WIKIPEDIA

Linkage disequilibrium

In population genetics, **linkage disequilibrium** is the non-random association of alleles at different loci in a given population. Loci are said to be in linkage disequilibrium when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly.[1]

Linkage disequilibrium is influenced by many factors, including selection, the rate of recombination, the rate of mutation, genetic drift, the system of mating, population structure, and genetic linkage. As a result, the pattern of linkage disequilibrium in a genome is a powerful signal of the population genetic processes that are structuring it.

In spite of its name, linkage disequilibrium may exist between alleles at different loci without any genetic linkage between them and independently of whether or not allele frequencies are in equilibrium (not changing with time).[1] Furthermore, linkage disequilibrium is sometimes referred to as **gametic phase disequilibrium**;[2] however, the concept also applies to asexual organisms and therefore does not depend on the presence of gametes.

Contents

Formal definition

Suppose that among the gametes that are formed in a sexually reproducing population, allele A occurs with frequency p_A at one locus (i.e. p_A is the proportion of gametes with A at that locus), while at a different locus allele B occurs with frequency p_B. Similarly, let p_{AB} be the frequency with which both A and B occur together in the same gamete (i.e. p_{AB} is the frequency of the AB haplotype).

The association between the alleles *A* and *B* can be regarded as completely random—which is known in statistics as *independence*—when the occurrence of one does not affect the occurrence of the other, in which case the probability that both *A* and *B* occur together is given by the product $p_A p_B$ of the probabilities. There is said to be a linkage disequilibrium between the two alleles whenever p_{AB} differs from $p_A p_B$ for any reason.

The level of linkage disequilibrium between *A* and *B* can be quantified by the *coefficient of linkage disequilibrium* D_{AB}, which is defined as

$$D_{AB} = p_{AB} - p_A p_B\,,$$

provided that both p_A and p_B are greater than zero. Linkage disequilibrium corresponds to $D_{AB} \neq 0$. In the case $D_{AB} = 0$ we have $p_{AB} = p_A p_B$ and the alleles *A* and *B* are said to be in *linkage equilibrium*. The subscript "AB" on D_{AB} emphasizes that linkage disequilibrium is a property of the pair {*A*, *B*} of alleles and not of their respective loci. Other pairs of alleles at those same two loci may have different coefficients of linkage disequilibrium.

Linkage disequilibrium in asexual populations can be defined in a similar way in terms of population allele frequencies. Furthermore, it is also possible to define linkage disequilibrium among three or more alleles, however these higher-order associations are not commonly used in practice.[1]

Measures derived from D

The coefficient of linkage disequilibrium D is not always a convenient measure of linkage disequilibrium because its range of possible values depends on the frequencies of the alleles it refers to. This makes it difficult to compare the level of linkage disequilibrium between different pairs of alleles.

Lewontin[3] suggested normalising *D* by dividing it by the theoretical maximum difference between the observed and expected allele frequencies as follows:

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

where

$$D_{\max} = \begin{cases} \max\{-p_A p_B,\, -(1-p_A)(1-p_B)\} & \text{when } D < 0 \\ \min\{p_A(1-p_B),\, (1-p_A)p_B\} & \text{when } D > 0 \end{cases}$$

An alternative to D' is the correlation coefficient between pairs of loci, expressed as

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

Example: Two-loci and two-alleles

Consider the haplotypes for two loci A and B with two alleles each—a two-locus, two-allele model. Then the following table defines the frequencies of each combination:

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

Note that these are relative frequencies. One can use the above frequencies to determine the frequency of each of the alleles:

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

If the two loci and the alleles are independent from each other, then one can express the observation A_1B_1 as "A_1 is found and B_1 is found". The table above lists the frequencies for A_1, p_1, and forB_1, q_1, hence the frequency of A_1B_1 is x_{11}, and according to the rules of elementary statistics $x_{11} = p_1q_1$.

The deviation of the observed frequency of a haplotype from the expected is a quantity[4] called the linkage disequilibrium[5] and is commonly denoted by a capital D:

$$D = x_{11} - p_1q_1$$

The following table illustrates the relationship between the haplotype frequencies and allele frequencies and D.

	A_1	A_2	Total
B_1	$x_{11} = p_1q_1 + D$	$x_{21} = p_2q_1 - D$	q_1
B_2	$x_{12} = p_1q_2 - D$	$x_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

Role of recombination

In the absence of evolutionary forces other than random mating, Mendelian segregation, random chromosomal assortment, and chromosomal crossover (i.e. in the absence of natural selection, inbreeding, and genetic drift), the linkage disequilibrium measure D converges to zero along the time axis at a rate depending on the magnitude of the recombination rate c between the two loci.

Using the notation above, $D = x_{11} - p_1q_1$, we can demonstrate this convergence to zero as follows. In the next generation, x'_{11}, the frequency of the haplotype A_1B_1, becomes

$$x'_{11} = (1 - c)\, x_{11} + c\, p_1q_1$$

This follows because a fraction $(1-c)$ of the haplotypes in the offspring have not recombined, and are thus copies of a random haplotype in their parents. A fraction x_{11} of those are A_1B_1. A fraction c have recombined these two loci. If the parents result from random mating, the probability of the copy at locus A having allele A_1 is p_1 and the probability of the copy at locus B having allele B_1 is q_1, and as these copies are initially in the two different gametes that formed the diploid genotype, these are independent events so that the probabilities can be multiplied.

This formula can be rewritten as

$$x'_{11} - p_1 q_1 = (1-c)\,(x_{11} - p_1 q_1)$$

so that

$$D_1 = (1-c)\,D_0$$

where D at the n-th generation is designated as D_n. Thus we have

$$D_n = (1-c)^n\ D_0 .$$

If $n \to \infty$, then $(1-c)^n \to 0$ so that D_n converges to zero.

If at some time we observe linkage disequilibrium, it will disappear in the future due to recombination. However, the smaller the distance between the two loci, the smaller will be the rate of convergence of D to zero.

Example: Human leukocyte antigen (HLA) alleles

HLA constitutes a group of cell surface antigens as MHC of humans. Because HLA genes are located at adjacent loci on the particular region of a chromosome and presumed to exhibit epistasis with each other or with other genes, a sizable fraction of alleles are in linkage disequilibrium.

An example of such linkage disequilibrium is between HLA-A1 and B8 alleles in unrelated Danes[6] referred to by Vogel and Motulsky (1997).[7]

Because HLA is codominant and HLA expression is only tested locus by locus in surveys, LD measure is to be estimated from such a 2x2 table to the right.[7][8][9][10]

Table 1. Association of HLA-A1 and B8 in unrelated Danes[6]

			Antigen j		**Total**
			+	−	
			$B8^+$	$B8^-$	
Antigen i	+	$A1^+$	$a = 376$	$b = 237$	C
	−	$A1^-$	$c = 91$	$d = 1265$	D
Total			A	B	N
			No. of individuals		

expression (+) frequency of antigen i :

$$pf_i = C/N = 0.311\,;$$

expression (+) frequency of antigen j :

$$pf_j = A/N = 0.237\,;$$

frequency of gene i :

$$gf_i = 1 - \sqrt{1 - pf_i} = 0.170,$$

and

$$hf_{ij} = \text{estimated frequency of haplotype } ij = gf_i \; gf_j = 0.0215.$$

Denoting the '—' alleles at antigen i to be 'x,' and at antigen j to be 'y,' the observed frequency of haplotype xy is

$$o[hf_{xy}] = \sqrt{d/N}$$

and the estimated frequency of haplotype xy is

$$e[hf_{xy}] = \sqrt{(D/N)(B/N)}.$$

Then LD measure Δ_{ij} is expressed as

$$\Delta_{ij} = o[hf_{xy}] - e[hf_{xy}] = \frac{\sqrt{Nd} - \sqrt{BD}}{N} = 0.0769\,.$$

Standard errors ***SEs*** are obtained as follows:

$$SE \text{ of } gf_i = \sqrt{C}/(2N) = 0.00628,$$
$$SE \text{ of } hf_{ij} = \sqrt{\frac{(1-\sqrt{d/B})(1-\sqrt{d/D}) - hf_{ij} - hf_{ij}^2/2}{2N}} = 0.00514$$
$$SE \text{ of } \Delta_{ij} = \frac{1}{2N}\sqrt{a - 4N\Delta_{ij}\left(\frac{B+D}{2\sqrt{BD}} - \frac{\sqrt{BD}}{N}\right)} = 0.00367.$$

Then, if

$$t = \Delta_{ij}/(SE \text{ of } \Delta_{ij})$$

exceeds 2 in its absolute value, the magnitude of Δ_{ij} is statistically significantly large. For data in Table 1 it is 20.9, thus existence of statistically significant LD between A1 and B8 in the population is admitted.

Table 2 shows some of the combinations of HLA-A and B alleles where significant LD was observed among pan-europeans.[10]

Vogel and Motulsky (1997)[7] argued how long would it take that linkage disequilibrium between loci of HLA-A and B disappeared. Recombination between loci of HLA-A and B was considered to be of the order of magnitude 0.008. We will argue similarly to Vogel and Motulsky below. In case LD measure was observed to be 0.003 in Pan-europeans in the list of Mittal[10] it is mostly non-significant. If Δ_0 had reduced from 0.07 to 0.003 under recombination effect as shown by $\Delta_n = (1-c)^n\Delta_0$, then $n \approx 400$. Suppose a generation took 25 years, this means 10,000 years. The time span seems rather short in the history of humans. Thus observed linkage disequilibrium between HLA-A and B loci might indicate some sort of interactive selection.[7]

The presence of linkage disequilibrium between an HLA locus and a presumed major gene of disease susceptibility corresponds to any of the following phenomena:

- Relative risk for the person having a specific HLA allele to become suffered from a particular disease is greater than 1.[11]
- The HLA antigen frequency among patients exceeds more than that among a healthy population. This is evaluated by δ value[12] to exceed 0.
- 2x2 association table of patients and healthy controls with HLA alleles shows a significant deviation from the equilibrium state deduced from the marginal frequencies.

Table 2. Linkage disequilibrium among HLA alleles in Pan-europeans[10]

HLA-A alleles i	HLA-B alleles j	Δ_{ij}	t
A1	B8	0.065	16.0
A3	B7	0.039	10.3
A2	Bw40	0.013	4.4
A2	Bw15	0.01	3.4
A1	Bw17	0.014	5.4
A2	B18	0.006	2.2
A2	Bw35	-0.009	-2.3
A29	B12	0.013	6.0
A10	Bw16	0.013	5.9

(1) Relative risk

Relative risk of an HLA allele for a disease is approximated by the odds ratio in the 2x2 association table of the allele with the disease. Table 3 shows association of HLA-B27 with ankylosing spondylitis among a Dutch population.[13] Relative risk $x$of this allele is approximated by

Table 3. Association of ankylosing spondylitis with HLA-B27 allele[13]

		Ankylosing spondylitis		Total
		Patients	Healthy controls	
HLA alleles	$B27^+$	$a = 96$	$b = 77$	C
	$B27^-$	$c = 22$	$d = 701$	D
Total		A	B	N

$$x = \frac{a/b}{c/d} = \frac{ad}{bc}\ (= 39.7,\ \text{in Table 3}\ ).$$

Woolf's method[14] is applied to see if there is statistical significance. Let

$$y = \ln(x)\ (= 3.68)$$

and

$$\frac{1}{w} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\ (= 0.0703).$$

Then

$$\chi^2 = wy^2\ \ \left[= 193 > \chi^2(p = 0.001,\ df = 1) = 10.8\right]$$

follows the chi-square distribution with $df = 1$. In the data of Table 3, the significant association exists at the 0.1% level. Haldane's[15] modification applies to the case when either of$a,\ b,\ c,$ and d is zero, where replace x and $1/w$with

$$x = \frac{(a+1/2)(d+1/2)}{(b+1/2)(c+1/2)}$$

and

$$\frac{1}{w} = \frac{1}{a+1} + \frac{1}{b+1} + \frac{1}{c+1} + \frac{1}{d+1},$$

respectively.

Table 4. Association of HLA alleles with rheumatic and autoimmune diseases among white populations[11]

Disease	HLA allele	Relative risk (%)	FAD (%)	FAP (%)	δ
Ankylosing spondylitis	B27	90	90	8	0.89
Reactive arthritis	B27	40	70	8	0.67
Spondylitis in inflammatory bowel disease	B27	10	50	8	0.46
Rheumatoid arthritis	DR4	6	70	30	0.57
Systemic lupus erythematosus	DR3	3	45	20	0.31
Multiple sclerosis	DR2	4	60	20	0.5
Diabetes mellitus type 1	DR4	6	75	30	0.64

In Table 4, some examples of association between HLA alleles and diseases are presented.[11]

(1a) Allele frequency excess among patients over controls

Even high relative risks between HLA alleles and the diseases were observed, only the magnitude of relative risk would not be able to determine the strength of association.[12] δ value is expressed by

$$\delta = \frac{FAD - FAP}{1 - FAP}, \ \ 0 \leq \delta \leq 1,$$

where FAD and FAP are HLA allele frequencies among patients and healthy populations, respectively.[12] In Table 4, δ column was added in this quotation. Putting aside 2 diseases with high relative risks both of which are also with high δ values, among other diseases, juvenile diabetes mellitus (type 1) has a strong association with DR4 even with a low relative risk$= 6$.

(2) Discrepancies from expected values from marginal frequencies in 2x2 association table of HLA alleles and disease

This can be confirmed by χ^2 test calculating

$$\chi^2 = \frac{(ad-bc)^2 N}{ABCD} \ (= 336, \text{ for data in Table 3; } P < 0.001).$$

where $df = 1$. For data with small sample size, such as no marginal total is greater than 15 (and consequently $N \leq 30$), one should utilize Yates's correction for continuity or Fisher's exact test.[16]

Resources

A comparison of different measures of LD is provided by Devlin & Risch[17]

The International HapMap Project enables the study of LD in human populations online (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html). The Ensembl project integrates HapMap data with other genetic information from dbSNP.

Analysis software

- PLINK (http://zzz.bwh.harvard.edu/plink/) - whole genome association analysis toolset, which can calculate LD among other things
- LDHat (http://www.stats.ox.ac.uk/~mcvean/LDhat/)
- Haploview
- LdCompare (https://www.ncbi.nlm.nih.gov/pubmed/17148510)[18]— open-source software for calculating LD.
- SNP and Variation Suite (http://goldenhelix.com/products/SNP_Variation/index.html)- commercial software with interactive LD plot.
- GOLD (http://www.sph.umich.edu/csg/abecasis/GOLD/index.html) - Graphical Overview of Linkage Disequilibrium
- TASSEL (http://www.maizegenetics.net/tassel) -software to evaluate linkage disequilibrium, traits associations, and evolutionary patterns
- rAggr (http://raggr.usc.edu/) - finds proxy markers (SNPs and indels) that are in linkage disequilibrium with a set of queried markers, using the 1000 Genomes Project and HapMap genotype databases.
- SNeP (https://sourceforge.net/projects/snepnetrends/) - Fast computation of LD and Ne for large genotype datasets in PLINK format.

Simulation software

- Haploid (http://haploid.nongnu.org) — a C library for population genetic simulation (GPL)

See also

- Haploview
- Hardy-Weinberg principle
- Genetic linkage
- Co-adaptation
- Genealogical DNA test
- Tag SNP
- Association Mapping
- Family based QTL mapping

References

1. Slatkin, Montgomery (June 2008). "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124487). *Nature Reviews Genetics*. **9** (6): 477–485. doi:10.1038/nrg2361 (https://doi.org/10.1038%2Fnrg2361). PMC 5124487 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124487). PMID 18427557 (https://www.ncbi.nlm.nih.go

v/pubmed/18427557).

2. Falconer, DS; Mackay, TFC (1996). *Introduction to Quantitative Genetics* (4th ed.). Harlow, Essex, UK: Addison Wesley Longman. ISBN 0-582-24302-5.
3. Lewontin, R. C. (1964). "The interaction of selection and linkage. I. General considerations; heterotic models" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210557). *Genetics*. **49** (1): 49–67. PMC 1210557 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210557). PMID 17248194 (https://www.ncbi.nlm.nih.gov/pubmed/17248194).
4. Robbins, R.B. (1 July 1918). "Some applications of mathematics to breeding problems III" (http://www.genetics.org/cgi/reprint/3/4/375). *Genetics*. **3** (4): 375–389. PMC 1200443 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1200443). PMID 17245911 (https://www.ncbi.nlm.nih.gov/pubmed/17245911).
5. R.C. Lewontin & K. Kojima (1960). "The evolutionary dynamics of complex polymorphisms". *Evolution*. **14** (4): 458–472. doi:10.2307/2405995 (https://doi.org/10.2307%2F2405995). ISSN 0014-3820 (https://www.worldcat.org/issn/0014-3820). JSTOR 2405995 (https://www.jstor.org/stable/2405995).
6. Svejgaard A, Hauge M, Jersild C, Plaz P, Ryder LP, Staub Nielsen L, Thomsen M (1979). *The HLA System: An Introductory Survey, 2nd ed.* Basel; London; Chichester: Karger; Distributed by Wiley, ISBN 3805530498(pbk).
7. Vogel F, Motulsky AG (1997). *Human Genetics : Problems and Approaches, 3rd ed.*Berlin; London: Springer, ISBN 3-540-60290-9.
8. Mittal KK, Hasegawa T, Ting A, Mickey MR, Terasaki PI (1973). "Genetic variation in the HL-A system between Ainus, Japanese, and Caucasians," *In* Dausset J, Colombani J, eds. *Histocompatibility Testing, 1972,* pp. 187-195, Copenhagen: Munksgaard, ISBN 87-16-01101-5.
9. Yasuda, N; Tsuji, K (June 1975). "A counting method of maximum likelihood for estimating haplotype frequency in the HL-A system". *Jinrui Idengaku Zasshi*. **20** (1): 1–15. PMID 1237691 (https://www.ncbi.nlm.nih.gov/pubmed/1237691).
10. Mittal, KK (1976). "The HLA polymorphism and susceptibility to disease". *Vox Sang*. 31?-73 (3): 161–73. PMID 969389 (https://www.ncbi.nlm.nih.gov/pubmed/969389).
11. Gregersen PK (2009). "Genetics of rheumatic diseases," *In*Firestein GS, Budd RC, Harris ED Jr, McInnes IB, Ruddy S, Sergent JS, eds. (2009). *Kelley's Textbook of Rheumatology,* pp. 305-321, Philadelphia, PA: Saunders/Elsevier, ISBN 978-1-4160-3285-4.
12. Bengtsson, BO; Thomson, G (November 1981). "Measuring the strength of associations between HLA antigens and diseases". *Tissue Antigens*. **18** (5): 356–63. doi:10.1111/j.1399-0039.1981.tb01404.x (https://doi.org/10.1111%2Fj.1399-0039.1981.tb01404.x). PMID 7344182 (https://www.ncbi.nlm.nih.gov/pubmed/7344182).
13. Nijenhuis, LE (September 1977). "Genetic considerations on association between HLA and disease". *Hum. Genet*. **38** (2): 175–82. doi:10.1007/bf00527400 (https://doi.org/10.1007%2Fbf00527400). PMID 908564 (https://www.ncbi.nlm.nih.gov/pubmed/908564).
14. Woolf, B (June 1955). "On estimating the relation between blood group and disease". *Ann. Hum. Genet*. **19** (4): 251–3. doi:10.1111/j.1469-1809.1955.tb01348.x (https://doi.org/10.1111%2Fj.1469-1809.1955.tb01348.x). PMID 14388528 (https://www.ncbi.nlm.nih.gov/pubmed/14388528).
15. Haldane, JB (May 1956). "The estimation and significance of the logarithm of a ratio of frequencies". *Ann. Hum. Genet*. **20** (4): 309–11. doi:10.1111/j.1469-1809.1955.tb01285.x (https://doi.org/10.1111%2Fj.1469-1809.1955.tb01285.x). PMID 13314400 (https://www.ncbi.nlm.nih.gov/pubmed/13314400).
16. Sokal RR, Rohlf FJ (1981). *Biometry: The Principles and Practice of Statistics in Biological Research.* Oxford: W.H. Freeman, ISBN 0-7167-1254-7.
17. Devlin B.; Risch N. (1995). "A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping" (http://biostat.jhsph.edu/~iruczins/teaching/misc/gwas/papers/devlin1995.pdf) (PDF). *Genomics*. **29** (2): 311–322. doi:10.1006/geno.1995.9003 (https://doi.org/10.1006%2Fgeno.1995.9003). PMID 8666377 (https://www.ncbi.nlm.nih.gov/pubmed/8666377).

18. Hao K.; Di X.; Cawley S. (2007). "LdCompare: rapid computation of single- and multiple-marker r2 and genetic coverage" (http://bioinformatics.oxfordjournals.org/cgi/reprint/23/2/252). *Bioinformatics*. **23** (2): 252–254. doi:10.1093/bioinformatics/btl574 (https://doi.org/10.1093%2Fbioinformatics%2Fbtl574). PMID 17148510 (https://www.ncbi.nlm.nih.gov/pubmed/17148510).

Further reading

- Hedrick, Philip W. (2005). *Genetics of Populations* (3rd ed.). Sudbury, Boston, Toronto, London, Singapore: Jones and Bartlett Publishers. ISBN 0-7637-4772-6.
- Bibliography: Linkage Disequilibrium Analysis (https://web.archive.org/web/20071013084046/http://www.nslij-genetics.org/ld/) : a bibliography of more than one thousand articles on Linkage disequilibrium published since 1918.