Black Sea Journal of Agriculture
doi: 10.47115/bsagriculture.1571251

Open Access Journal
e-ISSN: 2618 – 6578

Review
Volume 8 - Issue 1: 103-107 / January 2025

LINKAGE DISEQUILIBRIUM IN ANIMAL GENETICS – DEFINITION, MEASURES AND APPLICATIONS

Godswill Arinzechukwu IWUCHUKWU[1]*, Marvellous OYEBANJO[2], Uğur ŞEN[1]

[1]Ondokuz Mayıs University, Faculty of Agriculture, Department of Agricultural Biotechnology, 55105, Samsun, Türkiye
[2]University of Ibadan, Department of Animal Science, Animal Breeding and Genetics Unit, Ibadan, Nigeria

Abstract: The non-random connection of alleles at various loci is known as linkage disequilibrium (LD). Combinations of alleles inside haplotypes occur at frequencies that differ from those expected on independence when two alleles at two distinct loci are in LD. When genetic variation at a locus is linked to a trait, it means that either the genetic variation at that locus directly impacts the phenotype of interest or the locus is in LD with the causal mutation. The level of LD, which dictates how many markers should be typed in a genome scan to discover a quantitative trait locus (QTL) using LD, is critical to the practicality of association studies. This review explores the origin of LD in genetics and how it applies to animal breeding and genetics.

Keywords: Linkage disequilibrium, Quantitative trait loci, Alleles, Haplotypes, Genetic variation

***Corresponding author:** Ondokuz Mayıs University, Faculty of Agriculture, Department of Agricultural Biotechnology, 55105, Samsun, Türkiye
E mail: godswill2014.gi@gmail.com (G. A. IWUCHUKWU)
Godswill Arinzechukwu IWUCHUKWU https://orcid.org/0009-0001-3621-7055
Marvellous OYEBANJO https://orcid.org/0000-0002-0175-7916
Uğur ŞEN https://orcid.org/0000-0001-6058-1140

Received: October 21, 2024
Accepted: December 06, 2024
Published: January 15, 2025

Cite as: Iwuchukwu GA, Oyebanjo M, Şen U. 2025. Linkage disequilibrium in animal genetics – definition, measures and applications. BSJ Agri, 8(1): 103-107.

1. Introduction

Consider 2 hypothetical markers, A and B that are on the same chromosomes. Alleles A1 and A2 are present in A, and alleles B1 and B2 are present in B. A1 B1, A1 B2, A2 B1, and A2 B2 are the four potential haplotypes of markers. If the population's frequencies of alleles A1, A2, B1, and B2 are all 0.5, we can anticipate the population's frequencies of the four haplotypes to be 0.25. Linkage disequilibrium (LD) occurs when the haplotype frequencies deviate from 0.25, indicating that the genes are not in random association. Two loci unlinked are possible to be in linkage disequilibrium in some populations (Kavuncu, 2021) - in fact, linkage disequilibrium between a marker and a QTL is essential if the QTL is to be found in either type of analysis (Mueller, 2004). The distinction is that linkage analysis only takes into account linkage disequilibrium within families, which can span tens of thousands of cM and is broken down by recombination after only a few generations. A marker must be in linkage disequilibrium (LD) with a QTL throughout the entire population for linkage disequilibrium mapping. The relationship must have persisted for a significant number of generations to be a property of the entire population; hence, the marker(s) and QTL must be closely related.

2. Measures of Linkage Disequilibrium

According to Hill (1981), one measure of LD is D, which can be calculated as (Equation 1):

$$D = Freq\ (A1B1) \times Freq\ (A2B2)\ - Freq\ (A1B2) \times Freq\ (A2B1) \tag{1}$$

where *Freq (A1alt indis*$_{B1}$*)* is the population frequency of the A1$_{B1}$ haplotype, and similarly for the other haplotypes.

The *D* statistic is highly reliant on the frequencies of individual alleles, making it ineffective for assessing the degree of LD between numerous loci (for example, at different points along the genome). Hill and Robertson (1968) suggested the *r²* statistic (Equation 2) is less dependent on the allele frequencies metric.

$$r^2 = \frac{D^2}{Freq\ (A1) \times Freq\ (A2) \times Freq\ (B1) \times Freq\ (B2)} \tag{2}$$

The frequency of the A1 allele in the population is *Freq (A1)*, and the same is true for the other alleles in the population. The value of *r²* ranges from 0 for a pair of loci with no linkage disequilibrium to 1 for a pair of loci in complete linkage disequilibrium.

For example, consider the following hypothetical allelic frequencies.

$Freq\ (A1) = Freq\ (A2) = Freq\ (B1) = \ Freq(B2) = 0.5$

The haplotype frequencies are:

$Freq\ (A1_{B1}) = 0.1$
$Freq\ (A1_{B2}) = 0.4$
$Freq\ (A2_{B1}) = 0.4$
$Freq\ (A2_{B2}) = 0.1$

$D\ = 0.1 \times 0.1 - 0.4 \times 0.4 = \ -0.15$

$D^2 = 0.0225$

The value of *r*2 is then

$$\frac{0.0225}{(0.5 \times 0.5 \times 0.5 \times 0.5)} = 0.36$$

This is a moderate level of *r*2.

D' is another often used pair-wise LD measure. The value of D is standardized by the highest value it can achieve to determine D' (Equation 3).

$$D' = \frac{|D|}{D_{max}} \quad (3)$$

where if D > 0, (Equation 4)

$$D_{max} = \min[Freq(A1)\{1 - Freq(B2)\}, \{1 - Freq(A2)\}Freq(B1)] \quad (4)$$

If D < 0 (Equation 5)

$$D_{max} = \min[Freq(A1) \times Freq(B2), \{1 - Freq(A2)\}\{1 - Freq(B2)] \quad (5)$$

For two reasons, the statistic *r*2 is recommended over D' as a measure of the amount of LD. Firstly, the *r2* between a marker and a (unobserved) QTL is the fraction of variation generated by alleles at a QTL that can be explained by markers. The decrease in *r*2 with distance represents how many markers or phenotypes are needed to discover QTL in an initial genome scan using LD. When compared to the sample size for testing the QTL itself, the sample size for detecting an ungenotyped QTL must be raised by a factor of 1/*r*2. D', on the other hand, performs a terrible job of forecasting needed marker density for a genome scan using LD. The second rationale for using *r*2 instead of D' to determine the level of LD is that D' is prone to be overstated when sample sizes are small or allele frequencies are low (McRae *et al*. 2002).

The LD measurements mentioned above are for bi-allelic markers. While they can be applied to multi-allelic markers like microsatellites, Zhao *et al*. (2005) suggested using the $\chi^{2'}$ (Equation 6) measure of LD for multi-allelic markers.

$$\chi^{2'} = \frac{1}{l-1}\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{D_{ij}^2}{Freq(A_i)Freq(B_j)} \quad (6)$$

$$D_{ij}^2 = Freq(A_iB_j) - Freq(A_i)Freq(B_j)$$

Freq (Ai) is the frequency of the ith allele at marker A, *Freq (Bj)* is the frequency of the jth allele at marker B, and *l* is the minimum of the number of alleles at marker A and marker B. Note that for bi-allelic markers, $\chi^{2'} = r^2$.

Zhao et al. (2005) study involved the use of simulation, which indicated a number of multi-allelic pair-wise measures of LD – and $\chi^{2'}$ was the most reliable predictor of useable marker-QTL LD; that is, the measure of QTL variance that can be explained by the marker. We may want to quantify the extent of LD across a chromosome region that contains several markers, yet statistics like *r*2 only consider two loci at a time. The chromosome segment homozygozity (CSH) is an alternative multi-locus definition of LD (Hayes et al., 2003). Consider an ancestral animal that lived many generations ago and has descendants now. The ancestor's chromosome is torn down with each generation until only little portions of chromosome that may be traced back to the common ancestor remain. By descent, these chromosomal regions are identical (otherwise called identical by descent, IBD). The likelihood that two chromosomal segments of the same size and location picked at random from the population originate from a common ancestor (i.e., IBD) without intervening recombination is the CSH. CSH refers to the length of a chromosomal segment, up to the entire chromosome length. The CSH cannot be determined directly from marker data but must be inferred from marker haplotypes for chromosomal segments.

Consider a chromosomal segment with marker locus A on the left end and marker locus B on the opposite end. Alleles A and B define the haplotype. Two of these segments are randomly selected from the population. The haplotype homozygosity (HH) is the likelihood that two haplotypes are identical by state (IBS). The two haplotypes can be IBS in one of two ways: one, either they descended from a common ancestor without intervening recombination and are thus identical by descent (IBD); or two, they are identical by state but not IBD. CSH is the likelihood of one. Given that the segment is not IBD, the likelihood of two is a function of the marker homozygotes. The haplotype homozygosity (HH) is calculated by adding the probabilities of one and two (Equation 7).

$$HH = CSH + \frac{(Hom_A - CSH)(Hom_B - CSH)}{1 - CSH} \quad (7)$$

where *HomA* and *HomB* are the homozygosities of marker A and marker B, respectively. When the haplotype homozygosities and individual marker homozygosities are observed from the data, this equation can be solved for CSH. The estimated haplotype homozygosity can be determined in a similar but more difficult manner for more than two markers.

Another advantage of employing multi-locus LD measurements over pair-wise measures is that they can be less variable. Two sampling mechanisms cause the variation in LD. The initial sampling process is based on finite population size and reflects the sampling of gametes to generate successive generations. The second sampling procedure is the selection of individuals from the population to be genotyped, which is determined by the sample size, n. The large variability of LD measurements is due to the first sampling step. Marker pairs located at different locations in the genome but separated by a comparable distance might have vastly varied r2 values, especially if the marker separation is small. This is because an ancestral recombination between one set of markers but not the other may have occurred by accident.

Because they aggregate information across numerous loci in a time interval, multi-locus estimates of LD can minimize variability by averaging some of the impacts of accidental recombinations. Hayes et al. (2003) used simulation to evaluate the variability of r^2 and CSH. They used a mutation-drift model with a constant N of 1000 to generate a chromosomal region of 10 cM containing 11 markers. They discovered that when at least four loci were included in the CS computation, CSH was less variable than r^2.

2.1. Origins of Linkage Disequilibrium in Livestock Populations

Migration, mutation, selection, a tiny finite population size, or other genetic processes can cause LD in a population. In an F2 QTL mapping experiment, LD is established between marker and QTL alleles by crossing two inbred lines; in an F2 QTL mapping experiment, LD is created between marker and QTL alleles by crossing two inbred lines.

The fundamental source of LD in livestock populations is widely thought to be finite population size. This is due to the fact that

i. most livestock populations have tiny effective population sizes, resulting in huge quantities of LD;
ii. LD caused by crossbreeding (migration) is substantial when crossing inbred lines but minimal when crossing breeds with similar gene frequencies, and it fades within a few generations (Goddard, 1991);
iii. mutations are likely to have occurred many generations ago; and
iv. while selection is most likely a major driver of LD, its impact is likely to be limited to specific genes, with little impact on the amount of LD 'averaged' across the genome.

2.2. LD Extent in Livestock and Human Populations

If LD is primarily caused by finite population size, it should be less severe in humans than in cattle, because the effective population size in people is around 10,000 (Kruglyak, 1999), whereas in livestock, effective population numbers might be as low as 100 (Riquet et al., 1999). The image is a little muddied by the fact that animal numbers have been substantially bigger, although the effective population size of Caucasians has been much smaller (following the out of Africa hypothesis). As a result, we should anticipate seeing that the r^2 values in livestock are significantly higher than in humans at long distances between markers, but the amount of LD is more equivalent at short distances. This is exactly what has been observed. Moderate LD ($r^2 \geq 0.2$ in humans, for example) often spans less than 5 kb (0.005 cM) depending on the group investigated (Dunning et al., 2000; Reich et al., 2001; Tenesa et al., 2007). In humans and cattle, however, very high levels of LD (e.g., $r^2 \geq 0.8$) only reach a short distance. The first whole-genome LD study in cattle, which used 284 microsatellite markers from 581 maternally inherited gametes in Dutch black and white dairy cows to quantify the extent and distribution of LD, was carried out, with high levels of LD extending over several tens of centimorgans (Farnir et al., 2000). LD in cattle has been confirmed in several following studies (Tenesa et al., 2003; Vallejo et al., 2003; Khatkar et al., 2006a; Odani et al., 2006). Only recently, a study in a large mildly selected cattle population from Western Africa conducted under an extensive breeding system revealed that LD extends over shorter distances than previous studies from developed countries, which was explained by increased selective pressure and/or an admixture process (Thévenon et al., 2007). All of these LD investigations used microsatellite loci that were highly informative but had a low locus density. With the conclusion of the bovine genome sequencing project, it is now possible to determine the extent of LD using dense single nucleotide polymorphism (SNP) marker maps, resulting in significantly higher resolution. SNP markers have minimal genotyping costs, in addition to their abundance in the genome (Snelling et al., 2005). (Hinds et al., 2005). Khatkar et al. (2006b) used SNP loci to generate a first- generation LD map of bovine chromosome 6 in Australian Holstein–Friesian cattle, and D′ to estimate the extent of LD. The distance over which LD is expected to be beneficial for association mapping was discovered to be 13.3 Mb, indicating that the range of LD in Holstein–Friesian dairy cow is broad. McKay et al. (2007) used 2670 SNP markers to build LD maps for eight cow breeds from the *Bos taurus* and *Bos indicus* subspecies, and found that the amount of LD (calculated using r^2) available for association analysis does not surpass 500 kb. The disparities in the degree of LD between McKay et al. (2007) and prior investigations were related to differences in LD reporting measures, notably D′ vs. r^2. Previous investigations have found that D′ overestimates the extent of LD (Ardlie et al., 2002; Ke et al., 2004), resulting in extensive LD at long intermarker distances (Farnir et al., 2000; Tenesa et al., 2003; Vallejo et al., 2003; Khatkar et al., 2006a; Odani et al., 2006). Du et al. (2007) used 4500 SNP markers genotyped in six lines of commercial pigs to determine the degree of LD in pigs. Because paternal haplotypes were over-represented in the population, only maternal haplotypes of commercial pigs were utilized to calculate r^2 between SNPs. According to the findings of their investigation, pigs may have significantly greater LD than cattle. The average value of r^2 for SNPs separated by 1 cM was roughly 0.2. In cattle, LD of this size barely extends 100 kb. The average r^2 in pigs at 100 kb was 0.371. Heifetz et al. (2005) investigated the degree of LD in several breeding chicken populations. They employed microsatellite markers and applied the statistics to determine the degree of LD. They discovered considerable LD over large distances in their populations. For example, 57% of marker pairs separated by 5-10 cM had an $\chi^{2`} \geq 0.2$ in one line of chickens and 28% in the other. Heifetz et al. (2005) pointed out that the lines they studied had small effective population sizes and were largely inbred, so the level of LD in other chicken populations with greater effective population sizes may differ significantly. The extent of LD in domestic sheep was studied by McRae et al. (2002). Because they employed the D' parameter rather than the r^2 parameter, it's impossible to compare their findings to those of other species. They discovered that high levels of LD lasted for tens of centimorgans and then dropped as marker distance increased. They also looked at D' bias under various conditions and discovered that D' can be skewed when uncommon alleles are present. To establish the true extent of LD, they suggested using the statistical significance of LD in conjunction with coefficients such as D'.

3. Conclusion

QTL mapping can now be done using linkage disequilibrium. The population level connections between markers and QTL are used in linkage disequilibrium (LD) mapping of QTL. Because there are little pieces of chromosome in the current population that are descended from the same common ancestor, these relationships occur. These chromosome segments with no intervening recombination will have identical marker alleles or haplotypes, and if there is a QTL inside the chromosome segment, they will have identical QTL alleles. The genome-wide association test with single marker regression is the simplest of the QTL mapping procedures that take advantage of LD. Due to the availability of tens of thousands of single nucleotide polymorphism (SNP) markers in cattle, pigs, chickens, and sheep soon, doing trials to map QTL in genome-wide scans using LD has recently become practical.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	G.A.I.	M.O.	U.Ş.
C	10	80	10
D	20	80	
S			100
DCP	50	50	
DAI		100	
L	20	80	
W	20	60	20
CR	30	40	30
SR	30	40	30
PM	40	30	30

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management.

Conflict of Interest

The authors declare that there is no conflict of interest.

References

Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet, 3(4): 299-309.

Du FX, Clutter AC, Lohuis MM. 2007. Characterizing linkage disequilibrium in pig populations. Int J Biol Sci, 3: 166-178.

Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu CF, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BAJ. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. Amer j Human Genet, 67: 1544-1554.

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M. 2000. Extensive genome-wide linkage disequilibrium in cattle. Genome Res, 10(2): 220-227.

Goddard ME. 1991. Mapping genes for quantitative traits using linkage disequilibrium. Genet Select Evol, 23: 131s-134s.

Hayes BJ, Visscher PM., McPartlan H, Goddard M E. 2003. A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size. Genome Res, 13: 635.

Heifetz EM, Fulton JE, O'Sullivan N, Zhao H, Dekkers JC, Soller M. 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. Genetics, 171: 1173-1181.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. Theor Appl Genet, 38: 226-231.

Hill WG. 1981. Estimation of effective population size from data on linkage disequilibrium. Genet Res, 38: 209-216.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole- genome patterns of common DNA variation in three human populations. Science, 307(5712): 1072-1079.

Kavuncu O. 2021. Populasyon genetiği ve kantitatif genetic. Nobel Yayınları, Ankara, Türkiye, pp: 222.

Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, Deloukas P. 2004. The impact of SNP density on fine-scale patterns of linkage disequilibrium. Hum Mol Genet, 13(6): 577-588.

Khatkar MS, Collins A, Cavanagh JA, Hawken RJ, Hobbs M, Zenger KR, Barris W, McClintock AE, Thomson PC, Nicholas FW, Raadsma HW. 2006a. A first-generation metric linkage disequilibrium map of bovine chromosome 6. Genetics, 174(1): 79-85.

Khatkar MS, Thomson PC, Tammen I, Cavanagh JA, Nicholas FW, Raadsma HW 2006b. Linkage disequilibrium on chromosome 6 in Australian Holstein-Friesian cattle. Genet Select Evol, 38(5): 463-477.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genet, 22: 139-144.

McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Dias Neto E, Gill CA, Gao C, Mannen H, Stothard P, Wang Z, Van Tassell CP, Williams JL, Taylor JF, Moore SS. 2007. Whole genome linkage disequilibrium maps in cattle. BMC Genet, 8: 74.

McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J. 2002. Linkage disequilibrium in domestic sheep. Genetics, 160: 1113-1122.

Mueller JC. 2004. Linkage disequilibrium for different scales and applications. Brief Bioinform, 5(4): 355-364.

Odani M, Narita A, Watanabe T, Yokouchi K, Sugimoto Y, Fujita

T, Oguni T, Matsumoto M, Sasaki Y. 2006. Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. Anim Genet, 37(2): 139-144.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjlan R, Farhadian SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature, 411: 199-204.

Riquet J, Coppieters W, Cambisano N, Arranz JJ, Berzi P, Davis SK, Grisart B, Farnir F, Karim L, Mni M, Simon P, Taylor JF, Vanmanshoven P, Wagenaar D, Womack JE, Georges M. 1999. Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. Genetics, 96: 9252-9257.

Snelling WM, Casas E, Stone RT, Keele JW, Harhay GP, Bennett GL, Smith TP. 2005. Linkage mapping bovine EST-based SNP. BMC Genom, 6: 74.

Tenesa A, Knott SA, Ward D, Smith D, Williams JL, Visscher PM. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. J Anim Sci, 81(3): 617-623.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res, 17: 520-526.

Thévenon S, Dayo GK, Sylla S, Sidibe I, Berthier D, Legros H, Boichard D, Eggen A, Gautier M. 2007. The extent of linkage disequilibrium in a large cattle population of western Africa and its consequences for association studies. Anim Genet, 38(3): 277-286.

Vallejo RL, Li YL, Rogers GW, Ashwell MS. 2003. Genetic diversity and background linkage disequilibrium in the North American Holstein cattle population. J Dairy Sci, 86(12): 4137-4147.

Zhao H, Nettleton D, Soller M, Dekkers JCM. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet Res, 86: 77-87.