LINKAGE DISEQUILIBRIUM IN SUBDIVIDED POPULATIONS[1]

MASATOSHI NEI AND WEN-HSIUNG LI

Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77025, and Department of Medical Genetics, University of Wisconsin, Madison, Wisconsin 53706

Manuscript received October 12, 1972

ABSTRACT

The linkage disequilibrium in a subdivided populaton is shown to be equal to the sum of the average linkage disequilibrium for all subpopulations and the covariance between gene frequencies of the loci concerned. Thus, in a subdivided population the linkage disequilibrium may not be 0 even if the linkage disequilibrium in each subpopulation is 0. If a population is divided into two subpopulations between which migration occurs, the asymptotic rate of approach to linkage equilibrium is equal to either r or $2(m_1 + m_2) - (m_1 + m_2)^2$, whichever is smaller, where r is the recombination value and m_1 and m_2 are the proportions of immigrants in subpopulations 1 and 2, respectively. Thus, if migration rate is high compared with recombination value, the change of linkage disequilibrium in subdivided populations is similar to that of a single random mating population. On the other hand, if migration rate is low, the approach to lnkage equilibrium may be retarded in subdivided populations. If isolated populations begin to exchange genes by migration, linkage disequilibrium may increase temporarily even for neutral loci. If overdominant selection operates and the equilibrium gene frequencies are different in the two subpopulations, a permanent linkage disequilibrium may be produced without epistasis in each subpopulation.

IN a single random mating population, linkage disequilibrium may be produced by two factors, namely, epistatic interaction in fitness between the loci concerned and genetic random drift due to finite population size. The first factor has been studied by WRIGHT (1952), KIMURA (1956), LEWONTIN and KOJIMA (1960) and others, while the second factor by HILL and ROBERTSON (1968), SVED (1968), and OHTA and KIMURA (1968a,b). However, most natural populations are not a single random mating unit but subdivided into local populations. In the following we shall show that subdivision of a population may also create linkage disequilibrium. In the case of two subpopulations, this problem was recently studied by SINNOCK and SING (1972).

Although the statistical description of linkage disequilibrium in subdivided populations is simple and holds true for any pattern of migration and selection, it does not give any information on the dynamics of linkage disequilibrium in populations. In order to have a rough idea about the effect of migration on linkage disequilibrium we shall study a simple mathematical model. It will be shown that under certain circumstances migration creates linkage disequilibrium with-

[1] Part of this work was done while the authors were affiliated with Brown University, Providence, R. I. Supported by Public Health Service grants GM-17719 and GM-20293 and National Science Foundation grant GB-21224 and GP-20868.

out epistatic selection. Our method will be deterministic rather than stochastic: a general stochastic treatment seems to be quite complicated.

Linkage Disequilibrium in a Subdivided Population

Consider a population of size N which is divided into k subpopulations. Let N_i be the size of the i-th subpopulation with $\sum_{i=1}^{k} N_i = N$. We consider two loci A and B at which pairs of alleles A:a and B:b are segregating respectively. Let the frequencies of the four types of gametes AB, Ab, aB and ab in the i-th subpopulation be P_i, Q_i, R_i and S_i ($P_i + Q_i + R_i + S_i = 1$) respectively. Then the linkage disequilibrium in the i-th population is given by

$$D_i = P_i S_i - Q_i R_i \tag{1}$$

The average value of linkage disequilibrium for all subpopulations is

$$\bar{D} \equiv E(D) = \sum_{i=1}^{k} w_i D_i, \tag{2}$$

where $w_i = N_i/N$ is the proportion of the i-th subpopulation. In terms of gamete frequencies, $\bar{D}$ may be expressed as

$$\begin{aligned} \bar{D} &= E(PS - QR) \\ &= \bar{P}\bar{S} - \bar{Q}\bar{R} + \mathrm{Cov}(P,S) - \mathrm{Cov}(Q,R), \end{aligned} \tag{3}$$

where Cov denotes the covariance. The frequencies of alleles A and B are given by $p_A = P + Q$ and $p_B = P + R$, respectively. Thus, the covariance between p_A and p_B is

$$\begin{aligned} \mathrm{Cov}(p_A, p_B) &= \mathrm{Cov}(P+Q, P+R) \\ &= \mathrm{Cov}(P, 1-Q-S) + \mathrm{Cov}(Q, P+R) \\ &= -\mathrm{Cov}(P,S) + \mathrm{Cov}(Q,R). \end{aligned} \tag{4}$$

Therefore, if we denote the linkage disequilibrium in the total population by D_T, i.e., $D_T = \bar{P}\bar{S} - \bar{Q}\bar{R}$, then we obtain the following simple formula:

$$D_T = \bar{D} + \mathrm{Cov}(p_A, p_B). \tag{5}$$

Namely, the linkage disequilibrium in the total population is the sum of the average linkage disequilibrium for subpopulations and the covariance of gene frequencies of the two segregating loci.

If the number of subpopulations is 2, then formula (5) reduces to

$$D_T = w_1 D_1 + w_2 D_2 + w_1 w_2 (p_{A_1} - p_{A_2})(p_{B_1} - p_{B_2}),$$

in which p_{A_i} and p_{B_i} are the frequencies of allele A and allele B in the i-th subpopulation, respectively. This is identical with SINNOCK and SING's (1972) formula, though they inadvertently put a negative sign in front of the last term. Note also that formula (5) can be easily extended to the case of multiple alleles if linkage disequilibrium is defined as $D_{ij} = P_{A_i B_j} - p_{A_i} p_{B_j}$, in which p_{A_i} and p_{B_j}

are the frequencies of the i-th allele at the A locus and the j-th allele at the B locus, respectively, and $P_{A_iB_j}$ is the frequency of the gamete type A_iB_j.

Formula (5) indicates that if $\mathrm{Cov}(p_A,p_B) \neq 0$, D_T may be 0 even if $\bar{D}$ is not 0, or conversely D_T may not be 0 even if $\bar{D}$ is 0.

Migration Between Two Populations

In a single random mating population with no selection, the linkage disequilibrium is reduced in every generation at a rate of r, where r is the recombination value between the two loci under consideration. Let us now study how this property is affected if the population is not a single random mating unit. For simplicity, we shall consider the simplest case where migration occurs between two populations. Let N_1 and N_2 be the sizes of populations 1 and 2, respectively. We assume that N_1 and N_2 are fairly large and remain constant for all generations. If the two populations exchange M individuals per generation, the proportions of migration in populations 1 and 2 are given by $m_1 = M/N_1$ and $m_2 = M/N_2$, respectively. Although mathematically m_1 and m_2 can take any value between 0 and 1, we confine our discussion to the case of $(m_1 + m_2)/2 \leq 0.5$, since in nature the mean of m_1 and m_2 are unlikely to be larger than 0.5. For example, $m_1 = m_2 = 0.5$ corresponds to the case of random mating in the whole population.

Let P_1,Q_1,R_1, and S_1 be the frequencies of the four gamete types AB, Ab, aB, and ab in a generation in population 1, respectively. We assume that migration occurs after mating. Then, after migration they become $(1 - m_1)P_1 + m_1P_2$, $(1 - m_1)Q_1 + m_1Q_2$, $(1 - m_1)R_1 + m_1R_2$ and $(1 - m_1)S_1 + m_1S_2$, respectively, where P_2, Q_2, R_2, and S_2 are the gamete frequencies in population 2. The gamete frequencies in the next generation, after meiosis, are given by

$$\begin{aligned} P'_1 &= (1 - m_1)(P_1 - rD_{11}) + m_1(P_2 - rD_{22}) \\ Q'_1 &= (1 - m_1)(Q_1 + rD_{11}) + m_1(Q_2 + rD_{22}) \\ R'_1 &= (1 - m_1)(R_1 + rD_{11}) + m_1(R_2 + rD_{22}) \\ S'_1 &= (1 - m_1)(S_1 - rD_{11}) + m_1(S_2 - rD_{22}) \end{aligned} \tag{6}$$

where $D_{11} = P_1S_1 - Q_1R_1$, $D_{22} = P_2S_2 - Q_2R_2$, and r is the recombination value between the A and B loci. Similar recurrence equations may be obtained also for the gamete frequencies in population 2. If random mating occurs in each population, the linkage disequilibria for populations 1 and 2 in the next generation are given by $D'_{11} = P'_1S'_1 - Q'_1R'_1$ and $D'_{22} = P'_2S'_2 - Q'_2R'_2$, respectively. Therefore, we obtain the following recurrence equations for linkage disequilibria:

$$\begin{aligned} D'_{11} &= (1 - m_1)(1 - m_1 - r)D_{11} + m_1(1 - m_1)D_{12} + m_1(m_1 - r)D_{22} \\ D'_{12} &= [(1 - m_1)(2m_2 - r) - m_2r]D_{11} + (1 - m_1 - m_2 + 2m_1m_2)D_{12} \\ &\quad + [(1 - m_2)(2m_1 - r) - m_1r]D_{22} \\ D'_{22} &= m_2(m_2 - r)D_{11} + m_2(1 - m_2)D_{12} + (1 - m_2)(1 - m_2 - r)D_{22} \end{aligned}$$

where

$$D_{12} = \begin{vmatrix} P_1 & Q_1 \\ R_2 & S_2 \end{vmatrix} + \begin{vmatrix} P_2 & Q_2 \\ R_1 & S_1 \end{vmatrix}$$

In matrix notation the above equation may be written as

$$\mathbf{D}(t+1) = \mathbf{M}\mathbf{D}(t) \tag{7}$$

where $\mathbf{D}(t)$ is the column vector of $D_{11}(t)$, $D_{12}(t)$, and $D_{22}(t)$, in which $D_{ij}(t)$ denotes the value of D_{ij} in the t-th generation, and

$$\mathbf{M} = \begin{bmatrix} (1-m_1)(1-m_1-r) & m_1(1-m_1) & m_1(m_1-r) \\ (1-m_1)(2m_2-r)-m_2r & 1-m_1-m_2+2m_1m_2 & (1-m_2)(2m_1-r)-m_1r \\ m_2(m_2-r) & m_2(1-m_2) & (1-m_2)(1-m_2-r) \end{bmatrix}$$

The eigenvalues for $\mathbf{M}$ are $\lambda_1 = 1-r$, $\lambda_2 = (1-m_1-m_2)^2$, and $\lambda_3 = (1-r)(1-m_1-m_2)$, and the corresponding eigenvectors are

$$V_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad V_2 = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \quad V_3 = \begin{bmatrix} m_1 \\ m_1-m_2 \\ -m_2 \end{bmatrix}$$

in which

$$a = m_1[(m_1+2m_2-m_1m_2-m_2^2-1)r + m_1(2-3m_1-3m_2) + m_1(m_1+m_2)^2]$$
$$b = [(r-m_1r-2m_2+2m_1m_2+m_2^2)a + m_1(r-m_1)c]/[m_1(1-m_1)]$$
$$c = m_2[(2m_1+m_2-m_1m_2-m_1^2-1)r + m_2(2-3m_1-3m_2) + m_2(m_1+m_2)^2].$$

Using the standard method of matrix algebra, it can be shown that the general formulas for $D_{11}(t)$, $D_{12}(t)$ and $D_{22}(t)$ are

$$\begin{aligned} D_{11}(t) &= l_1\lambda_1^t + al_2\lambda_2^t + m_1l_3\lambda_3^t \\ D_{12}(t) &= 2l_1\lambda_1^t + bl_2\lambda_2^t + (m_1-m_2)l_3\lambda_3^t \\ D_{22}(t) &= l_1\lambda_1^t + cl_2\lambda_2^t + m_1l_3\lambda_3^t \end{aligned} \tag{8}$$

where

$$\begin{aligned} l_1 &= C[(c-b)D_{11}(0) + aD_{12}(0) - aD_{22}(0) \\ &\quad - \frac{m_1}{m_1+m_2}\{(2c-b)D_{11}(0) + (a-c)D_{12}(0) - (2a-b)D_{22}(0)\}] \\ l_2 &= C[D_{11}(0) - D_{12}(0) + D_{22}(0)] \\ l_3 &= C[(2c-b)D_{11}(0) + (a-c)D_{12}(0) - (2a-b)D_{22}(0)]/(m_1+m_2) \end{aligned}$$

in which $C = (a-b+c)^{-1}$. Note that if the initial gamete frequencies are the same for the two populations, l_2 becomes 0.

Formula (8) shows that the asymptotic rate per generation of approach to the equilibrium values of D's, i.e., $D_{11} = 0$ and $D_{22} = 0$, is given either by $1-\lambda_1 = r$ or $1-\lambda_2 = 2(m_1+m_2) - (m_1+m_2)^2$, whichever is smaller. In a single random mating population the rate of approach to linkage equilibrium is known to be r per generation, that is, $D(t) = (1-r)^tD(0)$. Thus, if $2(m_1+m_2) - (m_1+m_2)^2 > r$, the asymptotic rate in subdivided populations is the same as that in a single random mating population. In particular, if $m_1+m_2 = 1$, then $\lambda_2 =$

$\lambda_3 = 0$, and D_{11} and D_{22} decrease at the rate of r per generation from the beginning. On the other hand, if $2(m_1 + m_2) - (m_1 + m_2)^2 < r$ and $D_{11}(0) + D_{22}(0) - D_{12}(0) \neq 0$, then the asymptotic rate is $2(m_1 + m_2) - (m_1 + m_2)^2$. Thus, the approach to equilibrium is retarded in subdivided populations. This is because if the initial gamete frequencies in the two populations are not the same, migration brings non-equilibrium gametes into one population from the other.

In a single random mating population linkage disequilibrium for neutral loci always decreases as generation proceeds, but in subdivided populations it may increase temporarily. Let us examine this problem in some detail. For simplicity, we assume that $D_{11}(0) = D_{22}(0) = 0$, $D_{12}(0) \neq 0$, and $m_1 = m_2 = m$. In this case (8) reduces to

$$\begin{aligned} D_{11}(t) = D_{22}(t) &= \frac{m(1-m)\,D_{12}(0)}{\lambda_1 - \lambda_2}(\lambda_1^t - \lambda_2^t), \\ D_{12}(t) &= \frac{2m(1-m)\,D_{12}(0)}{\lambda_1 - \lambda_2}\{\lambda_1^t + \left(1 - \frac{r}{2m(1-m)}\right)\lambda_2^t\}, \end{aligned} \tag{9}$$

where $\lambda_1 = 1 - r$ and $\lambda_2 = (1 - 2m)^2$. If $\lambda_1 = \lambda_2$, i.e. $r = 4m(1 - m)$, the above formulae degenerate. By using l'Hospital's rule, however, we obtain

$$\begin{aligned} D_{11}(t) &= m(1-m)D_{12}(0)t(1-r)^{t-1} \\ D_{12}(t) &= D_{12}(0)\{(1-r) + 2m(1-m)t\}(1-r)^{t-1} \end{aligned} \tag{10}$$

Formulae (9) and (10) indicate that the absolute values of $D_{11}(t)$ and $D_{22}(t)$ are 0 for $t = 0$ and first increase as t increases except for $m = 0.5$ and $m = 0$, and after reaching a maximum, they decline and eventually become 0. Thus, if two isolated populations start to have gene exchange by migration, the absolute value of linkage disequilibrium in a population may temporarily increase as generation proceeds even for neutral loci. The generation (t_m) at which the maximum or minimum linkage disequilibrium is attained is given by

$$\begin{aligned} t_m &= ln\frac{ln\lambda_2}{ln\lambda_1}/ln\frac{\lambda_1}{\lambda_2}, & \lambda_1 \neq \lambda_2 \\ t_m &= -1/ln(1-r), & \lambda_1 = \lambda_2 \end{aligned}$$

For example, if r is 0.02, t_m is 34.6 for $m = 0.01$, 12.3 for $m = 0.05$, and 7.3 for $m = 0.1$. Therefore, linkage disequilibrium may increase for a considerable number of generations.

In the present model no selection occurs, so that D_{11} and D_{22} eventually become 0. However, if overdominant selection operates without epistasis but with different equilibrium gene frequencies in the two populations, migration may produce non-zero equilibrium values of D_{11} and D_{22}, as will be discussed later.

DISCUSSION

In a computer simulation study FRANKLIN and LEWONTIN (1970) have suggested that if closely linked overdominant loci interact multiplicatively with respect to fitness, strong linkage disequilibria may be developed. A similar suggestion has been made by WILLS, CRENSHAW and VITALE (1970) in a study of truncation selection, though the prevalence of truncation selection in nature is questionable (CROW 1970; NEI 1971). However, a survey of linkage disequilibria among enzyme loci in Drosophila, conducted by MUKAI, METTLER and CHIGUSA (1971), has revealed no significant disequilibria. These authors have taken this as a suggestion for neutrality of enzymic genes.

Caution, however, should be made in this type of study, since natural populations are often subdivided into local mating groups. (MUKAI, METTLER and CHIGUSA [1971] apparently sampled their Drosophila genomes from a large randomly mating population.) As we have seen, if the covariance between gene frequencies is not 0, the total linkage disequilibrium in a subdivided population may be 0 even if the disequilibria in subpopulations are not 0, or conversely, it may not be 0 even if the disequilibria in all subpopulations are 0. In natural populations there is often a cline of gene frequency due to natural selection or due to the spreading of a new mutant gene from one place to the other. If two loci show such a cline of gene frequency, D_T in these loci would deviate from 0 even if there is no epistasis so that $D_i = 0$.

As an example, let us consider the frequencies of the P and Duffy blood group genes (P_1 and F_y^a alleles, respectively) in 37 villages of the Yanomama Indians, which were studied by GERSHOWITZ et al. (1972). Although there are no clear clines of gene frequencies in these villages, the P_1 and F_y^a gene frequencies are correlated with each other. Thus, we obtain $\mathrm{Cov}(P_1, F_y^a) = -0.0062$. In the present case information on D_T and $\bar{D}$ is not available. However, if we assume $\bar{D} = 0$, then $D_T = -0.0062$. The maximum possible linkage disequilibrium with the gene frequencies in this population is 0.165 (cf. LEWONTIN 1964). The ratio of -0.0062 to 0.165 is therefore -0.0375. The absolute value of this ratio is comparable to those obtained and found to be statistically significant by SINNOCK and SING (1972) for the same loci in a Michigan population.

It is worth noting that the correlation between the gene frequencies for two loci may be written as

$$r(p_A, p_B) = (D_T - \bar{D}) / \sqrt{Vp_A Vp_B} \qquad (11)$$

where Vp_A and Vp_B are the variances of gene frequencies of p_A and p_B, respectively. Thus, even if $D_T = 0$, $r(p_A, p_B)$ may not be 0.

In our study of the effect of migration on linkage disequilibrium we have assumed no selection, so that linkage disequilibrium eventually disappears. If there is epistatic selection, this is no longer true, and linkage disequilibrium may persist indefinitely. Selection complicates the mathematical model considerably, but it seems that the essential features of the effect of migration remain the same, as long as genotype fitnesses are the same for different populations.

However, if selection is different in different populations, a further complication may occur. Suppose that there is overdominant selection for two loci without epistasis in two different populations but the equilibrium gene frequencies for the two populations are different. In the absence of migration there will be no linkage disequilibrium at steady state in each population. If, however, migration occurs between the two populations, non-equilibrium gametes are introduced from one population to the other in each generation, so that permanent linkage disequilibria may be produced *without epistasis*. We believe that this is a new feature of linkage disequilibrium.

We thank DR. JAMES CROW and a referee of this paper for their valuable suggestions on the first draft.

LITERATURE CITED

CROW, J. F., 1970 Genetic loads and the cost of natural selection. pp. 128–177. In: *Mathematical Topics in Population Genetics*. Edited by K. KOJIMA. Springer Verlag, Berlin.

FRANKLIN, I. and R. C. LEWONTIN, 1970 Is the gene the unit of selection? Genetics **65**: 707–734.

GERSHOWITZ, H., M. LAYRISSE, Z. LAYRISSE, J. V. NEEL, N. CHAGNON and M. AYRES, 1972 The genetic structure of a tribal population, the Yanomama Indians. II. Eleven blood-group systems and the ABH-Le secretor traits. Ann. Hum. Genet. **35**: 261–269.

HILL, W. C. and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genetics (Der Zuchter) **38**: 226–231.

KIMURA, M., 1956 A model of genetic system which leads to closer linkage by natural selection. Evolution **10**: 278–287.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General consideration of heterotic models. Genet. **49**: 49–67.

LEWONTIN, R. C. and K. KOJIMA, 1960 The evolutionary dynamics of complex polymorphisms. Evolution **14**: 458–472.

MUKAI, T., L. E. METTLER and S. I. CHIGUSA, 1971 Linkage disequilibrium in a local population of *Drosophila melanogaster*. Proc. Nat. Acad. Sci. U.S. **68**: 1065–1069.

NEI, M., 1971 Fertility excess necessary for gene substitution in regulated populations. Genetics **68**: 169–184.

OHTA, T. and M. KIMURA, 1969a Linkage disequilibrium due to random genetic drift. Genet. Res. **13**: 47–55. ——, 1969b Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63**: 229–238.

SINNOCK, P. and C. F. SING, 1972 Analysis of multilocus genetic systems in Tecumseh, Michigan. II. Consideration of the correlation between nonalleles in gametes. Amer. J. Human Genet. **24**: 393–415.

SVED, J., 1968 The stability of linked systems with a small population size. Genetics **59**: 543–563.

WILLS, C., J. CRENSHAW and J. VITALE, 1970 A computer model allowing maintenance of large amounts of genetic variability in mendelian populations. I. Assumptions and results for large populations. Genetics **64**: 107–123.

WRIGHT, S., 1952 The genetics of quantitative variability. pp. 5–41. In: *Quantitative Inheritance*. Edited by E. C. R. REEVE and C. H. WADDINGTON. Her Majesty's Stationery Office, London.

Corresponding Editor: R. LEWONTIN