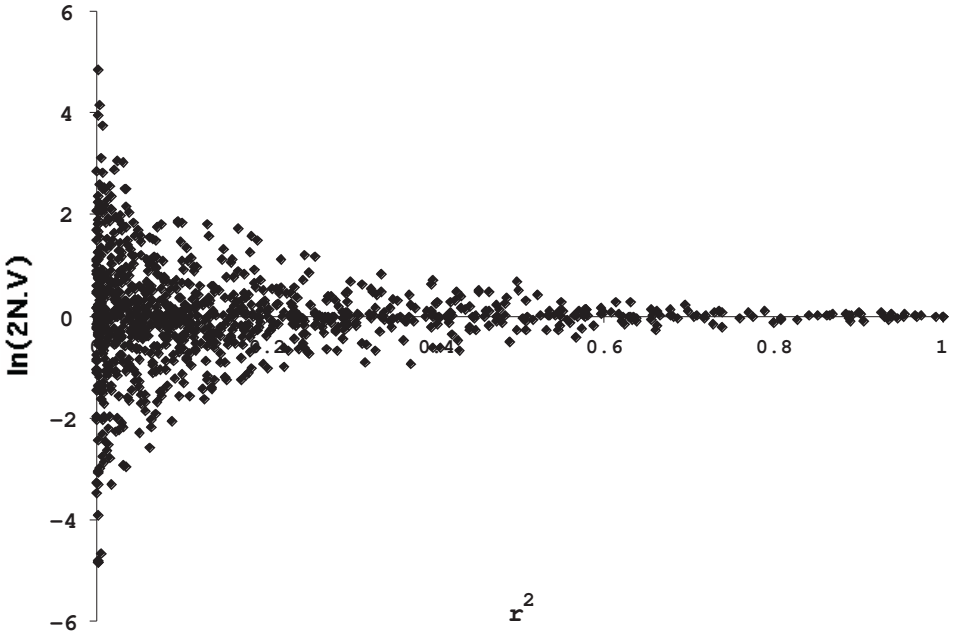


FIGURE 12. Plot of r^2 against $2NV$

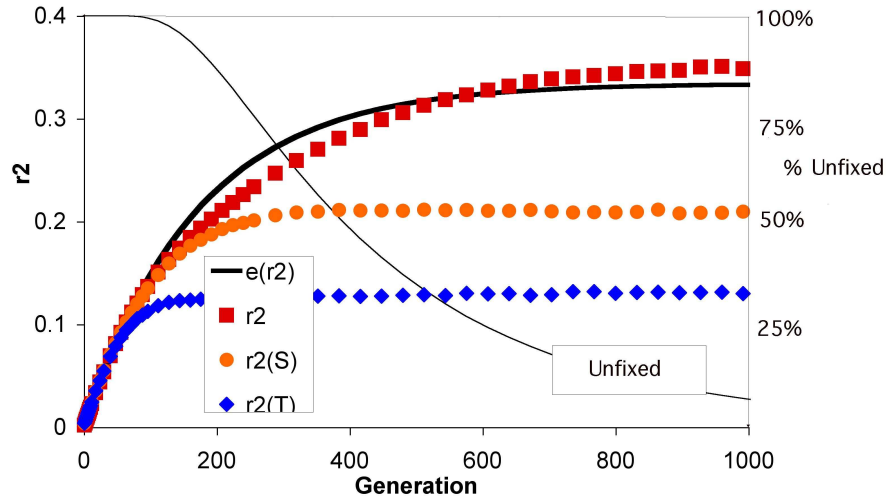


FIGURE 13. Comparison of sampling methods against r^2

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 13) expands the early generations, showing the divergence coinciding with the start of fixation.

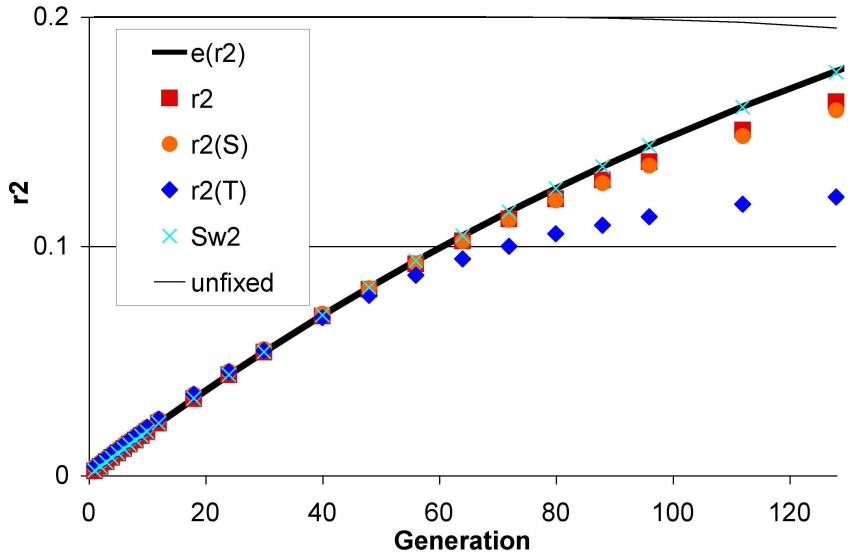


FIGURE 14. Early generations of Figure 13

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 15 shows the results from one such simulation, having the same starting parameters as Figure 13. The discrepancy between $r^2(S)$ and r^2 is reversed in this case. Extreme $r2(S)$ values are found in later generations, where few LIBD classes contribute segregating alleles.

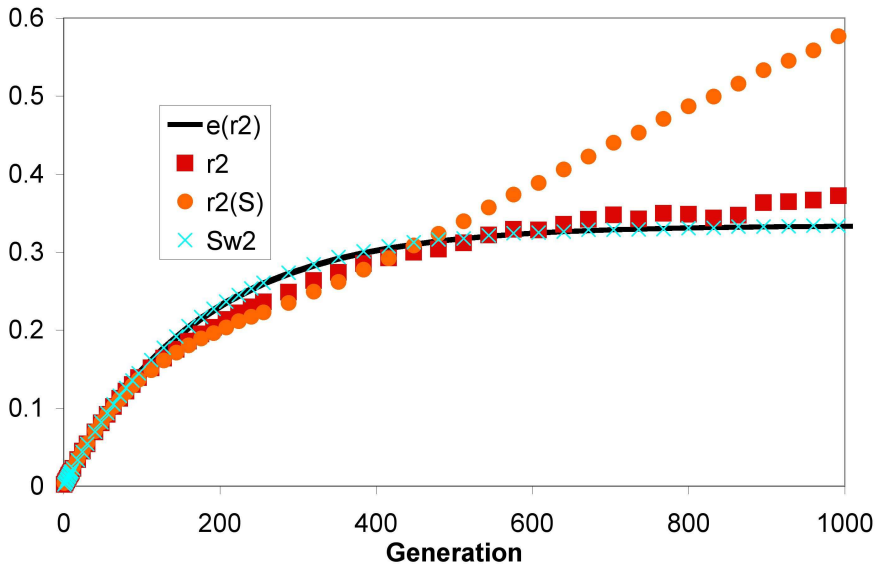


FIGURE 15. $r^2(S)$ plot including cases of fixation

The important comparison for present purposes is between $r^2(S)$ and $r^2(T)$ in Figure 13. This is repeated in Figure 16. 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 (19), introduces a substantial error.

Figure 16 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 (19). 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

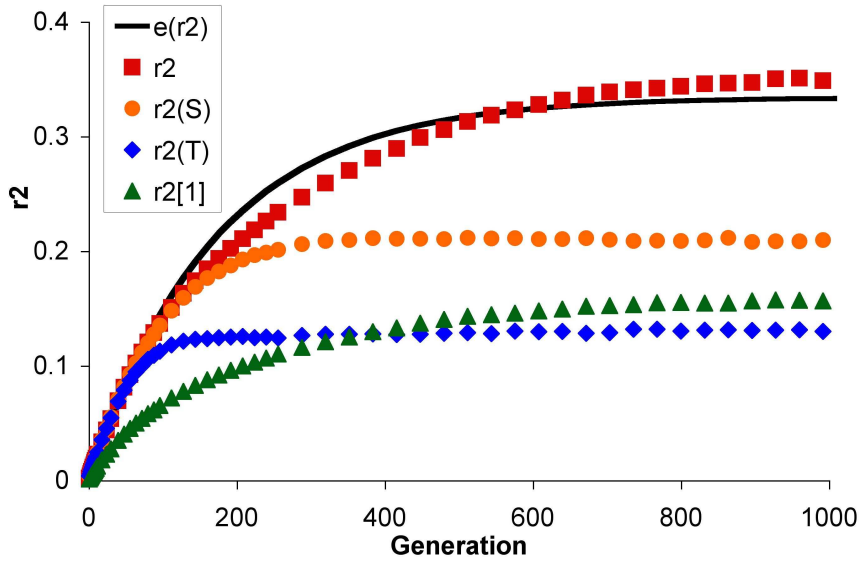


FIGURE 16. $r^2(S)$ versus $r^2(T)$ and $r^2[1]$

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 [20] 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.

5.4.3. *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 13). I'd be grateful for suggestions on these points.

6. LD UNDER A MUTATION MODEL

The simulations reported in Figure-10 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.

6.1. **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

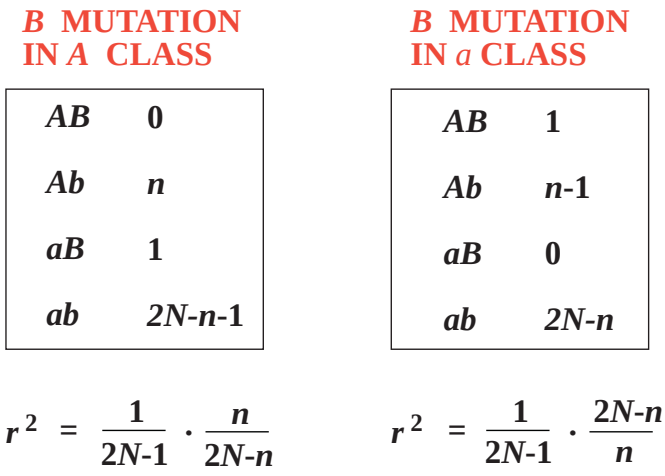


FIGURE 17. Population configuration after a single B mutation

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

which is approximately

$$\frac{ln(N)}{ln(N) + ln(2)} \tag{23}$$

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 17. 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 17 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.

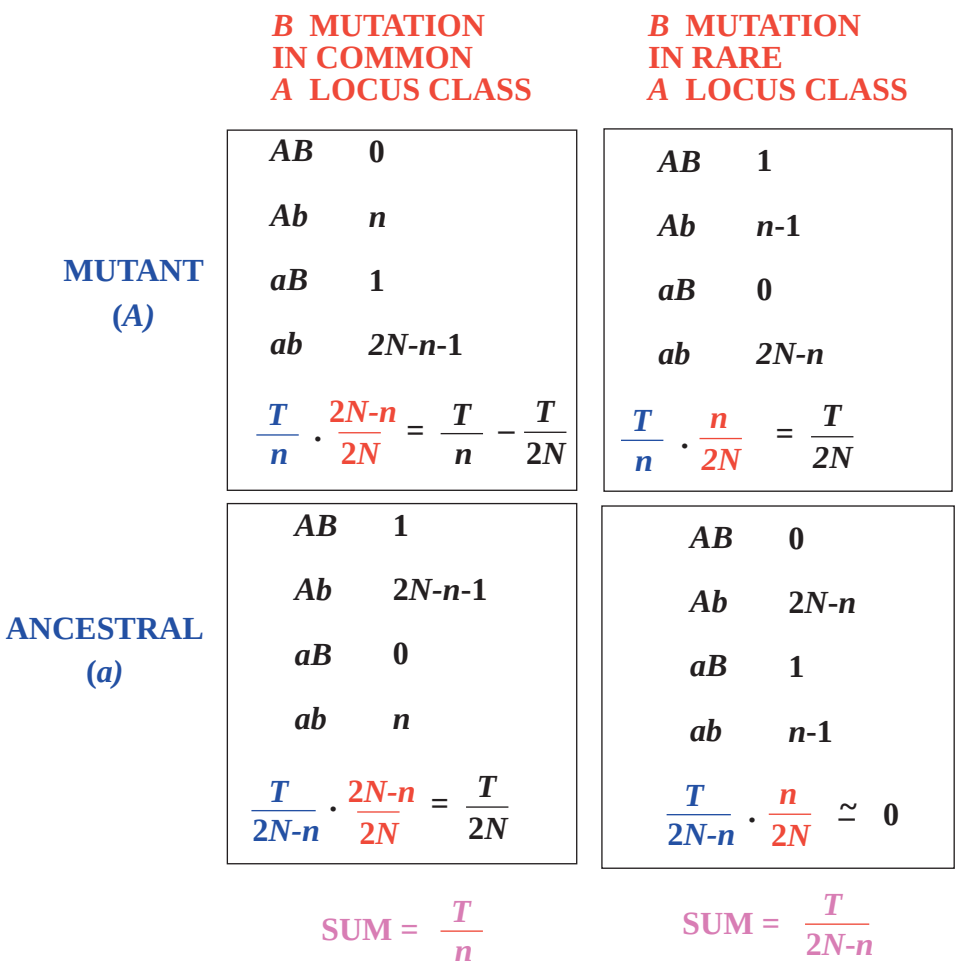


FIGURE 18. Probabilities of the various population configurations

Figure 18 extends Figure 17 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 18. 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 (23):

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

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 17, the mean value for common mutations is equal to

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

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

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

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) [23].

6.2. **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

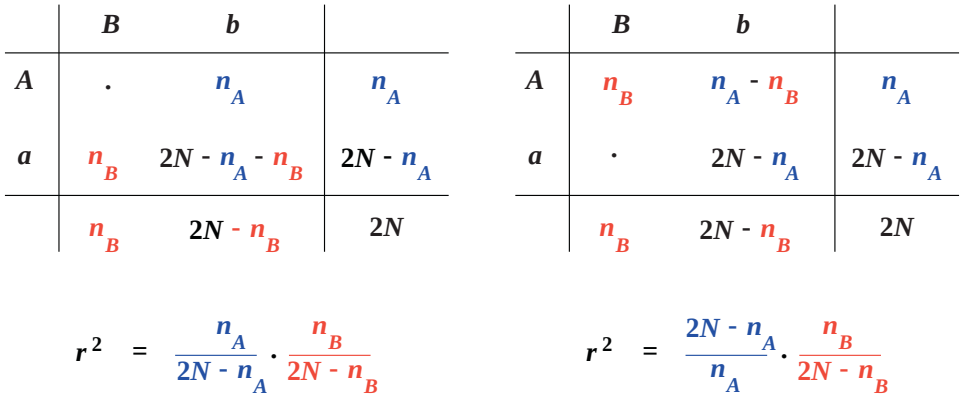


FIGURE 19. Genotypes following one Wright-Fisher generation

side of Figure 19 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

$$\frac{2N!}{n_A!n_B!(2N-n_A-n_B)!} \quad (\frac{n}{2N})^{n_A}.(\frac{1}{2N})^{n_B}.(1-\frac{n}{2N}-\frac{1}{2N})^{2N-n_A-n_B}$$

This expression needs to be multiplied by the associated value of r^2, $\frac{n_A}{2N-n_A}.\frac{n_B}{2N-n_B}$ (Figure 19). 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 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 17.

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 19 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

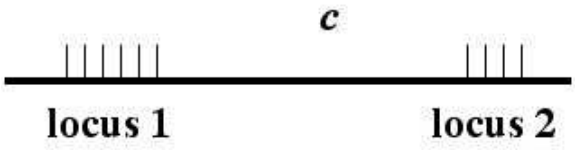
$$\frac{2N-n_A}{n_A}.\frac{n_B}{2N-n_B}$$

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 17. As previously, the sum needs to be corrected for unfixed populations by dividing by the 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.

6.3. **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,

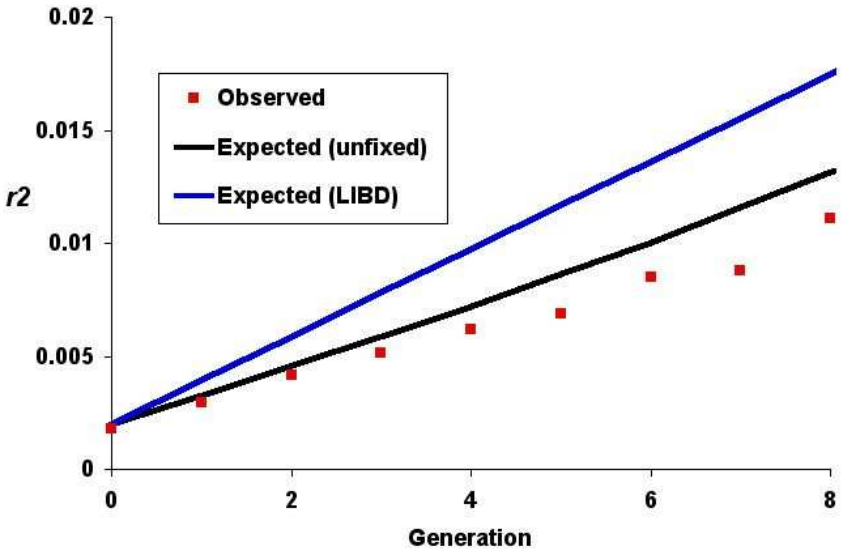


FIGURE 20. Observed and expected values of r^2 over 8 generations following production of a new mutant. Population size (N) is 256.

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 20 and Figure 21. 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 (11).

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 22 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

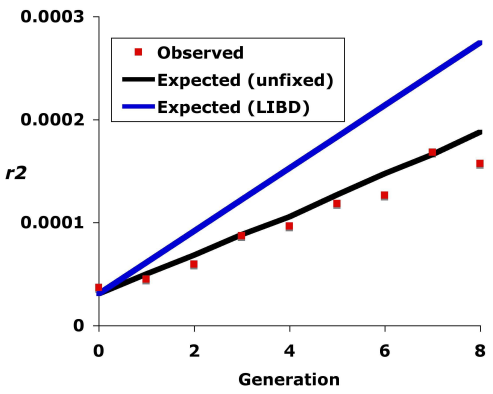


FIGURE 21. Observed and expected values of r^2 over 8 generations following production of a new mutant. Population size (N) is 16,384.

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 20 and 21, 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.

REFERENCES

[1] W F Bodmer and J Felsenstein. Linkage and selection: theoretical analysis of the deterministic two locus random mating model. *Genetics*, 57(2):237–65, Oct 1967.

[2] B Charlesworth, M T Morgan, and D Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303, 1993.

[3] J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Harper & Row, New York, 1970.

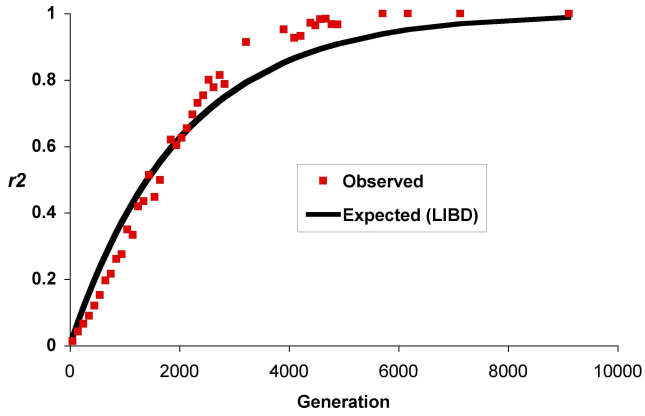


FIGURE 22. Observed and expected values of r^2 over 10,000 generations following production of a new mutant. Population size (N) is 1024. Observed values averaged over 100 data points to reduce chance fluctuations.

[4] L Excoffier and M Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, Sep 1995.

[5] J Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974 Oct.

[6] RA Fisher. On the interpretation of ï2 from contingency tables, and the calculation of p. *J. Roy. Statist. Soc.*, 85, 1922.

[7] I Franklin and R C Lewontin. Is the gene the unit of selection? *Genetics*, 65:707–734, 1970 Aug.

[8] O Frydenberg. Population studies of a lethal mutant in drosophila melanogaster. i. behaviour in populations wiuth discrete generations. *Hereditas*, 48, 1963.

[9] B Griffing. Theoretical consequences of truncation selection based on the individual phenotype. *Aust J Biol Sci*, pages 307–343, 1960.

[10] B Haubold and RR Hudson. Lian 3.0: detecting linkage disequilibrium in multilocus data. *Bioinfomatics*, 16:847–849, 2000.

[11] BJ Hayes, PM Visscher, HC McPartlan, and ME Goddard. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res*, 13:635–643, 2003.

[12] P. W. Hedrick. Gametic disequilibrium measures: proceed with caution. *Genetics*, 117:331–341, 1987.

[13] W G Hill. Correlation of gene frequencies between neutral linked genes in finite populations. *Theor Popul Biol*, 11:239–248, 1977 Apr.

[14] W G Hill. Estimation of effective population size from data on linkage disequilibrium. *Genet Res*, 38:209–216, 1981.

[15] W G Hill and A Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8:269–294, 1966.

[16] W G Hill and A Robertson. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, 38:226–231, 1968.

[17] S Karlin and M W Feldman. Linkage and selection: two locus symmetric viability model. *Theor Popul Biol*, 1(1):39–71, May 1970.

[18] M. Kimura and J. F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49, 1964.

[19] R. C. Lewontin. *The genetic basis of evolutionary change*. Columbia U.P., N.Y., 1974.

[20] GAT McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162:987–991, 2002.

[21] T Ohta. Associative overdominance caused by linked detrimental mutations. *Genetical Research*, 19:277–286, 1971.

[22] T. Ohta. Linkage disequilibrium with the island model. *Genetics*, 101:139–155, 1982.

[23] T. Ohta and M. Kimura. Linkage disequilibrium due to random genetic drift. *Genet.Res.*, 13:47–55, 1969.

[24] T. Ohta and M. Kimura. Development of associative overdominance through linkage disequilibrium in finite populations. *Genet Res*, 16(2):165–177, 1970.

[25] S Palsson and P Pamilo. The effects of deleterious mutations on linked, neutral variation in small populations. *Genetics*, 153(1):475–483, 1999 Sep.

[26] Petter Portin and Adam Wilkins. The evolving definition of the term "gene". *Genetics*, 205(4):1353–1364, Apr 2017.

[27] Pardis C Sabeti, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, Hans C Ackerman, Sarah J Campbell, David Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–7, Oct 2002.

[28] 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.

[29] P Stam. The distribution of the genome identical by descent in finite random mating populations. *Genet Res*, 35:131–135, 1980.

[30] J. A. Sved. The stability of linked systems of loci with a small population size. *Genetics*, 59:543–563, 1968.

[31] J. A. Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol*, 2:125–141, 1971.

[32] J. A. Sved. Heterosis at the level of the chromosome and at the level of the gene. *Theor Popul Biol*, 3(4), 1972.

[33] J A Sved. Correlation measures for linkage disequilibrium within and between populations. *Genet Res*, 91:183–192, 2009.

[34] J. A. Sved and M. W. Feldman. Correlation and probability methods for one and two loci. *Theor Popul Biol*, 4:129–132, 1973.

[35] J. A. Sved and B. D. Latter. Migration and mutation in stochastic models of gene frequency change. i. the island model. *J Math Biol*, 5(1), 1977.

[36] 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.

[37] John A Sved. The covariance of heterozygosity as a measure of linkage disequilibrium between blocks of linked and unlinked sites in hapmap. *Genet Res (Camb)*, 93(4):285–90, Aug 2011.

[38] John A Sved, Emilie C Cameron, and A Stuart Gilchrist. Estimating effective population size from linkage disequilibrium between

unlinked loci: Theory and application to fruit fly outbreak populations. *PLoS One*, 8(7):e69078, 2013.

[39] John A Sved and William G Hill. One hundred years of linkage disequilibrium. *Genetics*, 209(3):629–636, 07 2018.

[40] John A Sved, Allan F McRae, and Peter M Visscher. Divergence between human populations estimated from linkage disequilibrium. *Am J Hum Genet*, 83(6):737–743, 2008.

[41] H Tachida and C C Cockerham. Analysis of linkage disequilibrium in an island model. *Theor Popul Biol*, 29(2):161–97, Apr 1986.

[42] 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.

[43] Robyn S Waples. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, 7:167–184, 2006.

[44] RS Waples and C Do. LDNe: A program for calculating effective population size from data on linkage disequilibrium. *Molecular Ecology Notes*, 8:753–756, 2008.

[45] B S Weir. Inferences about linkage disequilibrium. *Biometrics*, 35:235–254, 1979.

[46] B S Weir and C C Cockerham. Behavior of pairs of loci in finite monoecious populations. *Theor Popul Biol*, 6(3):323–354, 1974 Dec.

[47] B S Weir and W G Hill. Effect of mating structure on variation in linkage disequilibrium. *Genetics*, 95(2):477–488, 1980.

[48] S Wright. Evolution in mendelian populations. *Genetics*, 16:97–159, 1931.

[49] Lei Zhao and Brian Charlesworth. Resolving the conflict between associative overdominance and background selection. *Genetics*, 203(3):1315–34, 07 2016.