Linkage Disequilibrium and Its Expectation in Human Populations

John A. Sved

School of Biological Sciences, University of Sydney, Australia

Linkage disequilibrium (LD), the association in populations between genes at linked loci, has achieved a high degree of prominence in recent years, primarily because of its use in identifying and cloning genes of medical importance. The field has recently been reviewed by Slatkin (2008). The present article is largely devoted to a review of the theory of LD in populations, including historical aspects.

Keywords: linkage disequilibrium, linked identity-by-descent, LD mapping, composite D, Hapmap

The Terminology of LD

It has been clear for many years that LD is prevalent across much of the human genome (e.g., Conrad 2006), and presumably any other organism when it is looked for (e.g., Farnir et al., 2000, in cattle). Given the widespread occurrence of the phenomenon, one might ask why such a term as 'disequilibrium' should be associated with it.

The term 'linkage disequilibrium' dates back to Lewontin & Kojima (1960). The term was introduced at a time when known genes were necessarily major genes, unknown at the molecular level, but known to have phenotypic, and potentially selective, consequences. Examples of LD among such major genes were even rarer, consisting of tightly linked 'supergenes' such as D, C, E in the Rh blood groups and the MNSs blood group system, and examples of genes tied up in inversions in *Drosophila*. The possibility of data from large numbers of closely linked single nucleotide human polymorphisms would have seemed remote.

Dominating 2-locus theory at that time was the finding dating back to Robbins (1918), reinforced in greater generality by the noted mathematician Hilda Geiringer (1944), that linked genes are, in the long run, expected to be associated at random in the population, regardless of the strength of the linkage. The theory behind this is given in the next section. Thus the default expectation, in the absence of any special force such as selective interactions (Lewontin & Kojima, 1960; Franklin & Lewontin, 1970), was for 'linkage equilibrium', the antithesis of LD. Essentially it was the class of genes being studied in the premolecular era that led to the emphasis on LD as some sort of abnormal expectation.

By the time that the possibility of LD mapping was realized (e.g., Ikonen, 1990) the terminology of LD was well established. Currently the term is increasingly used, with many thousands of PubMed references and over 100 Wikipedia references. Note also the confusion with 'lethal dose', 'learning disabilities' and other terms in literature searches involving the LD acronym. The chance of any more appropriate terminology seems to have passed, even if a simple and suitable term could be suggested. 'Allelic association' would seem a desirable term (e.g., Morton et al., 2001) but for the fact that the genes involved are, by definition, non-allelic. Other authors have preferred to use the term 'gametic phase disequilibrium' (e.g., Falconer & Mackay, 1996; Denniston, 2000) to take account of the possibility of LD between genes on different, a situation that can arise when many unlinked loci affect a selected quantitative character (Bulmer, 1971). From a utilitarian point of view, however, particularly in regard to LD mapping, it is the 'disequilibrium' rather than 'linkage' part of the terminology that needs to be replaced.

The Theory of LD

Robbins (1918) introduced the currently accepted measure of the departure from independence. Frequencies can be defined as follows:

The frequency of allele A at one locus is p_A.

The frequency of allele B at a second locus is p_B,

The frequency of the allele pair, or haplotype, AB is p_{AB}.

If the genes are combined at random in the population, then

$$p_{AB} = p_A p_B$$

Robbins used the symbol Δ to measure the departure from this expectation,

$$\Delta = p_{AB} - p_A p_B$$

It is this statistic, usually indicated as D, d or D_{AB} for genes A and B, that is nowadays used to measure LD,

Received 20 November, 2008; accepted 4 December, 2008.

Address for correspondence: John A. Sved, Biology A12, Sydney University NSW 2006, Australia. E-mail: jsved@usyd.edu.au

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

and given the name 'coefficient of LD' by Lewontin & Kojima (1960).

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

$$D' = (1 - c).\ D \qquad (1)$$

where D' is the corresponding coefficient one generation later (not to be confused with another LD measure introduced below). Robbins showed this by complete enumeration of matings, using symbols different to those used here.

Crow & Kimura in their 1970 textbook have a two-line derivation of the relationship as follows (see also Falconer & Mackay, 1996). With probability c any gamete produced in the current generation is a recombinant. The A gene is therefore combined with a random B gene, assuming random mating, giving the probability of AB in the gamete, and the next generation, as $p_A p_B$. Among 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

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

Noting that allele frequencies do not change between generations, so that $p'_A p'_B = p_A p_B$, this equation rearranges to give equation (1). Note that equation (1) predicts that unlinked genes, for which $c = ½$, are not expected to go to linkage equilibrium in a single generation. The reason for this apparently counter-intuitive result is that recombination is only effective in reducing LD in double heterozygote genotypes, and such genotypes occur with a maximum frequency of 50% in random mating populations, even with complete LD. All of these calculations assume that the population size is infinite. As becomes increasingly evident when dealing with human data, this is a crucial, and unlikely, assumption. The consequences of finite size are considered in detail in Section 2.

Measures of LD

The parameter D is the most straightforward measure of LD, but is very dependent on allele frequencies. The maximum range of D values is –0.25 to 0.25 when allele frequencies are 0.5. However, when allele frequencies are closer to 0 or 1, the range of values of D becomes much more restricted.

Numerous other parameters have been suggested. All of these normalize D in some way, so that the parameter is of the form D/C, where C is some function of gene frequencies or haplotype frequencies.

Perhaps the best known of these parameters is the correlation of frequencies, r, introduced by Hill & Robertson (1968). This is

$$r = \frac{D}{\sqrt{p_A(1 - p_A)p_B(1 - p_B)}}$$

The parameter r^2 relates directly to a 1 degree-of-freedom χ^2 testing the independence of frequencies at the two loci. It extends directly to a muliple degree-of-freedom χ^2 statistic for multiple alleles (Zhao et al., 2005). A desirable property of the correlation is that the r^2 parameter directly measures the strength of association between a causal variant and a genotyped (neutral) variant (e.g., Sved, 1968; Ardlie et al., 2002).

A second frequently used parameter is D' (Lewontin, 1964), in which D is divided by its maximum absolute value for the observed allele frequencies. The range of D', as for r, is –1 to 1.

Hedrick (1987) in an influential review has argued against the use of r and in favour of D', on the grounds that r cannot take the full range of values if allele frequencies are unequal. For example, for the case $p_A = 0.4$, $p_B = 0.1$, D has the range –0.04 to + 0.06, and r has the range –0.27 to + 0.41. D', on the other hand, still by definition has the range –1 to 1. See Wray (2005) for constraints on r^2 implied by allele frequencies.

Opposed to this view is the question of whether a measure of LD ought to be able to take the full range of values for this set of allele frequencies. The fact that allele frequencies at the two loci are unequal is not without information. It implies that there cannot be a complete correlation of frequencies. The D' parameter discards this information, whereas the r parameter takes it into account.

The question of which parameter should be used thus comes down to a question of whether it is necessary to take allele frequencies into account. From the point of view of understanding the dynamics of two loci in a population, it is not clear that it is useful to calculate LD conditional on allele frequencies. On the other hand for mapping purposes it is important to take into account the allele frequencies.

The question of what parameter is optimal for disease gene mapping has been widely debated. Devlin & Risch (1995) recommend the use of the parameter δ, defined as

$$\delta = \frac{D}{p_a p_{AB}}$$

where p_a is the frequency of a disease gene and p_{AB} is the frequency of the haplotype involving the normal gene and the more common allele at a linked marker locus. Morton et al. (2001), however, claim an advantage for ρ, defined as

$$\rho = \frac{D}{p_a p_B}$$

Both δ and ρ are closely related to D' in the situation of a rare disease gene arising in a population closely linked to an informative marker.

LD Estimated From Diploid Data

A problem in most cases of LD measurement comes from the diploid nature of the data. All of the above LD statistics assume that haploid data are available. Usually, therefore, computer programs are needed to infer haplotypes. Adkins (2004) for small data sets and

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

Marchini et al. (2006) for large data sets have shown that most such computer programs; for example, PHASE (http://www.stat.washington.edu/stephens/phase), perform well. However, the error rate can be as much as 5% to 6% if family data are not available.

Two methods are available for estimating LD directly from diploid data. The first uses the correlation between the number of alleles at each of two loci. The values at each of two loci, 0, 1 or 2, are respectively the x and y values. Under conditions of random mating or of simple inbreeding this correlation has the same expectation, r, as the correlation between alleles in haplotypes. The method is implemented in the computer program PLINK (http://pngu.mgh.harvard.edu/purcell/plink/).

A second, related, method, is the 'composite LD measure' (Weir, 1979 — attributed to P. M. Burrows). In converting genotype numbers to haplotype numbers, each possible haplotype is counted once. Genotypes such as $A_1A_2;B_1B_2$ contribute haplotypes $1/4A_1B_1$; $1/4A_1B_2$; $1/4A_2B_1$ and $1/4A_2B_2$. Two of these will be the actual haplotypes and two incorrect. However, the incorrect haplotypes are expected to be present at the product of the allele frequencies, and will thus dilute, but not bias, the actual haplotype frequencies.

Calculating haplotypes in this manner, assuming random mating, the expected coefficient of LD among these composite haplotype frequencies is expected to be just $D/2$. Similarly the expected correlation is $r/2$.

Furthermore numbers, rather than frequencies, of composite haplotype can be cumulated, with each genotype giving four possible haplotypes. This leads to a valid χ^2 test for association with 1d.f., even though the marginal numbers are necessarily multiples of 2. The same applies to a test for independence of two loci with multiple alleles, which gives a χ^2 with $(m - 1)(n - 1)$ df where there are m and n alleles at the two loci respectively.

The composite LD measure has been tested using the Hapmap data set (see next section). One such study (unpublished) used the data from chromosome 21. The average correlations for particular recombination values obtained using the diploid data set, doubled to take account of the diluting effect, were essentially indistinguishable from the r values calculated from the haploid data set assigned by Hapmap (Figure 1).

It should be noted that these two methods are only feasible for two loci. For more than two loci the number of possible haplotypes becomes large, and it is necessary to rely on assigned haplotypes.

Human Population Data

Sequencing of the human genome has opened up the possibility of studying LD between markers with known nucleotide positions. Earlier data sets on human polymorphism have been superseded by the Hapmap study (The International Hapmap Consortium, 2005, 2007). This study currently involves 3.1 million SNPs in 270 individuals from four populations (YRI — Yoruba in Ibadan, Nigeria; CEU — Northern and Western European ancestry from Utah; CHB — Han Chinese in Beijing; JPT — Japanese in Tokyo). The data are available for download at http://www.hapmap.org.

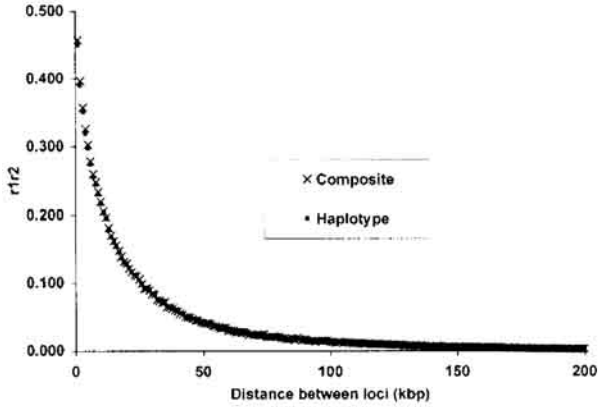

Figure 1

Comparison of values of r_1r_2 calculated using diploid data and assigned haploid data for the comparison of YRI (Africa) vs. CEU (Europe) using Hapmap chromosome 21 data.

Many conclusions related to LD have already emerged from the Hapmap study. For example it has been possible to identify pairs of individuals within populations who share a segment that can be traced as coming from a recent common ancestor, less than 10% of all sequence pairs in YRI but 20% to 30% in the other populations. Similarly large regions of homozygosity have been identified within individuals.

It has led to the identification of around 200 regions where natural selection has been identified by hitchhiking of linked SNPs (Sabeti et al., 2007). A genome-wide comparison of a subset of nonsynonymous versus synonymous mutations has also shown higher levels of differentiation for the nonsynonymous mutations. This has been interpreted as showing higher levels of purifying selection for protein-coding genes.

Regions of high and low recombination have also been identified. This has led to estimates of the distribution of crossing-over over the genome at the nucleotide level (Myers et al., 2005).

The main utility of the Hapmap study is in disease gene identification. Early studies using LD to identify disease genes; for example, MacDonald et al., (1991) for Huntington's disease, were necessarily based on markers with relatively low LD values. The Hapmap coverage is now sufficient that there is a high likelihood that a genome-wide assay for a single nucleotide disease gene will detect a mapped SNP in almost complete LD with the disease gene. Genome-wide association (GWA) studies now appear for a wide variety of traits including cancer susceptibility, diabetes, Alzheimer's, nicotine dependence and so on, with a PubMed 'GWA' search identifying more than 100 such studies in the past year.

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

Although the Hapmap study provides the most comprehensive set of SNPs, it is restricted to four samples. Other data sets have considered a wider set of populations, concentrating on smaller chromosomes or chromosome regions. Service et al. (2006) surveyed 200 individuals from 12 populations or isolates of various size for chromosome 22. They found fewer 'holes', regions of low LD, in the smaller isolates. Conrad et al. (2006) surveyed 927 individuals from 52 populations for 12 Mb of sequence from various chromosomal locations. Both studies support the Hapmap conclusions of higher levels of LD in non-African populations. Conrad et al. found that the four Hapmap populations contain more than 80% of the haplotypes found in their wider study.

Finite Size Effects

The theory given in Section 1.2 shows a slow rate of approach to linkage equilibrium for the most closely linked loci, when the multiplier term in equation (1) is very close to 1. However in the absence of any systematic force opposing the effect of recombination, such as selective interactions (Lewontin, 1964), there seemed no reason to doubt that linkage equilibrium would eventually be reached. This situation changed in 1968–1969 when three papers (Hill & Robertson, 1968; Ohta & Kimura, 1969; Sved, 1968) showed that finite-size effects (haplotype segregation) could potentially swamp the effects of recombination.

This effect now seems so obvious that it is difficult to believe that it was once ignored. It appears most starkly in the case of human disease genes in small isolates. Where the two mutant alleles of a homozygote are likely to have descended from a single original mutation, it is clear that homozygosity is also to be expected at any very closely linked locus. The increased frequency of double homozygotes implies that there cannot be linkage equilibrium, an argument that is expanded in a following section.

It is of interest to note that, as in much else of population genetics theory, the first statement of this effect goes back to one of the three founders of population genetics, R. A. Fisher, J. B. S Haldane and S. Wright; in this case, Haldane (1949). He showed that inbreeding, in general, leads to the association of homozygous genotypes at linked loci. This effect was investigated in detail by Bennett & Binet (1956) for the inbreeding system of mixed self-fertilization and random mating. However, it was not extended at the time to the case of inbreeding due to finite size.

The Expectation of LD in a Finite Population

Each of the three papers cited above gave expressions for the amount of LD expected in a finite population under the Wright-Fisher model. Hill & Robertson (1968) produced a very simple solution, noting that LD is increased by drift and decreased by recombination. They showed that a balance between the two forces occurs when

$$r^2 = \frac{1}{4Nc}$$

This equation becomes infinite for small values of $N c$, whereas the upper bound of r^2 should be 1. However Hill (personal communication) has shown that a small correction to the derivation leads to the equation

$$r^2 = \frac{1}{1+4Nc} \tag{2}$$

which is similar for large values of N c and has the correct upper bound for small values.

Sved (1968) considered a model in which the allele frequencies are held at 50% by heterozygote advantage at both loci. This led to the equation

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

which is the same as equation (2) for the special case when all allele frequencies are one half. This result was extended by Avery (1978).

Ohta & Kimura (1969) considered the same model as Hill & Robertson. Using difiusion equations, they calculated the expected values of $\sigma^2 = E[D^2]/E[p_A(1 - p_A)p_B(1 - p_B)$. The ratio of the expectations is not the same as the expectation of the ratio, although Ohta & Kimura noted that the two are not very different. Differences between the two were further explored by Hill (1977) and Song & Song (2007).

Ohta & Kimura (1969a) further calculated the mean steady state value of σ^2 under a two-locus infinite allele mutation model. The model makes the implicit assumption that only two alleles segregate at a locus at one time, making it equivalent to an infinite-site mutation model. Ohta & Kimura calculated that the limiting value for very low recombination values was 5/11, a value also derived by Hill & Weir (1988) and McVean (2002).

It is convenient to express the expected steady state value of the ratio of expectations in the form

$$E[\sigma^2] = \frac{1}{\alpha+4Nc} \tag{3}$$

where α is equal to 11/5 for the mutation model considered above. Note that equation (2) gives a value of $r^2 = 1$, assuming that σ^2 is equivalent to r^2. This is a limiting value, assuming that a steady state has been reached at all loci. Under a mutation model, however, some locus pairs will have reached the limiting value while others have lower values of σ^2 or r^2, leading to a mean value of around 1/2.

Calculations using Hapmap (Tenesa et al., 2007, Figure 1) confirm that the mean value of r^2 asymptotes to around 1/2 for closely linked loci on all chromosomes. However the most closely linked locus pairs were not considered in this calculation, and the limiting value may be somewhat higher (see The International Hapmap Consortium, 2007).

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

Fixation Bias

The ratio of expectations calculation conceals a difficulty of the calculation of the expected ratio. When fixation occurs at either locus, the value of r is indeterminate. Strictly speaking, therefore, any recurrence relationship between generations for a quantity such as r^2 needs to be made conditional on non-fixation (Hill & Robertson, 1968; Hill, 1977). There is no such requirement for the calculation of the ratio of expectations.

As argued above, LD develops because of chance fluctuations in a finite population. Such fluctuations can, by definition, sometimes be large and sometimes small. The larger fluctuations are those which lead to higher values of r^2. Correspondingly, however, higher fluctuations are also likely to lead to fixation, in which case r^2 cannot be measured. The rise in r^2 will therefore be less than expected if this effect is not taken into account.

A second, related, effect due to fixation, described as 'fixation bias' by Sved et al. (2008), can be seen when r^2 starts at a non-zero value in a population. This is most easily seen in a numerical example, taking haplotype frequencies as:

AB	0.4
Ab	0.1
aB	0.1
ab	0.4

The value of r in such a case comes to 0.6.

What is expected to happen in a finite population some generations after it has such values? Assuming that the loci are sufficiently closely linked so that recombination can be ignored, haplotypes can be treated as analogous to multiple alleles at a single locus. The four haplotypes will eventually be reduced to three by fixation, and then to two and eventually one. If there is only one haplotype, then no calculation of r values is possible. However at the preceding stage, when there are two haplotypes, r values can be calculated if the remaining haplotype pairs are either *AB* and *ab*, or alternatively *Ab* and *aB*. Any other haplotype pair involves fixation at either the *A* or *B* locus, in which case r is indeterminate.

Clearly the relative probability that *AB* and *ab* will be the last remaining haplotypes, as opposed to the *Ab* and *aB* haplotypes, is very high. Calculations given by Sved et al. (2008), based on theory by Littler (1975), show that the mean value of r, given that only two haplotypes remain, is $r(7 - r^2) / (3 + 5r^2)$, which comes to 0.92. Thus the value of r has increased very significantly. Note that reversing the haplotype frequencies to give an initial value of $r = -0.6$ would predict a final value of $r = -0.92$. It is the absolute value of r that is predicted to rise.

The value of r^2 increases in this case to its maximum value of 1, as predicted by (2). But the calculations leading to (2) predict only a rise in the mean value of r^2, not in the mean value of r. The effect of fixation bias is complementary to the well-established effect of haplotype segregation.

Estimating Effective Population Size

Equations (2) and (3) provide a means of estimating population size using levels of LD in a population. Hill (1981) discussed in detail the theory for doing this, including correcting for sample size. The equation in this case, modified from equation (3), is approximately

$$E[r^2] = \frac{1}{\alpha + 4Nc} + \frac{1}{n} \tag{4}$$

where n is the sample size. When dealing with values of $4N\ c > 1$, the $1/n$ term can make a substantial contribution. This term can be somewhat inaccurate, particularly for high values of r^2, for which a more accurate correction is given by Sved et al. (2008).

A problem in applying equation (4) is the large variance in r^2 values (Hill, 1977; Golding, 1984). This has been overcome to some extent by the large number of locus pairs available in such data sets as Hapmap, of the order of 2 x 10^7 pairs of loci with estimated recombination less than 0.1cM. Such a figure exaggerates the power of the data since most such locus pairs are not independent of each other, and the data set relies on a much smaller number of crossovers since the common ancestors.

Tenesa et al. (2007) calculated from the Hapmap data set an effective population size for YRI (Yoruba — Africa) of 7,500. This estimate is highly influenced by ancestral population sizes. Recent size expansion would be expected to reduce the value of r^2, but for the most closely linked loci any change in r^2 would be very slow. The estimated size for the CEU sample (Central European) was 3,100, presumably reflecting bottlenecks in the founder populations.

Note that these estimates are dependent on having the correct estimates for c, the recombination frequency. While overall estimates of recombination per chromosome are well characterized from counting of chiasmata (Kong et al., 2004), the known occurrence of recombination hotspots, and particularly recombination coldspots, can have marked effects on estimates of c. A fine scale map is available, the Oxstats map, taking into account population data (Myers et al., 2005).

Estimating Separation Times

LD estimates can also be used to estimate separation times of human populations. The theory, given by de Roos et al. (2008), is essentially infinite-size population theory, but relies on chance LD occurring in the ancestral population.

When separation between populations occurs, as in the out-of-Africa hypothesis, each individual locus pair is expected to show a certain level of LD. Overall levels of r^2 are expected to be maintained in the separate populations, perhaps even rising because of bottlenecking. However independently of population

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

size, levels of r are expected to fall by a fraction $(1 - c)$ in each generation as given by equation (1). The value of r_1r_2 from two independent populations is expected to fall by a fraction $(1 - c)^2$. After T generations, the value of r_1r_2 is expected to fall by a fraction $(1 - c)^{2T}$.

Sved et al. (2008) used this theory to estimate T, the number of generations separating the current African and European populations. They assumed that the current levels of r^2 reflect ancestral values of r^2 for the most closely linked loci in the YRI African population. The calculation led to estimates of T of lower than 1,000 generations, less than one-half of current estimates (Fagundes et al., 2007). Migration between populations was suggested as a possible reason for the low T estimate.

Fixation bias was an important factor in the calculation of separation times. The lowest recombination frequencies gave negative estimates of separation time, reflecting the fact that $r_{YRI}r_{CEU}$ values are greater than r^2_{YRI} values. Just such an effect is expected if fixation occurred during bottlenecking leading to the current CEU population. The fixation bias effect becomes smaller as recombination increases, although it still causes an approximate 10% bias for the largest recombination frequencies for which a positive $r_{YRI}r_{CEU}$ signal could be detected (Sved et al., 2008).

LD, LIBD and Homozygosity

The main focus of attention to date has been on measures of LD such as r. As pointed out above, there is a second, rather different, way of looking at LD. Where a disease gene is associated with a SNP or other genotype marker in a small population, this is clearly because the two genes have been inherited through the same pathways. In the case of recessive genes, this association is noted because of homozygosity at the disease gene locus, implying descent from a recent ancestor.

How is this passage of genes through the same pathways related to LD? And where does homozygosity come into the picture?

Linked Identity-by-Descent (LIBD)

The concept of Linked Identity-by-Descent, although not the specific name, was introduced in Sved (1971) and Sved & Feldman (1973). It is similar to the concept of chromosome segment homozygosity (CSH — Hayes et al., 2003).

Figure 2 shows the inheritance of linked markers. LIBD of two haplotypes implies the identity at both loci through the same pathways. Note the difference with the parameter F_{11} of Cockerham & Weir (1973), denoting IBD at each of two linked loci, but where either the same or different pathways can be involved.

The probability of LIBD can be given the symbol L. On each of the pathways of Figure 2 the probability of no crossovers is defined as f_{AB}. Since events in the two pathways are independent,

$$\mathrm{L} = f^2_{\mathrm{AB}} \qquad (5)$$

The Relationship Between LIBD and LD

It is well known that for a single locus, inbreeding can be defined either in terms of probability or correlation, specifically the probability of IBD or the correlation between uniting gametes. The simple relationship between the two can be seen, related to the argument of Crow & Kimura (1970), p 66. IBD implies a correlation of one. Conversely, non-IBD implies a correlation of zero. Assuming additivity, the overall correlation is simply the probability of IBD (Figure 2).

A similar argument can be made for the two-locus parameters. Amongst haplotypes descended from an

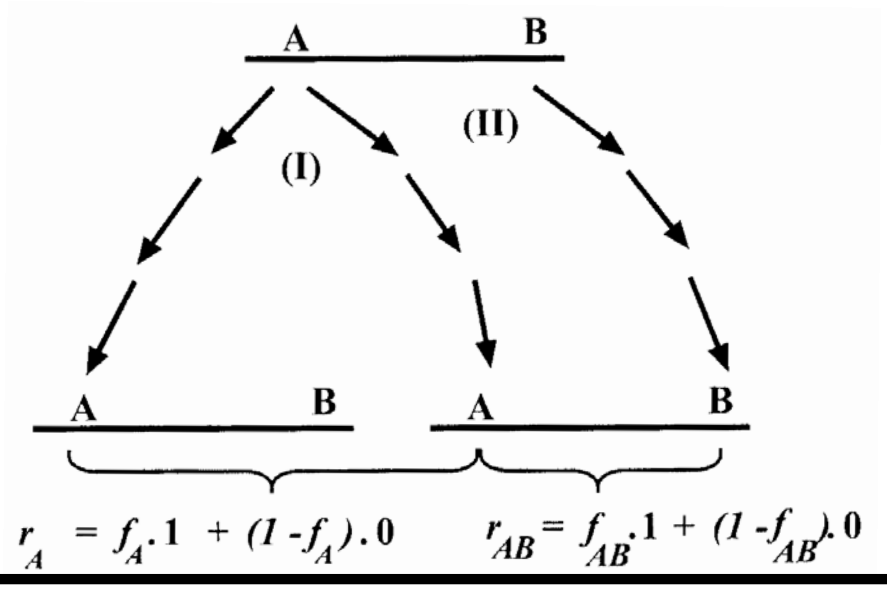
A
B
(II)
(I)
A
B
A
B
$r_A = f_A.1 + (1 - f_A).0$
$r_{AB} = f_{AB}.1 + (1 - f_{AB}).0$

Figure 2

A comparison of single locus and two locus correlation and probability parameters.

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

ancestral haplotype, absence of crossingover ensures a correlation of one. Assuming random mating, any crossover is sufficient to connect the *A* locus to a random *B* locus, implying a correlation of zero. Overall, therefore, the expected correlation, $E[r_{AB}]$, is equal to the probability of no crossingover f_{AB}. Taking into account Equation (5),

$$L = f_{AB}^2 = E[r_{AB}^2]. \quad (6)$$

The difference between the parameters *L* and r^2 needs to be emphasized. One is a probability parameter referring to the probability of no crossing over in a genealogy, the other is a correlation parameter referring to frequencies in the current population. Equality between the two implies that the probability and correlation parameters give alternative descriptions of the same LD process.

In fact, Hayes et al. (2003) pointed out that with the high density of markers now available, it is possible to directly estimate a parameter analogous to *L*. They showed that this can be estimated with greater accuracy in populations than possible for estimating r^2.

Equation (6) allows a simple calculation for $E[r^2]$, since the recurrence relationship for *L* under the Wright-Fisher model can readily be written down. At equilibrium, the value of *L* is approximately $1/(1 + 4N c)$. Equating r^2 and *L* then gives the same steady state expectation as given in equation (2).

The Relationship Between Homozygosity and LIBD

When considering joint homozygosity, the obvious parameter to use is the frequency of double homozygotes (Sabatti & Risch, 2002; Hayes et al., 2003). Alternatively, this frequency can be thought of as the probability that two haplotypes sampled from the population are identical in state. Sabatti & Risch considered in detail the frequency of double homozygotes in terms of the LD parameter *D*. The expectation includes terms in D^2 and also in $D(p_A - 1/2)(p_B - 1/2)$. No simple relationship can be found just in terms of *D* or *r*.

The present author (Sved, 1971) attempted to overcome this problem by looking at joint homozygosity in a different way. The definition again involved selecting pairs of haplotypes from a population with two alleles at each of two loci. However, the definition was in terms of homozygosity at the *B* locus conditional on selecting homozygous genes at the *A* locus. And rather than selecting homozygotes *AA* and *aa* with relative frequencies p^2_A and p^2_a, the definition required selecting homozygous genotypes with frequency p_A and p_a respectively. The probability of homozygosity at the *B* locus was then calculated in two ways, (1) using frequency parameters, (2) using a conditional LIBD probability. Equating the two led to the same relationship as equation (6), except that the LIBD probability was conditioned on selecting the same allele at the *A* locus in the pair of haplotypes.

Coalescence arguments (e. g. Hudson, 1985) are based on models incorporating both mutation and recombination. The above model assumes that all *A* alleles coalesce to a single ancestral allele, and similarly all *a* alleles to a different ancestral allele. A two-allele two-locus model thus has the contradictory requirement of low probabilities of mutation at both loci while at the same time requiring segregation at both loci. The necessity for conditioning thus appears to come from attempting to coerce a two-allele model into a coalescence framework.

Homozygosity Mapping

Although not directly related to any homozygosity measure, the topic of homozygosity, or autozygosity, mapping should be mentioned here. In this case recessive disease-causing loci are identified through increased homozygosity in a particular region (e.g., Wang et al., 2008).

Autozygosity mapping is now less precise than LD mapping, given the number of SNPs currently available for genome-wide association studies. It does possess one advantage, in that chromosomal regions of interest can be identified by studying just affected individuals, rather than needing a direct comparison with unaffected individuals.

Junctions and Identity of Chromosome Regions

In dealing with finite size LD, it is clear that two-locus measures can provide only a partial description. The use of multi-locus coefficients can partly overcome these deficiencies, and Hill & Weir (2007) have calculated expectations of three-locus LD measures under a Wright-Fisher model.

Ultimately, however, the LD structure of a population is determined by the lengths of identical segments in the population. In the terminology of Fisher (1954), it is the mapping of 'junctions', points of recombination between nonidentical segments, that is important. Theory of the number of recognisable junctions in a finite population has been given by Stam (1980) (see also Macleod et al., 2005). As mentioned above, the Hapmap study provides information at a sufficiently detailed level to allow determination of such junctions. This should allow the theory of Stam to be applied to obtain more detailed insights than currently available on the ancestry of present-day populations.

Acknowledgments

Comments from Peter Visscher, Bill Hill and Ian Franklin in the writing of this review are gratefully acknowledged.

References

Adkins, R. M. (2004). Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. *BMC Genetics, 5*, 22.

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews, Genetics, 3*, 299–309.

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

Avery, P. J. (1978). Effect of finite population-size on models of linked over-dominant loci. *Genetic Research, 31,* 239–254.

Bennett, J. H., & Binet, F. E. (1956). Association between Mendelian factors with mixed selfing and random mating. *Heredity, 10*, 51–56.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *American Naturalist, 105,* 201–211.

Cockerham, C. C., & Weir, B. S. (1973). Descent measures for two loci with some applications. *Theoretical Population Biology, 4,* 300–330.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics, 38,* 1251–1260.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in holstein-friesian, jersey and angus cattle. *Genetics, 179,* 1503–1512.

Denniston, C. (2000). Equivalence by descent, pedigree analysis with inbreeding and gametic phase disequilibrium. *Annals of Human Genetics, 64,* 61–82.

Devlin, B., & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics, 29,* 311–322.

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., & Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Science, USA, 104,* 17614–17619.

Falconer, D. S. & Mackay, T. F. C. (1996). *Quantitative genetics* (4th ed.). Essex, England: Longman.

Farnir, F., Coppieters, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., & Georges, M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Research, 10,* 220–227.

Fisher, R.A. (1954). A fuller theory of junctions in inbreeding. *Heredity, 8,* 187–197.

Franklin, I. R., & Lewontin, R. C. (1970). Is the gene the unit of selection? *Genetics, 65,* 707–734.

Geiringer, H. (1944). On the probability theory of linkage in Mendelian heredity. *Annals of Mathematical Statistics, 15*, 25–57.

Golding, G. B. (1984). The sampling distribution of linkage disequilibrium. *Genetics, 108,* 257–274.

Haldane, J. B. (1949). The association of characters as a result of inbreeding and linkage. *Annals of Eugenics, 15,* 15–23.

Hayes, B., Visscher, P., McPartlan, H., & Goddard, M. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research, 13,* 635–643.

Hedrick, P. W. (1987). Gametic disequilibrium measures, proceed with caution. *Genetics, 117,* 331–341.

Hill, W. G. (1977). Correlation of gene frequencies between neutral linked genes in finite populations. *Theoretical Population Biology, 11,* 239–248.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetic Research, 38,* 209–216.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium infinite populations. *Theoretical and Applied Genetics, 38,* 226–231.

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria infinite populations. *Theoretical Population Biology, 33,* 54–78.

Hill, W. G., & Weir, B. S. (2007). Prediction of multilocus inbreeding coefficients and relation to linkage disequilibrium in random mating populations. *Theoretical Population Biology, 72,* 179–185.

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

Ikonen, E., Palo, J., Ott, J., Gusella, J., Somer, H., Karila, L., Palotie, A., & Peltonen, L. (1990). Huntington disease in Finland, linkage disequilibrium of chromosome 4 rp haplotypes and exclusion of a tight linkage between the disease and d4s43 locus. *American Journal of Human Genetics, 46,* 5–11.

Kong, X., Murphy, K., Raj, T., He, C., White, P. S., & Matise, T. C. (2004). A combined linkage-physical map of the human genome. *American Journal of Human Genetics, 75,* 1143–1148.

Lewontin, R. C. (1964). The interaction of selection and linkage. i. general considerations; heterotic models. *Genetics, 49,* 49–67.

Lewontin, R. C. & Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution, 14,* 458–472.

Littler, R. A. (1975). Loss of variability at one locus in a finite population. *Mathematical BioSciences, 25,* 151–163.

MacDonald, M. E., Lin, C., Srinidhi, L., Bates, G., Altherr, M., Whaley, W. L., Lehrach, H., Wasmuth, J., & Gusella, J. F. (1991). Complex patterns of linkage disequilibrium in the Huntington disease region. *American Journal of Human Genetics, 49,* 723–734.

MacLeod, A. K., Haley, C. S., Woolliams, J. A., & Stam, P. (2005). Marker densities and the mapping of ancestral junctions. *Genetic Research, 85,* 69–79.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z., Munro, H., Abecasis, G., & Donnelly, P. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics, 78,* 437–450.

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press

McVean, G. (2002). A genealogical interpretation of linkage disequilibrium. *Genetics, 162,* 987–991.

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., & Collins, A. (2001). The optimal measure of allelic association. *Proceedings of the National Academy of Sciences, USA, 98,* 5217–5221.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science, 310,* 321–324.

Ohta, T. & Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetic Research, 13,* 47–55.

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

Robbins, R. B. (1918). Some applications of mathematics to breeding problems iii. *Genetics, 3,* 375–389.

Sabatti, C. & Risch, N. (2002). Homozygosity and linkage disequilibrium. *Genetics, 160,* 1707–1719.

Sabeti, P., Varilly, P., Fry, B., & et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature, 449,* 913–918.

Service, S., DeYoung, J., Karayiorgou, M., Roos, J. L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J. A., Heutink, P., Aulchenko, Y., Oostra, B., van Duijn, C., Jarvelin, M. R., Varilo, T., Peddle, L., Rahman, P., Piras, G., Monne, M., Murray, S., Galver, L., Peltonen, L., Sabatti, C., Collins, A., & Freimer, N. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for Genome-wide association studies. *Nature Genetics, 38,* 556–560.

Slatkin, M. (2008). Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. *Nature Reviews: Genetics, 9,* 477–485.

Song, Y. & Song, J. (2007). Analytic computation of the expectation of the linkage disequilibrium coefficient r2. *Theoretical Population Biology, 71,* 49–60.

Stam, P. (1980). The distribution of the genome identical by descent in finite random mating populations. *Genetic Research, 35,* 131–135.

Sved, J., A. McRae, A., & Visscher, P. (2008). Divergence between human populations estimated from linkage disequilibrium. *American Journal of Human Genetics (*November 12 Epub).

Sved, J. A. (1968). The stability of linked systems of loci with a small population size. *Genetics, 59,* 543–563.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology, 2,* 125–141.

Sved, J. A. & Feldman, M. W. (1973). Correlation and probability methods for one and two loci. *Theoretical Population Biology, 4,* 129–132.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research, 17,* 520–526.

The International Hapmap Consortium. (2005). A haplotype map of the human genome. *Nature, 437,* 1299–1320.

The International Hapmap Consortium. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature, 449,* 851–861.

Wang, S., Haynes, C., Barany, F., & Ott, J. (2008). Genome-wide autozygosity mapping in human populations. *Genetic Epidemiology* (September 22 Epub).

Weir, B. S. (1979). Inferences about linkage disequilibrium. *Biometrics, 35,* 235–254.

Wray, N. (2005). Allele frequencies and the r2 measure of linkage disequilibrium, impact on design and interpretation of association studies. *Twin Research and Human Genetics, 8,* 87–94.

Zhao, H., Nettleton, D., Soller, M., & Dekkers, J. C. M. (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and qtl. *Genetic Research, 86,* 77–87.

https://doi.org/10.1375/twin.12.1.35 Published online by Cambridge University Press