One Hundred Years of Linkage Disequilibrium

John A. Sved*,[1] and William G. Hill†
*Evolution and Ecology Research Centre, University of New South Wales, Sydney, 2052, Australia and †Institute of Evolutionary Biology, University of Edinburgh, EH9 3FL, United Kingdom

ABSTRACT One hundred years ago, the first population genetic calculations were made for two loci. They indicated that populations should settle down to a state where the frequency of an allele at one locus is independent of the frequency of an allele at a second locus, even if these loci are linked. Fifty years later it was realized what is obvious in retrospect, that these calculations ignored the effect of chance segregation of linked loci, an effect now widely recognized following the association of closely linked markers (SNPs) with rare genetic diseases. Linkage disequilibrium is now accepted as the norm for closely linked loci, leading to powerful applications in the mapping of disease alleles and quantitative trait loci, in the detection of sites of selection in the human genome, in the application of genomic prediction of quantitative traits in animal and plant breeding, in the estimation of population size, and in the dating of population divergence.

KEYWORDS LD; recombination; population size; genetic drift; GWAS; genomic prediction

THE humble beginnings of the study of linkage disequilibrium (LD) can be dated back to 1918, 10 years after the Hardy–Weinberg law introduced population genetics for a single locus. Robbins (1918), in volume 3 of *GENETICS*, developed the original theory for two loci, taking into account the then relatively new concepts of linkage and recombination. LD has now become a huge topic, with nearly 25,000 keyword citations in the most recent PubMed database. In this article we provide a history of the development of LD theory and explain and illustrate the many applications of LD in pure and applied genetics.

Early Theory

The basic theory for two loci, A and B, is simple (see Box 1). The frequency in a particular population of allele A at the first locus is p_A and of allele B at the second locus is p_B, and the combined haplotype AB has frequency p_{AB}. Robbins introduced the measure Δ, nowadays usually denoted D, to describe the extent to which alleles at the two loci depart from random combination, where $D = p_{AB} - p_A p_B$. Although Robbins used an earlier recombination parameter, essentially what he showed was that one generation of random mating reduces the value of D in the population to $D(1 - c)$, where c is the recombination rate and quantifies the amount of crossing over or recombination.

For each generation, the same factor $1 - c$ applies so that, after t generations, the value of D is reduced to $D(1 - c)^t$. Over time, therefore, for any pair of loci undergoing recombination, D approaches zero. So in a closed population and in the absence of selection and other forces, genes are expected to combine at random in the population, *i.e.*, to be in linkage equilibrium (LE).

Over the 50 years following Robbins' initial work, the theory was extended in various ways, most importantly incorporating selection. An influential article by Kimura (1965) showed that "quasi linkage equilibrium" could be attained under the assumption of weak selection. In contrast, it was shown that if particular gene combinations were favored, then LD could be maintained in populations (Lewontin and Kojima 1960). Such "equilibrium models" were studied in considerable detail, frequently in the context of the evolution of recombination (*e.g.*, Bodmer and Felsenstein 1967; Karlin and Feldman 1970). Franklin and Lewontin (1970) extended the theory, predicting the possibility of LD over large regions of the genome due to multiplicative selection interaction. Although much of the emphasis of these articles was on LD, the general conclusion was that such LD required strong epistatic selection compared with the amount of recombination, and therefore could apply only to a minority of locus pairs

doi: https://doi.org/10.1534/genetics.118.300642
Manuscript received February 8, 2018; accepted for publication April 15, 2018.
[1]Corresponding author: E-mail: j.sved@unsw.edu.au

Box 1 Definitions and Expected Changes in LD Measures

Frequencies in a population:
Frequency of allele A at first locus = p_A.
Frequency of allele B at second locus = p_B.
Frequency of allele pair (haplotype) AB = p_{AB}.

LD measures:

$$D = p_{AB} - p_A p_B\,;$$

where D is the coefficient of LD.

$$D' = D/\text{maximum value of } D;$$

is an LD measure designed to have a range from -1 to 1.

$$r^2 = D^2/[p_A p_B(1-p_A)(1-p_B)];$$

where r is the correlation of allele frequencies.

Expectation from generation $t-1$ to t (infinite population):

$$D[t] = (1-c)D[t-1];$$

where c is the recombination frequency.

Expected value at equilibrium in an infinite population:

$$E[D] = 0$$

Fifty Years Ago: LD Progresses from Exception to Expected

In the two-locus theory of that time it had been assumed implicitly that gene frequencies could be manipulated as if the population was infinite, thereby ignoring the possibility that LD could be produced solely by the joint segregation of linked genes. The emphasis on LD rather than LE as the norm changed when attention was drawn to the effect of finite size of populations (Hill and Robertson 1968; Sved 1968; Ohta and Kimura 1969). Such effects became obvious later following the study in human populations, for example in Finland (Hästbacka *et al.* 1992), of rare disease alleles and associated linked polymorphisms.

Of equal importance to the increasing emphasis on LD was the realization of the extent of closely linked polymorphic sites in populations. From a present-day point of view, it is difficult to appreciate the background of population genetics theory in the premolecular era. It was well known, from *Drosophila* for example, that there are many cases of very closely linked loci. What was less clear, however, was whether there are many cases of closely linked *polymorphic* loci in populations. In retrospect, the lack of thought given to this possibility seems surprising given the background of Fisher's multi-locus model for quantitative traits, Wright's concept of relationship as a correlation, and the subsequent widespread acceptance of the polygenic model of inheritance of continuous traits (*e.g.*, Falconer's 1960 textbook).

This situation changed following the work of Lewontin and Hubby (1966) in *Drosophila*, whose study has been the subject of a previous *Perspectives* article (Charlesworth *et al.* 2016), and a study by Harris (1966) in humans. These authors were the first to address systematically the question of what proportion of loci were polymorphic, focusing on loci where a protein product could be visualized on a gel. Their conclusion was that at least one third of such loci in both species were polymorphic, implying that there had to be many thousands of such loci, and that many would have to be closely linked. The first systematic study of polymorphism at the DNA level, at the *Adh* locus in *Drosophila* (Kreitman 1983), indicated that DNA polymorphism vastly exceeded the amount of detectable polymorphism at the protein level. Later studies at the genome level in humans (International HapMap Consortium 2005) and more generally have strongly borne out this early finding.

One result of this history is the usage of the term LD. In modern usage it usually applies to closely linked loci, where the idea that linked SNPs within linkage blocks are somehow in "disequilibrium" seems counterintuitive. The LD term is also used to describe the situation for unlinked loci (*e.g.*, see section on estimation of population size below), where the term is especially inappropriate. In retrospect, the term "allelic association" (see, *e.g.*, Morton *et al.* 2001) would probably have been more suitable.

Measures of LD

The range for D is -0.25 to 0.25, but it depends on allele frequencies (see Box 1): the maximum and minimum values can be attained only if the frequencies of both alleles (p_A and p_B) are 0.5. These allele frequencies in the population are referred to by Weir and Goudet (2017) as "allele probabilities" to clarify that these are the expected values of the allele proportion. Correspondingly, the haplotype frequencies are haplotype proportions in the population. If, for example, $p_A = 0.3$ and $p_B = 0.1$, the possible range is asymmetric and restricted to $-0.03 \le D \le 0.07$. Consequently, Lewontin (1964) introduced the quantity D', in which D is divided by its minimum and maximum values for the particular observed allele frequencies, so that D' can range from -1 to 1. Its sampling properties are unknown, however, so its use has declined. The use of other measures, *e.g.*, as discussed by Devlin and Risch (1995), has also declined.

The measure $r^2 = D^2/[p_A(1-p_A)p_B(1-p_B)]$, introduced by Hill and Robertson (1968), is simply the square of the conventional correlation of gene frequencies in the sample. It reduces some of the influence of allele frequency on its range: for $p_A = 0.3$, $p_B = 0.1$, for example, to $-0.22 \le r \le 0.51$; but if $p_A = p_B$, the full range from -1 to $+1$ for r (0–1 for r^2) is possible. As the Chi-square statistic with 1 d.f. for a test of correlation in a sample size of n haplotypes is equal to nr^2, it facilitates significance testing for departure from LE, albeit randomization tests are a simple alternative. Further, as discussed

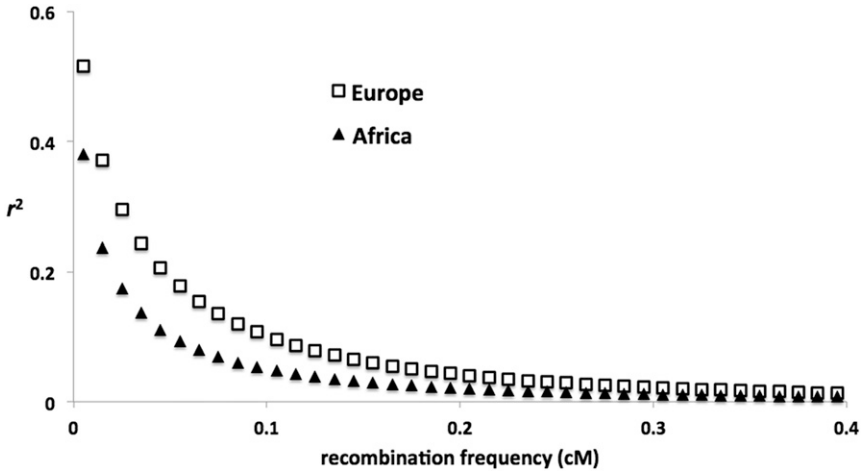

Figure 1 Summary of r^2 values from two populations averaged over all SNP pairs (International HapMap Consortium 2005; Sved *et al.* 2008).

subsequently, r^2 is relevant to the power of marker–trait association studies [genome-wide association studies (GWAS)].

An example showing the range of r^2 values in human populations is shown in Figure 1. r^2 values are dependent on allele frequencies, and SNPs with a minimum allele frequency <0.1 have been omitted for the figure. The higher LD values in European populations are expected if there was a reduction in population size (bottlenecking) during their establishment.

The measures of LD discussed to date involve only pairs of loci. The extension to more loci rapidly becomes very messy because the possible values of a three-locus quantity, *e.g.*, frequency(ABC) $- p_A p_B p_C$, have feasible boundaries dependent on both single- and two-locus haplotype frequencies. Although parametrizations and dynamics of frequency changes have been derived for multiple loci (Hill 1974a), the multi-locus disequilibria are rarely used. In random mating populations, haplotype frequencies and D can be estimated by iterative maximum likelihood for pairs (Hill 1974b) and for multiple loci (Hill 1975; Excoffier and Slatkin 1995) and explicitly for pairs (Weir and Cockerham 1979).

Other parameters have been used for different situations. Sabeti *et al.* (2002) defined the statistic extended haplotype homozygosity in measuring the decay of LD to determine sites of selection in the human genome, using multiple SNP data to define homozygous segments. Chromosome segment homozygosity, introduced by Hayes *et al.* (2003), is a similar measure, except that it uses a correction to infer identity by descent rather than homozygosity.

An important measure of LD introduced by Weir (1979) and used in population size estimation (see below) is the "composite LD measure," sometimes known as "Burrows' composite disequilibrium measure." It addresses the practical problem in diploid organisms that coupling and repulsion haplotype gametes cannot be distinguished in genotypes when both loci are heterozygous, so D cannot be estimated directly. However, a second D value can be calculated which considers not the gametes in the zygote (designated as the "coupling gametes") but rather the "repulsion gametes," the combination of the A gene from one parent and the B gene from the other parent. The sum of these two D values is the composite measure which can be calculated directly from the data. An equivalent measure to r^2 that does not assume random mating can be calculated by normalizing for gene and genotype frequencies (Weir 1979).

The Expectation of r^2 in Random Mating Populations

We now turn to the problem of predicting the expected magnitude of r^2 due to chance segregation as a function of parameters of the population, effective size, and the degree of linkage. The first approaches to this issue (Hill and Robertson 1968; Sved 1968) indicated that the expectation is a function of $1/(N_e c)$ and approximately equal to $1/(4N_e c)$ for large $N_e c$, where N_e is the effective population size.

A problem in these calculations is that r^2 is a ratio and is defined only when both loci are segregating, making it impossible to write down an exact forward recurrence relationship between generations. There is an extensive literature, in part to overcome this, notably the standardized LD quantity introduced by Ohta and Kimura (1969), $\sigma_D^2 = \mathrm{E}(D^2)/\mathrm{E}[p_A(1-p_A)p_B(1-p_B)]$, the ratio of expectations rather than $\mathrm{E}[r^2]$, the expectation of the ratio. The difference between $\sigma^2{}_D$ and $\mathrm{E}[r^2]$ is typically small.

Expectations of the components of σ_D^2, $\mathrm{E}[p_A(1-p_A)p_B(1-p_B)]$, and $\mathrm{E}(D^2)$ can be calculated by iteration of the moments over generations, requiring a third quantity, $\mathrm{E}[(1-2p_A)(1-2p_B)D]$, to obtain a closed form (Hill and Robertson 1968). Calculation can also be carried out by diffusion methods (Ohta and Kimura (1969), or by adopting a genealogical interpretation and using coalescent techniques (McVean 2002). A further complication in assessing data is that $\mathrm{E}(r^2)$ also depends on the current allele frequencies, and conditioning of the statistics on them may be needed (VanLiere and Rosenberg 2008). A full analysis has been given by Song and Song (2007).

A more general approach and analysis was undertaken by Weir and Cockerham (*e.g.*, Weir and Cockerham 1974, and

see also Weir 1979 for further review). Rather than initially setting up moments, they undertake analysis based on descent measures, probabilities that the two genes at two loci in an individual are descended from one, two, three, or four ancestral gametes, which are then identified by the individuals in which they are located. Together these provide a set of equations that enable iteration over generations, taking into account the mating system; selfing, for example, can be excluded or allowed. The methods developed by Weir and Cockerham require complicated notation, but in their hands it is a straightforward, formal, and powerful approach. The moments are functions of the allele frequencies in the base population and of the descent measures, and so can be obtained by iteration with results that are very close, except for very small populations, to those using the moments approach directly.

A simple, although less rigorous, approach to the expectation of r^2 was put forward by Sved and Feldman (1973). This was suggested by the treatment of inbreeding at a single locus, which can be defined using either the correlation between uniting gametes or the probability of identity by descent (Crow and Kimura 1970, section 3.2). For identical gametes, the correlation is one, otherwise zero, so the overall correlation is simply the probability of identity by descent. Extending the approach to two loci, the expected correlation r is equal to the probability of no recombination in a gamete. The probability of no recombination in either gamete of a pair, or the probability of linked identity-by-descent (L), estimates r^2. Its calculation is straightforward using recurrence, leading to an equilibrium at $L = 1/(1 + 4N_e c)$ for small c. The same recurrence relationship and equilibrium have been derived approximately but directly in terms of r^2 rather than L (Tenesa *et al.* 2007).

Population Subdivision and Assortative Mating

Nei and Li (1972) pointed out that LE requires dealing with a closed population. Just the act of mixing populations that are individually in LE will lead to LD in the combined population if gene frequencies are different in the subpopulations.

A related, but more complex, problem concerns expectations for LD due to drift in individual populations that are exchanging migrants. Different parameter sets and expectations for this case have been given by Ohta (1982), Tachida and Cockerham (1986), and Sved (2009).

Hedrick (2017) has also pointed to the effect of assortative mating in generating LD. In general, any departure from random mating can potentially lead to LD, although the level as measured using r^2 may be low.

The Multiple Applications of LD

So far we have looked at the description and prediction of the magnitude of LD, but not considered its uses. We consider five categories here:

1. Detecting sites of past selection in human populations.
2. Dating divergence of human and animal populations.
3. Estimation of effective population size in conservation biology.
4. GWAS.
5. Genomic prediction.

Of these categories, (4) and (5) are by far the largest areas of current interest.

Detecting sites of past selection in human populations

This method uses hitchhiking, the increase in frequency of a neutral mutation linked to an advantageous one (Smith and Haigh 1974), originally introduced without reference to LD. Sabeti *et al.* (2002) apply a similar principle, defining SNPs that are in high LD with a gene region and where the LD diminishes with increased distance from the region.

The test loses power when fixation of the newly selected gene is nearly complete, and the LD measure r^2 is undefined when complete fixation occurs. The availability of HapMap data from different populations in Africa, Asia, and Europe overcomes this difficulty, however, allowing the identification of sites of gene replacement that differ between populations. On a longer timescale, chimpanzees have been used as an outgroup to define selected regions in all human populations (Sabeti *et al.* 2007).

More than 20 chromosomal regions have been identified in this way, with many more regions showing evidence for lower levels of gene replacement. In many cases the functional gene substitutions have not been defined, but specific evidence for the substitution of genes affecting skin pigmentation, hair follicles, and resistance to Lassa virus were found (Sabeti *et al.* 2007).

Recently, Racimo *et al.* (2018) have proposed methods for inferring polygenic adaptation in complex traits by analyzing changes in genome frequency at multiple loci, and comparing the expected changes from this model with those expected from population history and simple genetic drift. These invoke the assumption that the genes analyzed are acting directly and that frequency changes do not arise through LD. Novembre and Barton (2018) recommend caution in interpreting the results.

The estimation of effective population size from LD

The expectation of r^2 is a function of effective population size N_e and recombination fraction c. Measurement of r^2 in a population from loci that are neutral for fitness should therefore lead to an estimate of population size, provided the recombination frequencies are known and the population size is constant (Hill 1981). Most other methods for estimation of population size from genetic data require measurement of gene frequencies in more than one generation.

In practice, because the methods are most useful in natural populations (often in species for which map distances are unknown) and because most pairs of loci are on different chromosomes, unlinked loci have been of most use for such measurement (Waples 2006). The expectation

given above, $E(r^2) = 1/(1 + 4N_e c)$, actually measures average N_e over the period of time during which the LD value settles down to an equilibrium, which takes much longer for closely linked loci than for loosely linked loci. Hayes *et al.* (2003) showed that the term $1/(2c)$ defines the time period relevant to population size estimation. Therefore, unlinked loci are most useful for measuring recent population size, in which case the composite r^2 measure is the method of choice. It may seem counterintuitive that unlinked loci can be in disequilibrium at all, but recombination can randomize gene combinations only in double heterozygote genotypes, which are expected to be less than half of the population.

The main difficulty with the measurement of LD for unlinked loci is that sample size tends to dominate the measured value of r^2 (Hill 1981). For unlinked loci, the expected value of the composite r^2 is $1/3N_e + 1/n$ (Weir and Hill 1980), where n is the sample size, which is likely to be small for wild populations. This difficulty can be overcome if enough highly variable markers, *e.g.*, microsatellite markers, are available. In practice, it seems that the method is sufficiently accurate only to distinguish between small and large population size (Wang 2016).

The recently developed multiple sequential Markovian coalescent (Schiffels and Durbin 2014) and pairwise sequentially Markovian coalescent methods, based on coalescence analysis of complete sequence data of a few individuals, may soon supersede the composite LD method. Currently they have been applied only on an evolutionary timescale in human populations. They require substantial genomic information and may not be applicable in many conservation studies, but have been used in a study of flycatchers by Nadachowska-Brzyska *et al.* (2016).

Dating population subdivision

A means of using LD to date the divergence between populations was proposed by de Roos *et al.* (2008) and Sved *et al.* (2008). A locus pair has a correlation equal to r_1 in one population, and a correlation equal to r_2 in a second population. The expected value of $r_1 r_2$ is then equal to $r^2(1 - c)^{2t}$, where r^2 is the square of the correlation in the ancestral population, c is the recombination frequency, and t is the number of generations since the populations separated. With knowledge of c, estimation of the value of r^2 in the ancestral population thus allows an estimate of t.

The method was used on HapMap data to estimate the number of generations since European populations diverged from African populations (Sved *et al.* 2008). The resulting estimate, ~1000 generations, is low compared to archaeological records, but is consistent with the notion of multiple waves of migration (Tassi *et al.* 2015).

GWAS

Mapping of disease genes in humans using association with SNP markers constitutes the earliest major GWAS application (see, *e.g.*, Altshuler *et al.* 2008 and Slatkin 2008) and has now expanded into a major research tool in human genetics and medicine and in the understanding of biological function (see review by Visscher *et al.* 2017 and the Web sites http://ldsc.broadinstitute.org, http://gwascentral.org, and http://www.ebi.ac.uk/gwas).

In GWAS, a test is made for each individual marker in turn (*e.g.*, by linear regression) of whether there is a significant difference in trait mean between alternative alleles at the marker. A significant difference indicates LD and that a trait gene is closely linked to that marker. As thousands of tests are undertaken, very stringent criteria for significance must be imposed to control for type-I errors, typically set at a rate of 5×10^{-8} for human data. As nearby markers are also likely to be in LD with each other, multiple hits occur, as exemplified in a Manhattan plot.

The power of an individual test depends on the effect of the trait gene and its frequency, formally on $E(r^2)$ times its additive variance plus $E(r^4)$ times its dominance variance, and on sample size (Weir 2008). As r^2 can take high values only when the marker and trait gene have near equal frequency, the power is likely to be low if the risk variant is uncommon and the marker has high heterozygosity. Indeed, the sites of largest effect are likely to have been at a selective disadvantage and are therefore rare. Eyre-Walker (2010) models some scenarios.

In such studies and indeed in all association tests, population substructure can lead to bias and false positives, so care to minimize these is needed. Many population studies record multiple health and phenotypic data on very many individuals (*e.g.*, the United Kingdom's Biobank). Summary statistics are made available to enable multiple other research groups to combine and use these data efficiently in subsequent analyses for specific projects.

GWAS have involved large and increasing resources. GWAS discoveries rose from <80 before 2008 to >10,000 by September 2016 (Visscher *et al.* 2017). Data sets can be used in GWAS for any trait on which records are included, so they are being combined. In early 2017, >30 summary association statistics of sample sizes of at least 20,000 were available (Pasaniuc and Price 2017). There has been extensive development of statistical and computational methodology to effect such advances. The successful hits in a GWAS study then provide a route for further study of gene action and understanding of the biochemistry and physiology of the loci identified, as well as the pathways through which they act.

Genomic prediction

Power and precision of identifying trait genes using GWAS can clearly be increased by fitting multiple markers, including those tightly linked to each other, and indeed the whole genome. Prediction of marker-associated effects and, from those, genotypic values (formally breeding values) for the trait on all individuals can, however, be undertaken simultaneously using whole genome marker data of all individuals included in the analysis. This approach was initially suggested by Meuwissen *et al.* (2001) in the context of selecting animals

in a dairy cattle improvement program. These predictions can be applied immediately to relatives and progeny as yet unborn based on their pedigree relationship, and predictions recomputed as data on more animals become available. Previously, young bulls were selected on their pedigree (parental records) and the most promising progeny were then tested, requiring long generation intervals and low selection intensity. Now, young bulls are selected on their genomic prediction and, consequently, rates of improvement have roughly doubled (Wiggans *et al.* 2017).

The increased accuracy of selection and opportunity for major modifications in the design and execution of breeding programs (*e.g.*, Hickey *et al.* 2017) is such that, in all livestock and increasingly in plant breeding (at least for outbreeding species), genomic prediction is becoming the norm, with clear benefit to society.

In contrast to classical GWAS, significance tests for individual genes are not required. Marker genotypes are now the independent variables in a multiple regression context, and individual animals' genetic merit, their "genomic prediction," are the dependent variables. Pedigree relationships among the animals are included in constructing the covariance matrix. To avoid overfitting, a random effects model is fitted for the vast number of marker-associated effects. The choice of its prior is an active and sometimes contentious issue, as it depends on the actual but unknown distribution of marker-associated effects. The priors that are used range from assuming the effects are all normally distributed with equal variance (now termed genomic best linear unbiased prediction or GBLUP) to Bayesian alternatives (Meuwissen *et al.* 2001).

One measure of the accuracy of genomic methods is the magnitude of the additive genetic variance accounted for by fitting just markers, the "genomic (or SNP) heritability" (Yang *et al.* 2017), compared with that from conventional analyses of quantitative traits based on pedigree. Critically, such estimates do not require a pedigree at all as this is provided by the SNPs. Early estimates differed quite substantially, creating an unproductive search for the "missing heritability" (Maher 2008). However, Yang *et al.* (2010) showed that much of this missing heritability was due to genes of small effect that could not be detected as significant in GWAS, but whose overall effect could be detected statistically. Conversely, conventional pedigree-based estimates can be biased upwards by common environment of sibs, maternal effects, and nonadditive gene action. Even so, the estimate of genomic heritability—just as for the prediction of breeding values using genomic prediction—is dependent on the statistical model fitted and on the actual distribution of gene effects in the population, which is of course unknown. De los Campos *et al.* (2015) discuss relevant concepts.

Consequences of the LD Revolution

The vast array of SNP and related markers now available, entirely unanticipated in earlier days, has led to increased recognition of the importance of LD among closely linked markers and the potential for its application to understanding the genetic basis of complex traits. As we discuss above, genomic methods using LD are now a major source of research activity and gene discovery in agriculture, human medicine, and health studies. Indeed, LD has provided a demand for research training, employment, and genome sequencing technology.

Acknowledgments

Peter Visscher and Naomi Wray provided incisive comments on an earlier draft of this article. We also acknowledge helpful advice from Ian Franklin, Mike Goddard, Bruce Weir, and an anonymous reviewer.

Literature Cited

Altshuler, D., M. J. Daly, and E. S. Lander, 2008 Genetic mapping in human disease. Science 322: 881–888. https://doi.org/10.1126/science.1156409

Bodmer, W. F., and J. Felsenstein, 1967 Linkage and selection: theoretical analysis of the deterministic two locus random mating model. Genetics 57: 237–265.

Charlesworth, B., D. Charlesworth, J. A. Coyne, and C. H. Langley, 2016 Hubby and Lewontin on protein variation in natural populations: when molecular genetics came to the rescue of population genetics. Genetics 203: 1497–1503. https://doi.org/10.1534/genetics.115.185975

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.

de los Campos, G., D. Sorensen, and D. Gianola, 2015 Genomic heritability: what is it? PLoS Genet. 11: e1005048. https://doi.org/10.1371/journal.pgen.1005048

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus Cattle. Genetics 179: 1503–1512. https://doi.org/10.1534/genetics.107.084301

Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29: 311–322. https://doi.org/10.1006/geno.1995.9003

Excoffier, L., and M. Slatkin, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12: 921–927. https://doi.org/10.1093/oxfordjournals.molbev.a040269

Eyre-Walker, A., 2010 Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. USA 107: 1752–1756. https://doi.org/10.1073/pnas.0906182107

Falconer, D. S., 1960 *Introduction to Quantitative Genetics*. Oliver and Boyd, Edinburgh.

Franklin, I. R., and R. C. Lewontin, 1970 Is the gene the unit of selection? Genetics 65: 701–734.

Harris, H., 1966 Enzyme polymorphisms in man. Proc. R. Soc. Lond. B Biol. Sci. 164: 298–310. https://doi.org/10.1098/rspb.1966.0032

Hästbacka, J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. 2: 204–211 (erratum: Nat. Genet. 2: 343). https://doi.org/10.1038/ng1192-204

Hayes, B. J., P. M. Visscher, H. C. McPartlan, and M. E. Goddard, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13: 635–643. https://doi.org/10.1101/gr.387103

Hedrick, P. W., 2017 Assortative mating and linkage disequilibrium. G3 (Bethesda) 7: 55–62. https://doi.org/10.1534/g3.116.034967
Hickey, J. M., T. Chiurugwi, I. Mackay, and W. Powell, 2017 Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat. Genet. 49: 1297–1303. https://doi.org/10.1038/ng.3920
Hill, W. G., 1974a Disequilibrium among several linked neutral genes in finite populations. 1. Mean changes in disequilibria. Theor. Popul. Biol. 5: 366–392. https://doi.org/10.1016/0040-5809(74)90059-8
Hill, W. G., 1974b Estimation of linkage disequilibrium in randomly mating populations. Heredity 33: 229–239. https://doi.org/10.1038/hdy.1974.89
Hill, W. G., 1975 Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 31: 881–888. https://doi.org/10.2307/2529813
Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209–216. https://doi.org/10.1017/S0016672300020553
Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226–231. https://doi.org/10.1007/BF01245622
International HapMap Consortium, 2005 A haplotype map of the human genome. Nature 437: 1299–1320. https://doi.org/10.1038/nature04226
Karlin, S., and M. W. Feldman, 1970 Linkage and selection: two locus symmetric viability model. Theor. Popul. Biol. 1: 39–71. https://doi.org/10.1016/0040-5809(70)90041-9
Kimura, M., 1965 Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection. Genetics 52: 875–890.
Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304: 412–417. https://doi.org/10.1038/304412a0
Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49: 49–67.
Lewontin, R. C., and J. L. Hubby, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54: 595–609.
Lewontin, R. C., and K. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution 14: 458–472.
Maher, B., 2008 Personal genomes: the case of the missing heritability. Nature 456: 18–21. https://doi.org/10.1038/456018a
McVean, G. A. T., 2002 A genealogical interpretation of linkage disequilibrium. Genetics 162: 187–191.
Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.
Morton, N. E., W. Zhang, P. Taillon-Miller, S. Ennis, P.-Y. Kwok *et al.*, 2001 The optimal measure of allelic association. Proc. Natl. Acad. Sci. USA 98: 5217–5221. https://doi.org/10.1073/pnas.091062198
Nadachowska-Brzyska, K., R. Burri, L. Smeds, and H. Ellegren, 2016 PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. Mol. Ecol. 25: 1058–1072. https://doi.org/10.1111/mec.13540
Nei, M., and W. Li, 1972 Linkage disequilibrium in subdivided populations. Genetics 75: 213–219.
Novembre, J., and N. H. Barton, 2018 Tread lightly interpreting polygenetic tests of selection. Genetics 208: 1351–1355. https://doi.org/10.1534/genetics.118.300786
Ohta, T., 1982 Linkage disequilibrium with the island model. Genetics 101: 139–155.
Ohta, T., and M. Kimura, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. 13: 47–55. https://doi.org/10.1017/S001667230000272X
Pasaniuc, B., and A. L. Price, 2017 Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. 18: 117–127. https://doi.org/10.1038/nrg.2016.142
Racimo, F., J. J. Berg, and J. K. Pickrell, 2018 Detecting polygenic adaptation in admixture graphs. Genetics 208: 1565–1584. https://doi.org/10.1534/genetics.117.300489
Robbins, R. B., 1918 Some applications of mathematics to breeding problems. III. Genetics 3: 375–389.
Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837. https://doi.org/10.1038/nature01140
Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918. https://doi.org/10.1038/nature06250
Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46: 919–925. https://doi.org/10.1038/ng.3015
Slatkin, M., 2008 Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9: 477–485. https://doi.org/10.1038/nrg2361
Smith, J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35. https://doi.org/10.1017/S0016672300014634
Song, Y. S., and J. S. Song, 2007 Analytic computation of the expectation of the linkage disequilibrium coefficient r2. Theor. Popul. Biol. 71: 49–60. https://doi.org/10.1016/j.tpb.2006.09.001
Sved, J. A., 1968 The stability of linked systems of loci with a small population size. Genetics 59: 543–563.
Sved, J. A., 2009 Correlation measures for linkage disequilibrium within and between populations. Genet. Res. 91: 183–192. https://doi.org/10.1017/S0016672309000159
Sved, J. A., and M. W. Feldman, 1973 Correlation and probability methods for one and two loci. Theor. Popul. Biol. 4: 129–132. https://doi.org/10.1016/0040-5809(73)90008-7
Sved, J. A., A. F. McRae, and P. M. Visscher, 2008 Divergence between human populations estimated from linkage disequilibrium. Am. J. Hum. Genet. 83: 737–743. https://doi.org/10.1016/j.ajhg.2008.10.019
Tachida, H., and C. C. Cockerham, 1986 Analysis of linkage disequilibrium in an island model. Theor. Popul. Biol. 29: 161–197. https://doi.org/10.1016/0040-5809(86)90008-0
Tassi, L., S. Ghirotto, M. Mezzavilla, S. T. Vilaca, L. De Santi *et al.*, 2015 Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. Investig. Genet. 6: 13. https://doi.org/10.1186/s13323-015-0030-2
Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.
VanLiere, J. M., and N. A. Rosenberg, 2008 Mathematical properties of the r2 measure of linkage disequilibrium. Theor. Popul. Biol. 74: 130–137. https://doi.org/10.1016/j.tpb.2008.05.006
Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy *et al.*, 2017 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. 101: 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005
Wang, J., 2016 A comparison of single-sample estimators of effective population sizes from genetic marker data. Mol. Ecol. 25: 4692–4711. https://doi.org/10.1111/mec.13725
Waples, R. S., 2006 A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv. Genet. 7: 167–184. https://doi.org/10.1007/s10592-005-9100-y
Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics 35: 235–254. https://doi.org/10.2307/2529947

Weir, B. S., 2008 Linkage disequilibrium and association mapping. Annu. Rev. Genomics Hum. Genet. 9: 129–142. https://doi.org/10.1146/annurev.genom.9.081307.164347

Weir, B. S., and C. C. Cockerham, 1974 Behavior of pairs of loci in finite monoecious populations. Theor. Popul. Biol. 6: 323–354. https://doi.org/10.1016/0040-5809(74)90015-X

Weir, B. S., and C. C. Cockerham, 1979 Estimation of linkage disequilibrium in randomly mating populations. Heredity 42: 105–111. https://doi.org/10.1038/hdy.1979.10

Weir, B. S., and J. Goudet, 2017 A unified characterization of population structure and relatedness. Genetics 206: 2085–2103. https://doi.org/10.1534/genetics.116.198424

Weir, B. S., and W. G. Hill, 1980 Effect of mating structure on variation in linkage disequilibrium. Genetics 95: 477–488.

Wiggans, G. R., J. B. Cole, S. M. Hubbard, and T. S. Sonstegard, 2017 Genomic selection in dairy cattle: the USDA experience. Annu. Rev. Anim. Biosci. 5: 309–327. https://doi.org/10.1146/annurev-animal-021815-111422

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42: 565–569. https://doi.org/10.1038/ng.608

Yang, J., J. Zeng, M. E. Goddard, N. R. Wray, and P. M. Visscher, 2017 Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. 49: 1304–1310. https://doi.org/10.1038/ng.3941

Communicating editor: A. S. Wilkins