Jakob Mueller
is affiliated to the National Centre for Genetic Epidemiological Methods. His research interests include population and evolutionary genetics in animals and humans as well as gene mapping for complex traits through linkage and association studies.

Linkage disequilibrium for different scales and applications

Jakob C. Mueller

Received (in revised form): 27th October 2004

Abstract

Assessing the patterns of linkage disequilibrium (LD) has become an important issue in both evolutionary biology and medical genetics since the rapid accumulation of densely spaced DNA sequence variation data in several organisms. LD deals with the correlation of genetic variation at two or more loci or sites in the genome within a given population. There are a variety of LD measures which range from traditional pairwise LD measures such as D' or r^2 to entropy-based multi-locus measures or haplotype-specific approaches. Understanding the evolutionary forces (in particular recombination) that generate the observed variation of LD patterns across genomic regions is addressed by model-based LD analysis. Marker type and its allelic composition also influence the observed LD pattern, microsatellites having a greater power to detect LD in population isolates than SNPs. This review aims to explain basic LD measures and their application properties.

Keywords: *linkage disequilibrium measures, recombination rate variation, positive selection, microsatellites, SNPs*

Jakob C. Mueller,
Institute for Human Genetics,
GSF – National Research Institute for Environment and Health,
85764 Neuherberg,
Germany

Tel.: +49 89 3187 3464
Fax.: +49 89 3187 3297
E-mail: jakob.mueller@gsf.de

INTRODUCTION

Testing for the presence of linkage disequilibrium (LD) and measuring its value are two important instruments of statistical genetics that have recently received a great deal of attention. Novel methods, which enable high-throughput genotyping of closely localised genetic markers on a given chromosome, certainly contributed to the renewed interest of linkage disequilibrium. In the past few decades LD has been utilised as a tool for genetic mapping of trait or disease loci in humans and model organisms.

LD is defined as the non-random gametic association of alleles at different loci in a population. Synonymous terms are 'allelic association' or 'gametic phase disequilibrium'. It should be noted that LD measures the allelic association in the same gamete, although there are relaxations to that. If the allelic association is additionally measured on the same chromosome – which is mostly the case – LD is considered a measure of chromosomal proximity or linkage of genetic loci. But there are also other factors that enhance the level of LD. Co-selection of two or more non-linked loci or recent admixture of populations with differing gametic frequencies can influence LD in the same way as true linkage.[1]

There are a number of reasons why it may be of interest to assess whether the allele distribution at several loci is at linkage disequilibrium or not. In general, population genetic models exhibit much simpler behaviour when there is no LD, ie linkage equilibrium. Each genetic locus can be independently modelled. On the other hand, when there is LD variation within a region, this information can be used to estimate the regional variation of the recombination rate.[2] In the field of population genetics, LD has been extensively used to describe demographic and evolutionary processes in plant and animal populations. For example, admixture or migration among populations was assessed by LD patterns.[3] There is also a causal relationship of LD with population size, natural selection and mutation.

LD plays a central role in mapping genes relevant for specific traits of interest

Design and interpretation of association studies depend on LD patterns

mainly in humans and useful animals and plants. In this approach, genetic variation at a set of marker loci in a sample of individuals is tested for association with a given phenotype.[4] If such an association is found between a particular marker locus and the phenotype, it suggests that either the variation at that marker locus affects the phenotype of interest, or that the variation of that marker locus is in LD with the true phenotype-related locus, which was not genotyped. The association signal and the pattern of LD around this signal pinpoint the chromosomal region within which the causal variant(s) for the phenotype should be searched. General LD patterns are therefore important for the design and interpretation of association studies.

There are a variety of LD measures available. The following sections group these measures in terms of their method of calculation and applicability. Web addresses for the estimation and testing of LD are presented in Table 1.

PAIRWISE LD MEASURES

Central to most LD calculations stands the linkage disequilibrium coefficient *D*, for which the layout and notation are shown in Figure 1. Consider two loci A and B, each locus having two possible alleles: A1 and A2 at locus A and B1 and B2 at locus B. The allele frequencies are denoted as *p* and naturally represent only sample estimates of some underlying population parameters, which are mostly unknown unless the total population have been scored. There are four possible allele combinations among these two loci, which could represent the four possible types of gametes in a sexually reproducing organism. If the two loci are physically linked on the same chromosome, this array specifically represents the four haplotypes, but this does not have to be the case. If the two loci assort completely independently (ie linkage equilibrium), the gametic frequencies are calculated by the products of the allele frequencies, eg the frequency of a gamete bearing the

Table 1: Web addresses for the estimation and testing of LD

Pairwise LD	
GOLD package (ldmax, haploxt)	http://csg.sph.umich.edu/pn/index.php?furl=/abecasis/GOLD/index.html
Haploview	http://www.broad.mit.edu/personal/jcbarret/haplo/index.php
GENEPOP/LinkDos	ftp://ftp.cefe.cnrs-mop.fr/genepop/
ARLEQUIN	http://lgb.unige.ch/arlequin/
GENETIX	http://www.univ-montp2.fr/%7Egenetix/genetix/genetix.htm
GDA	http://lewis.eeb.uconn.edu/lewishome/software.html
POPGENE	http://www.ualberta.ca/~fyeh/index.htm
DnaSP	http://www.ub.es/dnasp/
The R Project for Statistical Computing	http://www.r-project.org/
InfoGeneMap	http://www.lnt.ei.tum.de/download/InfoGeneMap/
Multi-locus LD	
MultiLocus	http://www.agapow.net/software/multilocus/
lds (ε calculator)	http://www.bioinf.mdc-berlin.de/~capella/eld/eld.htm
HapGraph	http://bioinformatics.med.utah.edu/~alun/software.html
Haplotype-specific LD	
EHH calculator	http://ihg.gsf.de/cgi-bin/mueller/webehh.pl
Model-based LD and recombination	
LDhat	http://www.stats.ox.ac.uk/~mcvean/LDhat/LDhat.html
LDMAP	http://cedar.genetics.soton.ac.uk/public_html/helpld.html
HOTSPOTTER and PHASE	http://www.stat.washington.edu/stephens/software.html
infs and sequenceLD	http://www.maths.lancs.ac.uk/~fearnhea/software/

Alleles at locus A		Alleles at locus B: B1	B2	
A1		A1B1	A1B2	
	Actual	p_{A1B1}	p_{A1B2}	
	Expected	$p_{A1}p_{B1}$	$p_{A1}p_{B2}$	p_{A1}
A2		A2B1	A2B2	
	Actual	p_{A2B1}	p_{A2B2}	
	Expected	$p_{A2}p_{B1}$	$p_{A2}p_{B2}$	p_{A2}
		p_{B1}	p_{B2}	1

Figure 1: Association between two alleles at each of two loci, showing the actual gametic frequencies and the expected gametic frequencies when the loci are in linkage equilibrium. The marginal frequencies represent the allele frequencies

The basic LD measure *D* needs to be standardised for comparison

alleles A1 and B1 is given by the product $p_{A1} \times p_{B1}$. A simple and basic component of many disequilibrium measures is the difference (D) between the actual gametic frequency and the expected gametic frequency when the loci are independent. With $p_{A?B?}$ being the actual gametic frequencies, there are four different (for each gamete) expressions, which all calculate the same D value:

$$\begin{aligned} D &= p_{A1B1} - p_{A1}\,p_{B1} \\ &= p_{A2B2} - p_{A2}\,p_{B2} \\ &= -(p_{A1B2} - p_{A1}\,p_{B2}) \\ &= -(p_{A2B1} - p_{A2}\,p_{B1}) \end{aligned}$$

In a biallelic system, the deviation between actual and expected gametic frequencies in the coupling phase must be equal but opposite in sign to those in the repulsion phase; hence, for the last two expressions the change of sign.

The additive linkage disequilibrium coefficient D, however, is constrained in the value it may take by the underlying allele frequencies of the two loci, since actual gametic frequencies cannot be negative. For instance, since $p_{A1B2} = p_{A1}\,p_{B2} - D \geqslant 0$, it follows that $D \leqslant p_{A1}\,p_{B2}$, and so on. In order to compare LD quantities among different pairs of loci with differing allele frequencies, several standardisation methods have been proposed. For a comparison of the properties of such standardised coefficients see Hedrick,[5] Lewontin,[6] Devlin and Risch[7] or Morton *et al.*[8] One way of standardisation is provided by dividing the coefficient D by its maximum value given the allele frequencies:[9]

$$D' = D/D_{\max}$$

with $D_{\max} = \min(p_{A1}\,p_{B2},\ p_{A2}\,p_{B1})$ if $D > 0$ and $D_{\max} = \max(-p_{A1}\,p_{B1}, -p_{A2}\,p_{B2})$ if $D < 0$. This procedure always makes D'-values range between 0 and 1.

It can be shown that D measures the statistical association of alleles in forming gametes, and is related to the well-known Pearson correlation coefficient r for a 2×2 table:[10]

$$r = D/(p_{A1}\,p_{A2}\,p_{B1}\,p_{B2})^{1/2}$$

The squared coefficient of determination r^2 is often used to remove the arbitrary sign introduced, when the marker alleles are arbitrarily labelled.

Significance testing for the LD coefficient D follows testing for independence in a 2×2 contingency table as shown in Figure 1. The usual methods for this type of test can be employed: a chi-square test, a likelihood ratio test or Fisher's exact test.[11,12] If the sample sizes or, more specifically, the expected frequencies in the cells of the contingency table are small, the asymptotic properties of the $\chi^2$and likelihood ratio test statistics are unlikely to apply. The significance of observed values of any statistics can alternatively be obtained by permuting the alleles of one of the loci with respect to the other locus alleles, keeping the allele frequencies constant. In this case, the p-value for the statistic is the proportion of permutations, which result in equal to or more extreme values of the statistic. Properties of multiple allele LD measures are explored by Weir and Cockerham[13] and Weir.[11] Tests are simply a generalisation for the

goodness-of-fit tests on a 2×2 table to more than two rows and columns ($r \times c$ tables).

When the gametic phase is unknown, genotypic LD measures can be used

Complications for the calculation of gametic LD coefficients arise when only genotype data in diploid individuals are scored, but the haplotype phases across loci remain unknown. This is generally the case when unrelated individuals are analysed by standard genotyping devices. Family recruitment or experimental methods may help to estimate the phase in diploid data, but these methods are currently time-consuming and expensive. For example, if the experimenter determines that the genotype of an individual at locus A is A1/A2 and at locus B is B1/B2, it will be unknown whether the individual's genotype is made up from an A1B1 haplotype and an A2B2 haplotype or instead is made up from an A1B2 haplotype and an A2B1 haplotype. That is, gametic phase of individuals that are heterozygous at two or more loci cannot be directly specified. However, under the assumption of random mating, in which genotypic frequencies are assumed to be the products of gametic frequencies, it is possible to obtain maximum likelihood estimates of gametic frequencies in an iterative procedure called expectation-maximisation (EM) algorithm.[14–16] With the estimated gametic frequencies, we can proceed with the standard LD calculations described above.

The assumption of random mating, however, might not be valid for a specific population. In such a case, one could use alternative Bayesian methods for haplotype reconstruction that are relatively robust to deviations of the random mating assumption[17] or use composite genotypic disequilibrium measures.[11] Composite LD coefficients do not distinguish between the two possible gametic phases in double heterozygotes, but rather jointly consider their deviations from random association. The composite measure Δ is defined as

$$\Delta = p_{\mathrm{A1B1}} + p_{\mathrm{A1/B1}} - 2\,p_{\mathrm{A1}}\,p_{\mathrm{B1}}$$

with $p_{\mathrm{A1/B1}}$ being the non-gametic frequency, ie the frequency that allele A1 of locus A and allele B1 of locus B within an individual come from different gametes. For an extension of the composite LD measure to the multi-allelic case see Schaid.[18] Alternative methods simply test the random association of single-locus genotypes across two or more loci, and do not intend to estimate gametic phases at all. They hypothesise that a two-locus genotypic frequency is equal to the product of corresponding one-locus genotypic frequencies.[19] Likewise, entropy-based measures such as the 'mutual information' can be used as basic measures of the information dependency between loci.[20] These measures using genotypic data are relatively free of assumptions, which might be advantageous in population samples of unknown substructure where mating is not guaranteed to be random.

D′ and r^2 may be applied for different purposes

The most frequently used LD coefficients D' and r^2 have very different properties and may be applied for different purposes. D' and its confidence bounds is useful to assess the probability for historical recombination in a given population, whereas r^2 is useful in the context of association studies. The parameter D' reaches its maximum value of 1 if one or more of the four gametes in Figure 1 is not observed. When in the absence of recurrent and/or backward mutation (infinite-site model, which is supposed to be appropriate for single nucleotide polymorphisms, SNPs) only three gametes are detected, there is no need to infer historical recombination. Each of the three gametes can be deduced from any starting haplotype just by single mutations. But when all four gametes between a pair of loci are observed, D' will be less than 1 and the only explanation is the occurrence of at least one historical recombination event. A deviation of D' from 1 thus gives evidence for historical recombination. However, D' values are known to fluctuate upwards when a small number of samples or rare alleles are examined. It

is therefore suggested that confidence intervals of D' should be relied on rather than point estimates.[21] Several methods have been proposed for estimating the confidence interval of D'.[21–24]

The LD coefficient r^2 is arguably the most relevant measure for association mapping, because there is a simple relationship between r^2 and the sample size required to detect association between a trait and marker loci. Suppose an LD of r^2 was measured between a causal locus and a nearby marker locus. Then, to achieve the same power to detect association at the marker locus as we would have at the causal locus, we need to increase our sample size by a factor of $1/r^2$.[25]

MULTI-LOCUS LD MEASURES

Multi-locus LD measures assess the background levels of LD

Extending test statistics from the two-locus case to more than two loci is relatively straightforward and formulae for tests are outlined.[11,12,19] However, to extend the estimation of the standard two-locus LD coefficients to the multi-locus case is difficult, and only specific coefficients, such as the allele-specific coefficients of gametic disequilibria, are described.[11] These specific measures are not very useful to describe the general LD structure across a chromosomal region with several markers. The combined analysis of all pairwise LD measures across a region is also not able to detect simultaneous allele associations among multiple loci. As an illustration, assume a sequence of four markers, each with two alleles labelled 1 and 2. Only four haplotypes (1122, 1221, 2112, 2211) each with equal frequency of 25 per cent are observed out of the possible 16 haplotypes. These four haplotypes comprise a block of LD since only a quarter of all possible haplotypes occurs; however, no measure based on D will detect LD between pairs of adjacent markers.

To measure the background levels of LD, various coefficients based on haplotypes have been proposed. They all rely on the same rationale as two-locus measures by calculating the difference in value between the observed state and the expected one under linkage equilibrium. Normalisation is generally achieved by dividing by the expected value. The unit to measure the states is usually some sort of diversity measure. One common approach employs the variance of pairwise distances between haplotypes, ie the number of loci at which they differ.[12,26–28] This measure is defined as the index of association I_A:

$$I_A = (V_o - V_e)/V_e$$

where V_o is the observed variance of pairwise distances and V_e is the variance expected under linkage equilibrium. Thus, it essentially tests to what extent haplotypes that are the same at one locus are more likely than random to be the same at other loci. A modification of I_A that removes the dependency on the number of loci was proposed by Agapow and Burt.[29] In diploid organisms, the index I_A is used to measure the variance differences between haplotypes within individuals.[30] Significance tests can be based either on randomisation tests[31] or on analytical approximations of the variance.[32]

Homozygosity of haplotypes, ie the probability of selecting two identical haplotypes at random from the population, can also be used to measure the level of LD among two or more loci. The basic coefficient H for three markers, for instance, is

$$H = H_{ABC} - H_A H_B H_C$$

with H_A, H_B and H_C being the homozygosities of single markers A, B and C respectively, and H_{ABC} being the haplotype homozygosity. The product of single marker homozygosities refers to the expected state under linkage equilibrium. Various standardisation methods and properties of H including test procedures have been thoroughly discussed by Sabatti and Risch.[33] An excess of either homozygosity or heterozygosity (the complement of homozygosity) signals a

departure from the gametic phase equilibrium.

A third measure of multi-locus LD employs entropy as a measure for the information content or non-structure of a haplotype array.[34] The normalised entropy difference ε is defined as:

$$\varepsilon = (S_E - S_B)/S_E$$

where S_E is the expected entropy in the equilibrium case and S_B is the observed entropy. It has been shown by the authors that ε can be interpreted as a multi-locus extension of the LD coefficient r^2, and a procedure for testing its significance is proposed.

A different approach to show the joint distribution of alleles at associated loci is based on graphical model estimation.[35] This method allows for non-contiguous and overlapping LD groups, which is important when analysing dense genetic data in which not only recombination but also mutation and population history shape the association structure among loci. An additional appeal of this approach is that categorical phenotypes can be included in the same analysis.

Signatures of selection are indicated by unusual haplotype-specific LD patterns

HAPLOTYPE-SPECIFIC LD

Multi-locus LD measures describe the general LD for an array of several haplotypes within a chromosomal region. However, each haplotype may have its own evolutionary history, and one may be interested in the LD structure of a specific haplotype, because this haplotype showed, for instance, a strong association with a phenotype of interest. The LD pattern and relative length of this haplotype can be assessed by a method that uses extended haplotype homozygosities.[36,37] Haplotype homozygosity is known from the previous section as a multi-locus measure of LD.[33] The procedure starts by calculating the homozygosity of a population of haplotypes that comprise the single specified core haplotype of interest. This is done in a stepwise manner by including more and more markers on both sides of the specified core region, increasing the length of haplotypes for which the homozygosity is calculated. In other words, the extended haplotype homozygosity (EHH) estimates the level of haplotype splitting owing to recombination and mutation on both sides of a specified core region. An attractive aspect of this approach is that the various core haplotypes at a locus serve as internal controls for one another at the same chromosomal region. This is important given the variability of local recombination rates across regions.

In combination with the core haplotype frequency, the extended haplotype homozygosity may serve as an indicator of recent positive selection or severe population bottlenecks. Frequent core haplotypes with an unusually high long-range LD are supposed to be positively selected in large populations. This is explained as follows: new non-selected variants (represented by a specified core haplotype) require a long time to reach high frequency in the large population, and LD around the variants will decay substantially during this period owing mainly to recombination. By comparison, positive selection or demographic bottlenecks can cause an unusually rapid rise in haplotype frequency, occurring over a short enough time that recombination is inefficient and results in long-range LD.[38] Signatures of selection could corroborate association signals, because it is suggested that natural selection in ancestral populations played a role at loci influencing susceptibility to common complex diseases.

MODEL-BASED LD MEASURES

In evolutionary studies, recombination plays a key role, because it destroys the simple cladistic behaviour of genealogical relationships among extant haplotypes of the analysed genomic region. Phylogenetic reconstructions, therefore, tend to subdivide genomic regions such that recombination can be ignored within the subparts. Characterising the rate and position of recombination is also

important for the design of association studies in terms of recombination-free blocks in which genetic markers have to be analysed and interpreted jointly. For example, it would help to define representative 'tag' markers for other loci which are transferable among populations. There is, however, no straightforward relationship between the above-described LD measures and the recombination rate. Methods of LD analysis based on explicit evolutionary models, instead, provide powerful tools for estimating population recombination rates.

In models based on the coalescent theory, ie models that statistically describe the genealogical history of a sample of chromosomes,[39] the key parameter in determining the extent of linkage disequilibrium is the product of the recombination rate r and the effective population size N_e, often termed the population recombination rate $N_e r$. Without prior information about one of these two parameters, it is impossible to estimate them separately. One method estimates the approximate likelihood of observing the LD patterns in the data under a range of population recombination rates within the framework of coalescent theory.[40,41] This method has been extended to allow for recurrent mutation.[42] The likelihoods are combined across pairs to provide a point estimate of the population recombination rate $N_e r$ and significance is tested by permutation methods.

Recombination rate variation can be estimated by model-based LD measures

Another approach employs the Malecot model that incorporates the main evolutionary processes of recombination, genetic drift, migration and mutation. An extension of this probabilistic model describes the decline of allelic association with increasing physical distance d between genetic markers.[43] This method is used to fit the Malecot parameter ε and construct LD maps with a map location in εd LD units.[8,44]

In addition, statistical models have been developed that relate patterns of LD directly to the underlying recombination process. The method considers each sampled haplotype in turn and attempts to construct it as a mosaic of previously considered haplotypes.[45] The average length of mosaic pieces is used to estimate the local background recombination rate, and the positions of breaks in the mosaic to estimate the location and intensity of recombination hot spots.

All these methods apply to the estimation of recombination variation at kilobase scales in large surveys of densely genotyped genomic regions. First surveys revealed evidence for substantial fine-scale variation in recombination rates across the human genome corroborating sparse experimental results.[2,46]

MARKER TYPE AND LD PATTERNS

LD patterns observed in natural populations are the result of a complex interplay between biological factors, such as recombination and mutation, and the population's demographic and evolutionary history. The structure and the effective size of populations as well as the selective regime (co-selection of loci, selective sweeps) are important determinants for regional LD patterns. It is therefore not surprising that substantial variation in LD among genomic regions and populations analysed have been reported.[47] Variation may even be traced back to individual differences in haplotype lengths. A wealth of algorithms based on the described LD measures has been developed to define high LD regions.[48]

From a practical point of view, type and informativeness of analysed markers also influence LD patterns. Single nucleotide polymorphisms (SNPs) and microsatellites (STR) are the most commonly used markers because of their abundance. There are nearly 10 million human SNP sites (dbSNP[49]) and ~12,000 microsatellites[50] in the public domain. There are not many studies that directly compare these two marker types. A general difference is that microsatellites have multiple alleles and higher heterozygosities (informativeness) as well

as a higher mutation rate than SNPs. Recurrent mutation in microsatellites can explain the lower levels of LD for tightly linked markers, and the more recent origin of microsatellite alleles can explain the slower observed decay rate of LD with physical distance.[50] Intervals across which LD was detected using microsatellite markers were significantly wider than those detected using SNPs.[51] Microsatellite pairs up to ~3 Mb apart were found to exhibit significant LD in a Finish sub-isolate, whereas SNPs revealed LD over only ~0.5 Mb. This observation would suggest the use of microsatellites in isolated populations for large-scale gene mapping approaches.[52] SNPs were inferior to microsatellites, even when the information from 3 to 5 SNPs was combined.[51] The increased long-range LD in microsatellites may partly be attributed to the higher information content,[25] but also to differing biological characteristics of these markers. However, further investigations are needed to reveal the distinct LD properties of SNPs and microsatellites.

Microsatellites showed long-range LD in a population isolate

Different LD patterns are also expected for alleles with different frequencies. In a neutral model of evolution, common alleles are generally older than rare alleles and there has been more historical opportunity for recombination to break down ancestral haplotypes. LD patterns calculated on markers that bear only high-frequency alleles will therefore emphasise the older history in comparison to low-frequency markers.

CONCLUSION

LD analysis has a wide range of applications. Population geneticists utilise LD analyses to assess the population structure and population history. In particular, model-based and haplotype-specific LD measures describe the variation patterns of recombination, mutation and natural selection across the genome and thus enhance our understanding of genome evolution. In an ideal case, the evolutionary history (demographic and selection history) of a given gene can be reconstructed. The understanding of such evolutionary processes may improve phylogenetic reconstructions based on those genes.

Another recent focus of LD analysis is in mapping complex trait loci. The international HapMap project[53] generates genome-wide and densely spaced sequence variation data in several human populations from Asia, Africa and Europe. Several local projects on limited genomic regions are also under way. This type of data will certainly promote multi-locus LD measures in order to assess the variability of background correlation across genomic regions. In regions with high LD, a low number of representative markers (tag-markers) will sufficiently capture the information of sequence variation in that region.[48] Genotyping effort for genome-wide association studies will thus be substantially reduced.

Acknowledgments

I am grateful to Jack Favor for helpful discussions. The author is supported by the German BMBF funded project 'Bioinformatics for the functional analysis of mammalian genomes – BFAM'.

References

1. Hartl, D. L. and Clark, A. G. (1989), 'Principles of Population Genetics', Sinauer Associates, Inc., Sunderland, MA.
2. McVean, G. A. T., Myers, S. R., Hunt, S. *et al.* (2004), 'The fine-scale structure of recombination rate variation in the human genome', *Science*, Vol. 304, pp. 581–584.
3. Mueller, J. (1998), 'Genetic population structure of two cryptic *Gammarus fossarum* types across a contact zone', *J. Evol. Biol.*, Vol. 11, pp. 79–101.
4. Lewis, C. M. (2002), 'Genetic association studies: design, analysis and interpretation', *Brief. Bioinformatics*, Vol. 3, pp. 146–153.
5. Hedrick, P. W. (1987), 'Gametic disequilibrium measures: Proceed with caution', *Genetics*, Vol. 117, pp. 331–341.
6. Lewontin, R. C. (1988), 'On measures of gametic disequilibrium', *Genetics*, Vol. 120, pp. 849–852.
7. Devlin, B. and Risch, N. (1995), 'A comparison of linkage disequilibrium measures for fine-scale mapping', *Genomics*, Vol. 29, pp. 311–322.
8. Morton, N. E., Zhang, W., Taillon-Miller, P.

et al. (2001), 'The optimal measure of allelic association', *Proc. Natl Acad. Sci. USA*, Vol. 98, pp. 5217–5221.

9. Lewontin, R. C. (1964), 'The interaction of selection and linkage. I. General considerations; heterotic models', *Genetics*, Vol. 49, pp. 49–67.
10. Hill, W. G. and Robertson, A. (1968), 'Linkage disequilibrium in finite populations', *Theor. Appl. Genet.*, Vol. 38, pp. 226–231.
11. Weir, B. S. (1996), 'Genetic Data Analysis II', Sinauer Associates, Inc., Sunderland, MA.
12. Hudson, R. R. (2001), 'Linkage Disequilibrium and Recombination', in Balding, D. J., Bishop, M. and Cannings, C., Eds, 'Handbook of Statistical Genetics', John Wiley and Sons, Ltd., Chichester, pp. 309–324.
13. Weir, B. S. and Cockerham, C. C. (1978), 'Testing hypotheses about linkage disequilibrium with multiple alleles', *Genetics*, Vol. 88, pp. 633–642.
14. Hill, W. G. (1974), 'Estimation of linkage disequilibrium in randomly mating populations', *Heredity*, Vol. 33, pp. 229–239.
15. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Stat. Soc. B*, Vol. 39, pp. 1–38.
16. Excoffier, L. and Slatkin, M. (1995), 'Maximum likelihood estimation of molecular haplotype frequencies in a diploid population', *Mol. Biol. Evol.*, Vol. 12, pp. 921–927.
17. Stephens, M., Smith, N. J. and Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *Amer. J. Hum. Genet.*, Vol. 68, pp. 978–989.
18. Schaid, D. J. (2004), 'Linkage disequilibrium testing when linkage phase is unknown', *Genetics*, Vol. 166, pp. 505–512.
19. Zaykin, D., Zhivotovsky, L. and Weir, B. S. (1995), 'Exact tests for association between alleles at arbitrary numbers of loci', *Genetica*, Vol. 96, pp. 169–178.
20. Mueller, J. C., Bresch, E., Dawy, Z. *et al.* (2003), 'Shannon's mutual information applied to population-based gene mapping', *Amer. J. Hum. Genet.*, Vol. 73, Suppl., p. 610.
21. Gabriel, S. B., Schaffner, S. F., Nguyen, H. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science*, Vol. 21, pp. 2225–2229.
22. Zapata, C., Alvarez, G. and Carollo, C. (1997), 'Approximate variance of the standardized measure of gametic disequilibrium D', *Amer. J. Hum. Genet.*, Vol. 61, pp. 771–774.
23. Ayres, K. L. and Balding, D. J. (2001), 'Measuring gametic disequilibrium from multilocus data', *Genetics*, Vol. 157, pp. 413–423.
24. Kim, S. K., Zhang, K. and Sun, F. (2004), 'A comparison of different strategies for computing confidence intervals of the linkage disequilibrium measure D', *Pacific Symp. Biocomput.*, 2004, pp. 128–139.
25. Pritchard, J. K. and Przeworski, M. (2001), 'Linkage disequilibrium in humans: Models and data', *Amer. J. Hum. Genet.*, Vol. 69, pp. 1–14.
26. Sved, J. A. (1968), 'The stability of linked systems of loci with small population size', *Genetics*, Vol. 59, pp. 543–563.
27. Brown, A. D. H., Feldman, M. W. and Nevo, E. (1980), 'Multilocus structure of natural populations of *Hordeum spontaneum*', *Genetics*, Vol. 96, pp. 523–536.
28. Maynard Smith, J., Smith, N. H., O'Rourke, M. and Spratt, B. G. (1993), 'How clonal are bacteria?', *Proc. Natl Acad. Sci. USA*, Vol. 90, pp. 4384–4388.
29. Agapow, P.-M. and Burt, A. (2001), 'Indices of multilocus linkage disequlibrium', *Mol. Ecol. Notes*, Vol. 1, pp. 101–102.
30. Chakraborty, R. (1984), 'Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample', *Genetics*, Vol. 108, pp. 719–731.
31. Burt, A., Carter, D. A., Koenig, G. L. *et al.* (1996), 'Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*', *Proc. Natl Acad. Sci. USA*, Vol. 93, pp. 770–773.
32. Haubold, B., Travisano, M., Rainey, P. B. *et al.* (1998), 'Detecting linkage disequilibrium in bacterial populations', *Genetics*, Vol. 150, pp. 1341–1348.
33. Sabatti, C. and Risch, N. (2002), 'Homozygosity and linkage disequilibrium', *Genetics*, Vol. 160, pp. 1707–1719.
34. Nothnagel, M., Fürst, R. and Rohde, K. (2002), 'Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks', *Human Heredity*, Vol. 54, pp. 186–198.
35. Thomas, A. and Camp, N. J. (2004), 'Graphical modeling of the joint distribution of alleles at associated loci', *Amer. J. Hum. Genet.*, Vol. 74, pp. 1088–1101.
36. Sabeti, P. C., Reich, D. E., Higgins, J. M. *et al.* (2002), 'Detecting recent positive selection in the human genome from haplotype structure', *Nature*, Vol. 419, pp. 832–837.
37. Mueller, J. C. and Andreoli, C. (2004), 'Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype

homozygosity', *Bioinformatics*, Vol. 20, pp. 786–787.

38. Bersaglieri, T., Sabeti, P. C., Patterson, N. *et al.* (2004), 'Genetic signatures of strong recent positive selection at the *Lactase* gene', *Amer. J. Hum. Genet.*, Vol. 74, pp. 1111–1120.
39. Hudson, R. R. (1983), 'Testing the constant rate neutral allele model with protein sequence data', *Evolution*, Vol. 37, pp. 203–217.
40. Hudson, R. R. (2001), 'Two-locus sampling distributions and their application', *Genetics*, Vol. 159, pp. 1805–1817.
41. Fearnhead, P. N. and Donnelly, P. (2002), 'Approximate likelihood methods for estimating local recombination rates', *J. R. Stat. Soc. Ser. B*, Vol. 64, pp. 657–680.
42. McVean, G., Awadalla, P. and Fearnhead, P. (2002), 'A coalescent-based method for detecting and estimating recombination from gene sequences', *Genetics*, Vol. 160, pp. 1231–1241.
43. Malecot, G. (1973), 'Isolation by Distance', in Morton, N. E., Ed., 'Genetic Structure of Populations', Univ. Press of Hawaii, Honolulu, pp. 72–75.
44. Maniatis, N., Collins, A., Xu, C.-F. *et al.* (2002), 'The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis', *Proc. Natl Acad. Sci. USA*, Vol. 99, pp. 2228–2233.
45. Li, N. and Stephens, M. (2003), 'Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data', *Genetics*, Vol. 165, pp. 2213–2233.
46. Crawford, D. C., Bhangale, T., Li, N. *et al.* (2004), 'Evidence for substantial fine-scale variation in recombination rates across the human genome', *Nat. Genet.*, Vol. 36, pp. 700–706.
47. Wall, J. D. and Pritchard, J. K. (2003), 'Haplotype blocks and linkage disequilibrium in the human genome', *Nature Reviews, Genetics*, Vol. 4, pp. 587–597.
48. Cardon, L. R. and Abecasis, G. R. (2003), 'Using haplotype blocks to map human complex trait loci', *Trends Genetics*, Vol. 19, pp. 135–140.
49. URL: http://www.ncbi.nlm.nih.gov/SNP/index.html
50. Abecasis, G. R., Noguchi, E., Heinzmann, A. *et al.* (2001), 'Extent and distribution of linkage disequilibrium in three genomic regions', *Amer. J. Hum. Genet.*, Vol. 68, pp. 191–197.
51. Varilo, T., Paunio, T., Parker, A. *et al.* (2003), 'The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories', *Human Mol. Genet.*, Vol. 12, pp. 51–59.
52. Varilo, T. and Peltonen, L. (2004), 'Isolates and their potential use in complex gene mapping efforts', *Curr. Opinion Genet. Develop.*, Vol. 14, pp. 316–323.
53. The International HapMap Consortium (2003), 'The international HapMap project', *Nature*, Vol. 426, pp. 789–796.