Linkage Disequilibrium

Peter JP Croucher, *University of California at Berkeley, Berkeley, California, USA*

Advanced article

Article Contents

Online posting date: 15th April 2013

When two or more polymorphic loci are studied in a population, the interaction between the loci is often expressed in terms of linkage disequilibrium (LD). The loci are in LD if their respective alleles do not associate independently (randomly). LD does not necessarily imply physical linkage, however most often the loci considered are on the same chromosome and the degree of over- or underrepresentation of an expected haplotype measures the extent of LD between a specific pair of alleles. Multilocus patterns of LD are often visualised graphically, revealing local blocks of high LD. LD is generated by mutation but may also be generated and maintained by population processes including selection, drift and admixture. Genomic rearrangements, such as inversions may also influence LD patterns. Recombination overtime acts to reduce LD and this relationship may be used to date alleles.

Measures of Linkage Disequilibrium (LD)

The genetics of two or more loci are considered in terms of haplotype (or gamete) frequencies. Consider a pair of segregating loci, one with alleles A and a at frequencies p_A and p_a and one with alleles B and b at frequencies p_B and p_b. If the alleles associate independently (there is no LD) then the frequencies of the four possible haplotypes or gametes, AB, Ab, aB and ab, are given by the products of the allele frequencies p_Ap_B, p_Ap_b, p_ap_B and p_ap_b. In the presence of LD some of these haplotypes will be more frequent than expected and some will be more rare; this difference from expectation is measured with the LD coefficient D (Lewontin and Kojima, 1960) as tabulated below in **Table 1**. **See also**: Polymorphisms: Origins and Maintenance; Population Genetics: Multilocus

Consequently, $D = p_{AB}p_{ab} - p_{Ab}p_{aB}$. This is often expressed more simply as $D = p_{Ab} - p_Ap_b$. D can take values between -0.25 and $+0.25$ and it is common practise to work with D^2 to eliminate problems with sign. Normalised coefficients of LD are often employed, most commonly D' and r^2. D' (Lewontin, 1964) takes values between -1 and $+1$ and is less dependent on allele frequencies than is D. D' is given by dividing D by its maximum possible numerical value for the given allele frequencies:

$$D' = D/\min(p_ap_B, p_Ap_b) \text{ if } D>0 \text{ or } D' = D/\min(p_Ap_B, p_ap_b) \text{ if } D<0 \quad [1]$$

D' has the useful property that if $D'=1$ this indicates that one (or more) of the four possible haplotypes is absent. This is equivalent to the 'four-gamete test' (Hudson and Kaplan, 1985) which suggests that under an infinite sites model there can be at most four haplotypes or gametes between two sites (loci). Assuming there is no recurrent or back mutation then the only way for all four haplotypes to be present (or $D'<1$) is if at least one recombination event has occurred. Unfortunately, $D'=1$ is likely to occur by chance in small samples and values of $D'<1$ have no clear interpretation in terms of the extent of recombination. Therefore D' is often considered along with other normalised coefficients such as r^2, δ and the LD-LOD (the $\log_{10}$ of the odds of LD between two loci (the probability of the data given LD/the probability of the data given no LD), also known as the Z-score and analogous to the measure originally defined for linkage analysis, see Morton, 1955). **See also**: Linkage Analysis

The square of the allelic correlation coefficient, r^2 (Hill and Robertson, 1968) is commonly used and takes values from 0 to 1. It is calculated by dividing D^2 by the product of all four allele frequencies:

$$r^2 = (p_{AB}p_{ab} - p_{Ab}p_{aB})^2/(p_Ap_ap_Bp_b) \quad [2]$$

Table 1 The LD coefficient D is the difference between the observed and the expected haplotype frequencies

Haplotype	Frequency
AB	$p_{AB} = p_Ap_B + D$
Ab	$p_{Ab} = p_Ap_b - D$
aB	$p_{aB} = p_ap_B - D$
ab	$p_{ab} = p_ap_b + D$

eLS subject area: Evolution & Diversity of Life

How to cite:
Croucher, Peter JP (April 2013) Linkage Disequilibrium. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005427.pub3

Numerous additional measures of LD have been proposed, especially within the context of mapping loci associated with disease (see, e.g. Devlin and Risch, 1995; Morton *et al.*, 2001). The conditional probability δ_A that a haplotype carries an allele A, given that it carries allele B at another locus, is sometimes used in disease association mapping (Bengtsson and Thomson, 1981):

$$\delta_A = p_A + D/p_B \quad [3]$$

It is important to note that all of these measures depend upon D, which is defined for a specific pair of alleles. When only two alleles exist at each of a pair of loci, all haplotype specific values of $|D|$ are equivalent. When more than two alleles are present then D must be considered with respect to particular pairs of alleles. **See also**: Single Nucleotide Polymorphism (SNP)

Sources of LD

Populations are not finite. Random drift generates LD because not all haplotypes are sampled proportionately from generation to generation. Consequently, population history (e.g. bottlenecks and founder effects) can generate marked differences in the extent of LD between populations. Long-distance LD in Caucasian and Asian populations is elevated relative to many African populations presumably as a result of a bottleneck when these populations left Africa (Schmegner *et al.*, 2005). **See also**: Genetic Drift in Human Populations; Genetic Load; Genetic Variation: Polymorphisms and Mutations; Population Genetics: Historical Aspects; Population History and Linkage Disequilibrium

Frequently, LD exists between loci simply because an insufficient number of generations have passed to allow recombination to randomise the haplotypes in the population. When a new mutation arises on a chromosome it will initially be in complete LD with one of the alleles of any neighbouring polymorphism (the four-gamete test, above). **See also**: Recombination and Human Genetic Diversity

LD can be generated by selection. If a particular combination of alleles at two loci is more advantageous than carrying one or other of the alleles alone then that haplotypic combination will increase in frequency and hence increase LD (Felsenstein, 1965). In addition, epistatic effects, where the fitness of an allele at one locus depends on that at another locus, can lead to the preferential selection of certain haplotypes and therefore the maintenance of LD. The human leucocyte antigen (HLA) genes may be one example maintained by balancing selection. **See also**: Disease Associations: Human Leukocyte Antigen (HLA) and Apolipoprotein E (APOE) Gene; Epistasis; Genetic Variation: Polymorphisms and Mutations; Histocompatibility Antigen Complex of Man; Identifying Regions of the Human Genome that Exhibit Evidence for Positive Selection; Major Histocompatibility Complex (MHC); Major Histocompatibility Complex: Disease Associations; Population Genetics: Multilocus; Population History and Linkage Disequilibrium

LD also results when differentiated populations (i.e. the allele frequencies at loci differ) merge (admixture), when individuals with different genotypes mate nonrandomly (inbreeding) or where chromosomal inversions occur. **See also**: Chromosome Rearrangement Patterns in Mammalian Evolution; Gene Flow, Haplotype Patterns and Modern Human Origins; Population History and Linkage Disequilibrium

Multilocus Models and Interaction

The HLA system represents a multilocus system of highly variable, tightly linked genes involved in immune defence and self-recognition. Many haplotypes are over-represented and this high level of LD may be the result of selection for heterozygote advantage (e.g. Hraber *et al.*, 2007). The HLA is a classical, albeit complex, example of a coadapted gene complex. Much current attention focuses on measures of LD for multiple-locus (and to a smaller extent multiple-allele) systems. Many two-locus measures can be extended to multilocus measures, e.g. D and D' (Ayres and Balding, 2001; Weir, 1996). However, most approaches do not account for the higher-order LD present in multiple-locus systems and little practical use has been made of higher-order LD coefficients (Slatkin, 2008). Some novel measures such as the haplotype-based normalised entropy difference (Nothnagel *et al.*, 2002), together with 'top-down' mathematical decomposition (Gorelick and Laubichler, 2004), attempt to incorporate these difficulties.

Rather than utilise multilocus LD coefficients, pairwise LD coefficients among multiple loci are more commonly visualised graphically using software such as Haploview (Barrett *et al.*, 2005; see **Figure 1**) or GOLD (Abecasis and Cookson, 2000). The development of a fine-scale map of human LD by the International HapMap Consortium (2007) confirmed that the human genome consists of blocks or regions of high LD which are delineated by areas of higher recombination ('hotspots'). Tools such as Haploview were specifically designed to enable the visualisation of these blocks (see **Figure 1**). The block structure suggested that genotyping only 'haplotype-tagging' variants within these blocks could be used as a means to reduce genotyping costs in disease association studies; however, the need for this approach is now dimished with ultra-high-throughput genotyping technologies. **See also**: Blocks of Limited Haplotype Diversity; Disease Associations: Human Leukocyte Antigen (HLA) and Apolipoprotein E (APOE) Gene; Epistasis; Genome-Wide Association Studies; HapMap Project; Histocompatibility Antigen Complex of Man; Major Histocompatibility Complex: Disease Associations; Microarrays and Single Nucleotide Polymorphism (SNP) Genotyping; Multilocus Linkage Analysis; Population Genetics: Multilocus; Recombination and Human Genetic Diversity

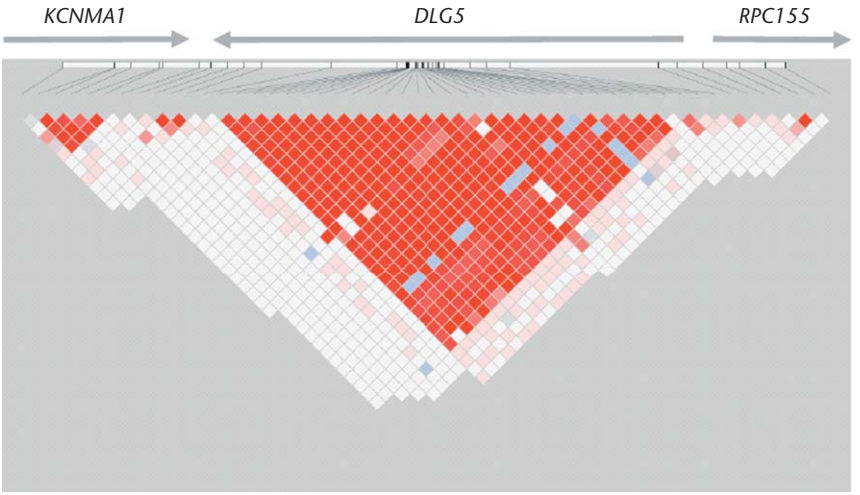

Figure 1 Blocks of LD across the human *DLG5* gene and its flanking region. Graphical pairwise LD among single nucleotide polymorphisms (SNPs) as output by the software Haploview (Barrett *et al.*, 2005). SNPs in the gene *DLG5* occupy block of strong LD ($D' > 0.8$) as defined by the dark red squares. White square indicates weak LD and the blue squares indicate high D' values but low LOD scores. Reproduced with permission from Stoll *et al.*, 2004. © Nature Publishing Group.

Although LD is typically considered among physically linked markers, loci that are separated by large distances or are on different chromosomes may also exhibit 'LD' or allelic association, for instance, because of higher-level interactions or epistasis. Consequently, the term gametic phase disequilibrium is still sometimes used to specify that physically linked markers are being examined. **See also**: Epistasis; Population Genetics: Multilocus

Decay as a Function of Age

Given an infinite population and no selection, recombination will act over successive generations to reduce the amount of LD between two physically linked markers. LD decays exponentially at a rate that depends on the linkage distance or recombination fraction, r, such that from one generation to the next $D_{n+1} = (1-r)D_n$. If the loci are unlinked ($r = 0.5$) then D will halve each generation, but if r is small, as it may be for closely linked markers, then substantial levels of LD can remain for hundreds of generations. The persistence of LD between closely spaced markers is frequently exploited in attempts to map disease-causing mutations. Projects, such as the International HapMap Project (The International HapMap Consortium, 2007), attempt to maximise this potential and have revealed that local patterns of LD and recombination are typically block like (see 'Multilocus Models and Interaction'). The dependency of LD on the recombination fraction can be used to date mutations based upon their allele frequencies (Slatkin and Rannala, 2000; see also Keinan and Clark, 2012). **See also**: Blocks of Limited Haplotype Diversity; Human Variation Databases; Linkage Analysis; Linkage and Association Studies; Population Genetics: Multilocus; Population History and Linkage Disequilibrium; Recombination and Human Genetic Diversity; Susceptibility Genes: Detection

References

Abecasis GR and Cookson WO (2000) GOLD – graphical overview of linkage disequilibrium. *Bioinformatics* **16**(2): 182–183.

Ayres KL and Balding DJ (2001) Measuring gametic disequilibrium from multilocus data. *Genetics* **157**: 413–423.

Barrett JC, Fry B, Maller J and Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**(2): 263–265. Epub 5 Aug 2004.

Bengtsson BO and Thomson G (1981) Measuring the strength of association between HLA antigens and diseases. *Tissue Antigens* **18**: 356–363.

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

Felsenstein J (1965) The effect of linkage on directional selection. *Genetics* **52**: 349–363.

Gorelick R and Laubichler MD (2004) Decomposing multilocus linkage disequilibrium. *Genetics* **166**: 1581–1583.

Hill WG and Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**: 226–231.

Hraber P, Kuiken C and Yusim K (2007) Evidence for human leukocyte antigen heterozygote advantage against hepatitis C virus infection. *Hepatology* **46**(6): 1713–1721.

Hudson RR and Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

Keinan A and Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**: 740–743.

Lewontin RC (1964) The interaction of selection and linkage I. General considerations; heterotic models. *Genetics* **49**: 49–67.

Lewontin RC and Kojima K (1960) The evolutionary dynamics of complex polymorphisms. *Evolution* **14**: 458–472.

Morton NE (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7**: 277–318.

Morton NE, Zhang W, Taillon-Miller P *et al*. (2001) The optimal measure of allelic association. *Proceedings of the National Academy of Sciences of the USA* **98**: 5217–5221.

Nothnagel M, Fürst R and Rhode K (2002) Entropy as a measure of linkage disequilibrium over multilocus haplotype blocks. *Human Heredity* **54**: 186–198.

Schmegner C, Hoegel J, Vogel W and Assum G (2005) Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the European population. *Human Genetics* **118**: 276–286.

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

Slatkin M and Rannala B (2000) Estimating allele age. *Annual Review of Genomics and Human Genetics* **1**: 225–249.

Stoll M, Corneliussen B, Costello CM *et al*. (2004) Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nature Genetics* **36**(5): 476–480.

Weir BS (1996) *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates.

Further Reading

Clark AG, Wang X and Matise T (2010) Contrasting methods of quantifying fine structure of human recombination. *Annual Review of Genomics and Human Genetics* **11**: 45–64.

Daly M, Rioux JD, Schaffner SF, Hudson TJ and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* **29**: 229–232.

Goldstein DB (2001) Islands of linkage disequilibrium. *Nature Genetics* **29**: 109–111.

Jorde JB (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Research* **10**: 1435–1444.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**: 139–144.

Lewontin RC (1988) On measures of gametic disequilibrium. *Genetics* **120**: 849–852.

Maynard Smith J (1989) *Evolutionary Genetics*. Oxford, UK: Oxford University Press.

Nordborg M and Tavaré S (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**: 83–90.

Reich DE, Cargill M, Bolk S *et al*. (2001) Linkage disequilibrium in the human genome. *Nature* **441**: 199–204.