ENCYCLOPEDIA OF LIFE SCIENCES

Population genetics: multilocus (article reference code: 1783)

Sergey Gavrilets
Departments of Ecology and Evolutionary
Biology and Mathematics
University of Tennessee, Knoxville, TN 37996 USA
phone: (423) 974-8136
fax: (423) 974-3067
email: gavrila@tiem.utk.edu

Keywords: haplotype, linkage equilibrium and disequilibrium, multilocus selection

Haplotypes
Linkage equilibrium and disequilibrium
Additive, multiplicative, epistatic fitness regimes
Causes of linkage disequilibrium
Extent of disequilibrium in nature

Multilocus population genetics is the study of the distribution and dynamics of genetic variation at several loci in biological populations.

Haplotypes

Even the simplest known biological organisms have thousands of genes and millions of DNA base pairs. This makes it necessary to study the distribution and dynamics of genetic variation at many loci (or genetic markers) simultaneously. A particular combination of genes (or genetic markers) for a DNA segment under consideration is referred to as haplotype. In what follows different haplotypes will be represented by sequences of bold letters corresponding to different alleles (for example **A, aB, ABC**) whereas the frequencies of these haplotypes in the population will be represented by variable x with a corresponding subscript (for example $x_{\mathbf{A}}, x_{\mathbf{aB}}, x_{\mathbf{ABC}}$). If a diploid population is at Hardy-Weinberg equilibrium, then genotype frequencies can be found as products of the corresponding haplotype frequencies. In this case, the population state can be described in terms of the haplotype frequencies.

Linkage equilibrium and disequilibrium

If there is no statistical association of alleles in haplotypes, then the frequency of a haplotype is equal to the product of the corresponding allele frequencies. In this case, one says that the population is at *linkage equilibrium*. For example, in such a population $x_{\mathbf{aB}} = x_{\mathbf{a}}x_{\mathbf{B}}, x_{\mathbf{ABC}} = x_{\mathbf{A}}x_{\mathbf{B}}x_{\mathbf{C}}$ and so on. If there is statistical association of alleles in haplotypes, one says that the population is at *linkage disequilibrium*. [The term "linkage disequilibrium" is somewhat misleading because even unlinked loci can be in "linkage disequilibrium" and because other factors besides linkage can affect the degree of statistical association of different alleles. Alternative terms used are "gametic (dis)equilibrium" and "gametic phase (dis)equilibrium".]

A standard measure of the statistical association between a pair of alleles **A** and **B** at two different loci is linkage (or gametic) disequilibrium $D_{\mathbf{AB}}$ defined as the deviation of the haplotype frequency from that expected at linkage equilibrium:

$$D_{\mathbf{AB}} = x_{\mathbf{AB}} - x_{\mathbf{A}}x_{\mathbf{B}}.$$

For a population at linkage equilibrium $D_{\mathbf{AB}} = 0$. The coefficient D can be positive or negative depending on whether alleles **A** and **B** are in *coupling* disequilibrium (haplotypes **AB** are over-represented) or in *repulsion* disequilibrium (haplotypes **AB** are underrepresented). Large linkage disequilibrium values mean that haplotype frequencies cannot be found, even approximately, as products of the corresponding allele frequencies. In the diallelic case that is when there are only two alleles at each locus under consideration, say alleles **A** and **a** at the first locus and alleles **B** and **b** at the second locus, the above definition of $D_{\mathbf{AB}}$ is equivalent to

$$D_{\mathbf{AB}} = x_{\mathbf{AB}}x_{\mathbf{ab}} - x_{\mathbf{Ab}}x_{\mathbf{aB}}.$$

Thus, D can be found as the difference of the products of the frequencies of complimentary haplotypes. The range of possible values of D depends on the allele frequencies: D cannot be larger than $\min(x_Ax_b, x_ax_B)$ and cannot be smaller than $\max(-x_Ax_B, -x_ax_b)$. ["min" and "max" mean "the smaller of" and "the larger of", respectively.] Besides D, other measures of linkage disequilibrium have been used. One such measure is the square of the correlation coefficient between the presence of alleles defined as

$$R^2 = \frac{D^2}{x_Ax_Bx_ax_b}.$$

This measure has a range from zero (linkage equilibrium) to one (complete linkage disequilibrium).

Recombination re-shuffles alleles reducing the absolute value of linkage disequilibrium. Assume that the population size is sufficiently large, mating is random, and that selection and other factors are absent. If $D_{\mathbf{AB}}(t)$ is linkage disequilibrium at generation t, then in the next generation

$$D_{\mathbf{AB}}(t+1) = (1 - r_{AB})D_{\mathbf{AB}}(t),$$

where r_{AB} is the probability of recombination between the two loci under consideration ($0 \le r_{AB} \le 0.5$). This shows that as time increases, the population approaches the state of linkage equilibrium (at which $D_{\mathbf{AB}} = 0$). The approach is very rapid for unlinked loci (for which $r_{AB} = 0.5$) but can be very slow for closely linked loci (for which $r_{AB} << 0.5$). For example, with $r = 0.001$, it will take about 700 generations of random mating to reduce D by half. Some degree of inbreeding or asexual reproduction in the population will reduce the rate of decay of disequilibrium in a manner similar to that resulting from physical linkage.

There are several ways to introduce higher order disequilibria between specific alleles at more than two loci. A common approach is to use indicator variables l_i equal to 1 if the corresponding allele is present at the i-locus and equal to 0 otherwise, and to define the n-th order disequilibrium between n alleles at different loci as the n-th order covariance:

$$D_{12\ldots n} = E[(l_1 - p_1)(l_2 - p_2)\ldots(l_n - p_n)]$$

(Slatkin 1972). Here p_i is the frequency of the allele at the i-th locus and $E[\ldots]$ means mathematical

expectation. With this definition, the linkage disequilibrium between alleles **A, B** and **C** at three different loci is

$$D_{\mathbf{ABC}} = x_{\mathbf{ABC}} - x_{\mathbf{A}} D_{\mathbf{BC}} - x_{\mathbf{B}} D_{\mathbf{AC}} - x_{\mathbf{C}} D_{\mathbf{AB}} - x_{\mathbf{A}} x_{\mathbf{B}} x_{\mathbf{C}},$$

where $D_{\mathbf{BC}}, D_{\mathbf{AC}}$ and $D_{\mathbf{AB}}$ are the corresponding pairwise linkage disequilibria. The linkage disequilibrium between alleles **A, B, C** and **D** at four different loci is

$$\begin{aligned} D_{\mathbf{ABCD}} = x_{\mathbf{ABCD}} &- x_{\mathbf{A}} D_{\mathbf{BCD}} - x_{\mathbf{B}} D_{\mathbf{ACD}} - x_{\mathbf{C}} D_{\mathbf{ABD}} - x_{\mathbf{D}} D_{\mathbf{ABC}} \\ &- x_{\mathbf{A}} x_{\mathbf{B}} D_{\mathbf{CD}} - x_{\mathbf{A}} x_{\mathbf{C}} D_{\mathbf{BD}} - x_{\mathbf{A}} x_{\mathbf{D}} D_{\mathbf{BC}} \\ &- x_{\mathbf{B}} x_{\mathbf{C}} D_{\mathbf{AD}} - x_{\mathbf{B}} x_{\mathbf{D}} D_{\mathbf{AC}} - x_{\mathbf{C}} x_{\mathbf{D}} D_{\mathbf{AB}} - x_{\mathbf{A}} x_{\mathbf{B}} x_{\mathbf{C}} x_{\mathbf{D}}. \end{aligned}$$

The set of allele frequencies and linkage disequilibria of different orders provides an alternative way to describe the state of a population at Hardy-Weinberg equilibrium. If the population under consideration is at linkage equilibrium, its analysis is significantly simplified because one has to concentrate only on allele frequencies. However different factors operating in natural populations are expected to introduce and maintain linkage disequilibrium. If the loci are in a state of linkage disequilibrium, the changes in allele frequency in one locus are not independent of changes at another locus. Linkage disequilibrium is relevant to a variety of evolutionary topics including molecular evolution, phenotypic evolution, sexual selection, evolution of sex and recombination, and speciation. Analysis of linkage disequilibrium is extremely useful and important in medical genetics where it can provide high resolution in the mapping of disease genes. Linkage disequilibrium is extensively used in different marker-assisted selection protocols used in agricultural breeding for the improvement of quantitative characters.

Additive, multiplicative, epistatic fitness regimes

Analysis of the dynamics of the genetic structure of populations under different selection regimes has been a focus of multilocus population genetics. Biological organisms are different with respect to various characteristics affecting fitness (e.g. viability, fecundity, fertility, attractiveness for the opposite sex etc). Multilocus theory has been mostly developed in the context of *viability selection* where fitness (viability) is defined as the probability of survival to the age of reproduction. Under this form of selection, populations are at Hardy-Weinberg equilibrium, and, thus, the population state can be completely characterized in terms of haplotype frequencies.

Let variables I, J and K denote different L-locus haplotypes. The genotype of a diploid organism can be described by a pair of these variables, for example JK. In a randomly mating diploid population with discrete non-overlapping generations experiencing viability selection defined by genotype fitnesses w_{JK}, the dynamics of the frequency x_I of haplotype I are described by a general recurrence equation

$$x_I(t+1) = \frac{\sum_J \sum_K w_{JK} x_J(t) x_K(t) R(J, K \to I)}{\overline{w}},$$

where

$$\overline{w} = \sum_J \sum_K w_{JK} x_J(t) x_K(t)$$

is the mean fitness of the population and $R(J, K \to I)$ is the probability that a genotype formed

from haplotypes J and K produces gamete I.

The case of two diallelic loci has been studied in much detail. In this case there are four different haplotypes, say **AB, Ab, aB** and **ab**. Let x_1, x_2, x_2 and x_4 be the corresponding haplotype frequencies. If there are no *cis-trans* effects (that is the fitnesses of both double heterozygotes **AB/ab** and **Ab/aB** are identical: $w_{14} = w_{23}$), the dynamic equations for the haplotype frequencies can be rewritten as

$$x_I(t+1) = \frac{w_I}{\overline{w}} x_I(t) \mp \frac{r w_{14} D}{\overline{w}}$$

where r is the rate of recombination between the loci, D is the linkage disequilibrium, $w_I = \sum_J w_{IJ} x_J$ is the induced (or average) fitness of haplotype I, and the sign is $+$ for $I = 2, 3$ and $-$ for $I = 1, 4$. The dynamic equations given above have been used to approach three main questions: 1) the maintenance of genetic variation under selection; 2) levels of linkage disequilibrium expected, and 3) the dependence of the outcome of evolutionary dynamics on initial conditions. Exact analytical results exist for several different fitness regimes.

Table 1: Additive fitnesses in the two-locus two-allele case.

	BB	**Bb**	**bb**
AA	$a_1 + b_1$	$a_1 + b_2$	$a_1 + b_3$
Aa	$a_2 + b_1$	$a_2 + b_2$	$a_2 + b_3$
aa	$a_3 + b_1$	$a_3 + b_2$	$a_3 + b_3$

In the *additive fitness regime*, the fitness, w, of an organism is found by summing up the contributions, w_i, of L individual loci:

$$w = w_1 + w_2 + \ldots + w_L.$$

The additive model may be a reasonable approximation if contributions of individual loci to fitness (or another trait under consideration) are small. Table 1 gives an example of additive fitnesses in the two-locus two-allele case ($L = 2$). Here, the entry in the i-th row and j-th columns shows the fitness of a genotype that has the specified genes in the first and second loci. Under additive fitness regime, genetic variation at a locus is maintained only if there is *overdominance* that is the heterozygous locus has higher fitness than homozygous loci. For instance using Table 1, genetic variation in locus A will be maintained only if $a_2 > a_1, a_3$. The population always evolves to a state of linkage equilibrium. With *underdominance* that is the heterozygous locus has lower fitness than homozygous loci (e.g. $a_2 < a_1, a_3$) either allele can be fixed. Which one will be fixed depends on initial conditions.

In the *multiplicative fitness regime*, the fitness, w, of an organism is found by multiplying the contributions, w_i, of individual loci:

$$w = w_1 \times w_2 \times \ldots \times w_L.$$

The multiplicative model may be a reasonable approximation if the individual loci contribute to fitness (or another trait under consideration) at different time moments. Table 2 gives an example of multiplicative fitnesses in the two-locus two-allele case. Under multiplicative fitness regime, genetic

Table 2: Multiplicative fitnesses in the two-locus two-allele case

	BB	**Bb**	**bb**
AA	a_1b_1	a_1b_2	a_1b_3
Aa	a_2b_1	a_2b_2	a_2b_3
aa	a_3b_1	a_3b_2	a_3b_3

variation at a locus is maintained only if there is overdominance. If genetic variation is maintained in both loci under consideration, then in general the population will evolve to a linkage equilibrium state if the rate of recombination between the loci is high enough, but to a linkage disequilibrium state if the recombination rate is small. Several different linkage disequilibrium states can be stable simultaneously and, thus, the outcome of evolution will depend on initial conditions. It is also possible that both linkage equilibrium state and several different linkage disequilibrium states are stable simultaneously.

Both the additive and multiplicative models imply that *epistasis* is absent - in these models fitness does not depend on interactions between alleles at different loci. Because of the mathematical difficulties in direct analyses of *epistatic fitness regimes*, in which alleles at different loci interact in controlling fitness, one uses different simplifications and approximations.

One approach is to introduce some symmetries in the model. A well-studied two-locus two-allele symmetric fitness model is given in Table 3. This model has a very rich spectrum of dynamic behaviors including existence of up to 4 simultaneously stable polymorphic equilibria and simultaneous stability of equilibria with $D = 0$ and $D \neq 0$ (Karlin and Feldman 1970; Hastings 1985). Simultaneous stability of different equilibria implies the dependence of the outcome of the dynamics on initial conditions (and history). The symmetric model has been generalized to more than two loci (e.g. Feldman et al. 1974; Christiansen 1990). Other approaches are to consider only some specific types of epistatic interaction (for example, additive-by-additive pairwise interactions as in Zhivotovsky and Gavrilets 1992), or to assume that each locus only interacts with a specified number of other loci (Kauffman 1993, Ch. 2), or to limit the number of different fitness values (for example, pulling all well-fit genotypes in one fitness class and all inviable genotypes in another fitness class as in Gavrilets 1997).

Table 3: Symmetric fitnesses in the two-locus two-allele case

	BB	**Bb**	**bb**
AA	δ	β	α
Aa	γ	η	γ
aa	α	β	δ

Some experimental data show that both the strength of selection and the degree of epistasis are generally weak. This observation prompted the development of models allowing only for weak selection and/or weak epistasis. The weak selection assumption is especially powerful for under this approximation one can neglect linkage disequilibrium and study evolutionary dynamics in terms of

allele frequencies. For example if the i-th locus is diallelic then the change in allele frequency p_i under weak viability selection can be approximated as

$$\Delta p_i = \frac{p_i q_i}{2} \frac{\partial \overline{w}}{\partial p_i}$$

where $q_i = 1 - p_i$. The last equation shows that the allele frequency stops changing ($\Delta p_i = 0$) if one allele is fixed ($p_i = 0$ or $q_i = 0$) or if the allele frequency is such that the mean fitness of the population is at a local maximum or minimum ($\partial \overline{w}/\partial p_i = 0$). The states at which the mean fitness of the population is at a local minimum are unstable to small perturbations. If selection is weak (relative to recombination) the population evolves to a state at which the mean fitness $\overline{w}$ is maximized.

Finally, extensive numerical studies of different epistatic fitness regimes have been performed (e.g. Turelli and Ginzburg 1983; Gimelfarb 1998). The most common approach is to randomly assign fitness values to the set of genotypes under consideration (for example, by picking them up from a uniform random distribution between 0 and 1) and iterating the corresponding dynamic equations (such as given above) for random sets on initial conditions to collect the statistics regarding possible dynamical regimes.

Although finding some general principle governing multilocus dynamics has proven to be very difficult some generalizations can be made. The overall conclusion of analytical and numerical studies of different epistatic regimes is that populations are expected to evolve to a stationary state (but see Hastings 1981), that there are rich possibilities for the maintenance of genetic variation, that polymorphic loci exhibit induced overdominance, that linkage disequilibrium should be present between closely linked loci, and that the outcome of evolution can significantly depend on initial conditions and history. The latter means that natural populations still can diverge genetically even if they are under very similar fitness regimes (that is exist under similar ecological conditions). Close linkage promotes the maintenance of genetic variation. Maintaining genetic variation in a number of loci need not be accompanied by a very heavy genetic load.

Causes of linkage disequilibrium

Selection is not the only factor causing linkage disequilibrium. In a population of a finite size N, random genetic drift will almost certainly result in nonrandom association between alleles at different loci. For a randomly mating population, the expected squared correlation R^2 between the presence of two linked alleles is

$$E[R^2] = \frac{1}{1 + 4Nr},$$

where r is the recombination rate (Hill and Robertson 1968). The sign of linkage disequilibrium generated by random genetic drift can be both positive and negative. Increasing the population size and the rate of recombination both decrease the expected value of R^2.

Migration or hybridization of two populations with different allele frequencies can cause linkage disequilibrium. For example, mixing individuals from two populations at linkage equilibrium each in a single population in the proportions $m_1 : m_2$ will result in linkage disequilibrium value

$$D = m_1 m_2 (x_{\mathbf{A},\mathbf{1}} - x_{\mathbf{A},\mathbf{2}})(x_{\mathbf{B},\mathbf{1}} - x_{\mathbf{B},\mathbf{2}}),$$

where $x_{\mathbf{A},\mathbf{i}}$ and $x_{\mathbf{B},\mathbf{i}}$ are the frequencies of allele **A** and **B** in the i-th population ($i = 1, 2$). Population subdivision also reduces the rate of decay of linkage disequilibrium. Interaction of non-epistatic selection and migration can maintain stable linkage disequilibrium if different selection regimes operates in the populations connected by migration.

At the molecular level, mutations are often unique. A mutant allele at a locus will be initially associated with a particular allele at a second locus that happened to be present when the mutation occurred. Thus, mutation is expected to generate some linkage disequilibrium. If the new mutant has a selective advantage, it may increase in frequency. The allele associated with it may be carried along ("hitchhike"). Linkage disequilibrium can be very easily generated in this manner. In a large population, this type of disequilibria will eventually disappear if there is any recombination between the loci. But for small recombination rates linkage disequilibrium is expected to last for a very long time.

Non-random mating may result in linkage disequilibrium. Often, individuals chose mates whose phenotype resembles their own (positive assortative mating). Positive assortative mating increases the coupling of alleles with similar effects, resulting in the proliferation of coupling haplotypes. Negative (or disassortative) mating leads to the proliferation of repulsion gametes, at which effects of different loci are balanced. Another factor that can cause a build-up of linkage disequilibrium is inbreeding.

Linkage disequilibrium observed in a population can be caused by factors (such as selection, random drift, migration, mutation) no longer present that acted in the founding population and generated linkage disequilibrium that has not yet had time to decay due to small rates of recombination.

Extent of disequilibrium in nature

From theoretical considerations, many factors such as selection, drift, mutation, selection at linked loci, and non-random mating may be responsible for generating linkage disequilibrium, whereas linkage and some other factors retard the rate of decay of linkage disequilibrium. There is a variety of statistical methods for estimating linkage disequilibrium and testing hypotheses about its value (e.g. Hedrick et al. 1978; Weir 1996; Lynch and Walsh 1998). For a few experimental systems, such as *Drosophila*, one can count haplotypes directly and use the observed haplotype frequencies in statistical procedures. However, for most natural populations, the only available information are the frequencies of multilocus genotypes in the population under study. This requires some special methods for the haplotype composition of heterozygotes cannot be resolved definitely. In general, unless gene frequencies are close to 0.5 at both loci and disequilibrium is strong, hundreds of individuals should be assayed to achieve a reasonable level of statistical power.

In outbreeding population pairs of polymorphic loci are usually found in linkage equilibrium, or nearly so. However, linkage disequilibrium is common among polymorphic sites within genes, because recombination rates there are very low. Some of the most extreme examples of linkage disequilibrium arise from studies of plants that are predominantly self-fertilizing. Linkage disequilibrium is also common in asexual populations. Hedrick et al. (1978) review many examples of linkage disequilibrium in natural populations. In panmictic human populations linkage disequilibrium is usually not noticeable for genetic markers at distances exceeding 1 cM, but in parts of the genome it is seen at distances as long as 1-2 cM, whereas for some markers, linkage disequilibrium is absent for very short distances (Peterson et al. 1995).

References

Christiansen FB (1990) Simplified models for viability selection at multiple loci. *Theoretical Population Biology* **37**: 39-54.

Feldman MW, Franklin I and Thomson G (1974) Selection in complex genetic systems. I. The symmetric equilibria of the three-locus symmetric viability model. *Genetics* **76**: 135-162.

Gavrilets S (1997) Evolution and speciation on holey adaptive landscapes. *Trends in Ecology and Evolution* **12**: 307-312.

Gimelfarb A (1998) Stable equilibria in multilocus genetic systems: a statistical investigation. *Theoretical Population Biology* **54**: 133-154.

Hastings A (1981) Stable cycling in discrete time genetic models. *Proceedings of the National Academy of Science* **78**: 7224-7225.

Hastings A (1985) Four simultaneously stable polymorphic equilibria in two-locus two-allele models. *Genetics* **109**: 255-261.

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

Karlin S and Feldman MW (1970) Linkage and selection - two locus symmetric viability model. *Theoretical Population Biology* **1**: 39-71.

Kauffman SA (1993) *The Origins of Order* New York: Oxford University Press.

Lynch M and Walsh B (1998) *Genetics and Analysis of Quantitative Traits* Sunderland, Mass: Sinauer Associates.

Peterson AC, DiRienzo A, Lehesjoki AE, Delachapelle A et al. The distribution of linkage disequilibrium over anonymous genome regions. *Human Molecular Genetics* **4**: 887-894.

Slatkin M (1972) On treating the chromosome as the unit of selection. *Genetics* **72**: 157-168.

Turelli M and Ginzburg L (1983) Should individual fitness increase with heterozygosity. *Genetics* **104**: 191-209.

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

Zhivotovsky LA and Gavrilets S (1992) Quantitative variability and multilocus polymorphism under epistatic selection. *Theoretical Population Biology* **42**: 254-283.

Further reading

Ewens W (1979) *Mathematical Population Genetics* Berlin: Springer-Verlag.

Futuyma DJ (1997) *Evolutionary Biology* Sunderland, Mass: Sinauer Associates.

Hastings A (1989) Deterministic multilocus population genetics: an overview *Lectures on Mathematics in the Life Sciences* **20**: 27-54.

Hedrick PW, Jain S and Holden L (1978) Multilocus systems in evolution In: Hecht MK, Steer WC and Wallace B (eds) *Evolutionary Biology, Vol. 11* pp. 101-184. New York: Plenum Press.

Karlin S (1975) General two-locus selection models: some objectives, results and interpretations. *Theoretical Population Biology* **7**: 364-398

Karlin S (1978) Theoretical aspects of multi-locus selection balance In: Levin SA (ed) *Studies in Mathematical Biology Part II: Populations and Communities*, pp. 503-587. Washington: Math. Assoc. Amer.

Liubich IU I (1992) *Mathematical Structures in Population Genetics* Berlin: Springer-Verlag.

Nagylaki T (1992) *Introduction to Theoretical Population Genetics* Berlin: Springer-Verlag.

Svirezhev IU M and Passekov VP (1990) *Fundamentals of Mathematical Evolutionary Genetics* Dordrecht: Kluwer Academic Publishers.

Glossary

Epistasis Allelic interaction resulting in that the effect of a gene on a trait (or fitness) depends on the allelic state of one or more alleles at different loci.

Genetic load is the relative difference between the fitness of the most fit genotype, w_{max}, and the mean fitness of the population, $\overline{w}$: $L = (w_{max} - \overline{w})/w_{max}$. In most evolutionary and population genetics considerations, genetic load is used as a measure of the amount of natural selection associated with a certain amount of genetic variability.

Haplotype A particular combination of alleles (or genetic markers) for the DNA segment under consideration.

Linkage disequilibrium A state of non-random association of alleles (or genetic markers) in a haplotype. Also a measure of the statistical deviation of two or more alleles in a haplotype from random association.

Linkage equilibrium A state of random association of alleles (or genetic markers) in a haplotype. For a population at linkage equilibrium haplotype frequencies can be found as products of the corresponding allele frequencies.