Correction. In the article “Recombinational bypass of pyrimidine dimers promoted by the recA protein of Escherichia coli” by Zvi Livneh and I. R. Lehman, which appeared in number RFII-RFIII-RFI-ss DNA 10, May 1982 of Proc. Natl. Acad. Sci. USA (79, 3171–3175), the reproduction of Fig. 2 did not do justice to the original photograph. A version at less contrast is shown here.

Fig. 2. Analysis by agarose gel electrophoresis of strand exchange between UV-irradiated phage φX174 ss DNA preparations with 0, 7, 21, and 70 dimers per molecule and φX174 linear duplex DNA. The reaction was carried out without an ATP-regenerating system. Lanes: M, DNA markers; 0–120, minutes of incubation.

Correction. In the article “Linkage disequilibrium due to random genetic drift in finite subdivided populations” by Tomoko Ohta, which appeared in number 6, March 1982 of Proc. Natl. Acad. Sci. USA (79, 1940–1944), the author requests that the following be noted. The 16th line of the Abstract should read “natural selection and limited migration, showing the latter as the main cause of the observed linkage disequilibrium.”

Linkage disequilibrium due to random genetic drift in finite subdivided populations
(subdivision of linkage disequilibrium/supergene/major histocompatibility complex polymorphism)

TOMOKO OHTA
National Institute of Genetics, Mishima, 411, Japan
Communicated by Motoo Kimura, December 11, 1981

ABSTRACT In order to clarify the mechanisms responsible for the observed linkage disequilibrium, such as that found between markers in the major histocompatibility complex of man and mouse, linkage disequilibrium between two linked loci was studied for a finite population with a subdivided structure. The infinite allele model was used. In analogy with the subdivision of the inbreeding coefficient, the linkage disequilibrium coefficient was subdivided, and various variance components of disequilibrium were defined. It was found that the disequilibrium components may get very large when migration is limited if the correlation of alleles at the two loci within a colony is taken relative to that of the entire population. In other words, with limited migration, random genetic drift of gamete types prevails in each colony. A possible test in which the variance components of disequilibrium are compared is suggested; the test discriminates between epistatic natural selection and limited migration, showing the former as the main cause of the observed linkage disequilibrium. It is pointed out that the major histocompatibility complex polymorphism is in accord with the population genetics model of multigene families that incorporate gene conversion or double unequal crossing-over between the loci in the supergene family.

There are two major mechanisms responsible for linkage disequilibrium or nonrandom association of alleles between two loci on a chromosome. They are epistatic natural selection and random genetic drift. The former increases the frequencies of favorable combinations of alleles in a population, and stable linkage disequilibrium is expected; the latter causes random fluctuation of gamete frequencies in the population and, hence, increases the variance of the linkage disequilibrium coefficient (see ref. 1 for review). Such random fluctuation would be enhanced if the population were subdivided into colonies (subpopulations) or if mating were not random in the population. Various attempts have been made to clarify such an effect (2–8). Most of these studies treat special transient cases or special types of mating schemes. Recently I have worked out a theory for treating linkage disequilibrium in a finite subdivided population when equilibrium is reached among various forces. The results of the mathematical treatments will be published elsewhere, but the analyses are summarized and will be further developed in this report. The results have an important bearing for understanding the factors responsible for the observed linkage disequilibrium, such as that found among the markers in the major histocompatibility complex (MHC) of man and mouse by Bodmer (9) and Klein (10).

MODEL
Let us consider a population (species) consisting of a finite number, n, of subpopulations (colonies). Each colony consists of N breeding individuals (effective size) and is subject to extinction with rate λ per generation, so that whenever the colony goes extinct, it is immediately replaced by a line derived from individuals of another colony in the population (11). In addition, each colony exchanges individuals with the entire population at a rate m per generation—i.e., Wright's island model (12) is assumed.

In my original analysis, the classical two-allele model with symmetric mutation rate was used. In this report, it is assumed that a mutation occurrence represents a new, not a preexisting, allele—i.e., the infinite allele model of Kimura and Crow (13) is assumed. This model is more realistic than the previous one because there are many allelic states at the molecular level. Let v be the mutation rate per locus per generation. Two loci, A and B, are assumed, and let c be the recombination fraction between them.

In this model, random drift occurs within each colony, and different chromosome types may spread in different colonies; in addition, the extinction/replacement of colonies may greatly accelerate random frequency drift in the total population. Before starting the analyses on two loci, it is useful to review the single-locus properties of this model. A necessary quantity is the probability of two randomly chosen genes being identical by descent. Let ϕi be the identity probability of two homologous genes within a single colony. Similarly, let ϕ0 be the identity probability of two homologous genes randomly chosen from different colonies. Maruyama and Kimura (11) studied transitional and equilibrium properties of ϕ0 and ϕi, considering haploid organisms. In the following, the approximate rate of change of ϕ0 and ϕi per generation by various forces are derived by assuming that the parameters, 1/(2N), m, 1/n, λ, and v, are much less than unity. Through random genetic drift within a colony, ϕi changes by the amount 1/(2N) − ϕi/(2N) in one generation. Through migration, it changes by the amount 2m(ϕ0 − ϕi). It decreases through mutation by the rate 2v. Therefore, for the change of ϕ0 in one generation, we have

$$\Delta \phi_0 = -2v(\phi_0 + m)\phi_1 + (1 - \phi_1)/(2N) + 2m\phi_2.$$  [1]

ϕ0 does not change by random drift within a colony but is influenced by the extinction/replacement process of colonies. Through extinction/replacement of colonies and through migration, ϕ0 changes by the amount 2(2m + λ)(ϕi − ϕ0)/n in one generation. This is because the fraction of the identity probability of different colonies 2(λ + m)/n, is replaced by that of one colony through these processes. ϕ0 also decreases through mutation at the rate 2v. Thus, the change of ϕ0 becomes

$$\Delta \phi_0 = -2v\phi_0 + 2(m + \lambda)(\phi_1 - \phi_0)/n.$$  [2]

At equilibrium, \(\Delta \phi_i = \Delta \phi_0 = 0\), and one gets the following identity probabilities.

<table>
<thead>
<tr>
<th>n</th>
<th>(\phi_0)</th>
<th>(\phi_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>0.19</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>0.15</td>
<td>0.37</td>
</tr>
<tr>
<td>5</td>
<td>0.13</td>
<td>0.34</td>
</tr>
<tr>
<td>6</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td>7</td>
<td>0.09</td>
<td>0.30</td>
</tr>
<tr>
<td>8</td>
<td>0.08</td>
<td>0.28</td>
</tr>
<tr>
<td>9</td>
<td>0.07</td>
<td>0.27</td>
</tr>
<tr>
<td>10</td>
<td>0.06</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Abbreviation: MHC, major histocompatibility complex.
\[
\phi_1 = \frac{1}{1 + 4N_v + 4N_{Nc}/(N_v + \lambda + m)}
\]
\[
\phi_2 = \frac{\phi_1}{1 + \nu/(\lambda + m)}
\]

where the hat (\(^\wedge\)) denotes an equilibrium value. Note that the corresponding second moments of the two-allele model are the expected values with respect to one allele, whereas the quantities \(\phi_1\) and \(\phi_2\) in Eqs. 3 and 4 are sums over all alleles in the population. Therefore, \(\phi_1\) and \(\phi_2\) may take the values between 0 and 1, whereas the corresponding moments of the two-allele model take the values between 1/4 and 1/2.

**VARIANCE COMPONENTS OF LINKAGE DISEQUILIBRIUM**

Various notations with respect to gamete and gene frequencies at the A and B loci are as follows. Let \(e_{ij}, k\) be the frequency of chromosome \(A_iB_j\) in the \(k\)th colony, and let \(x_{ik}\) and \(y_{jk}\) be the frequencies of \(A_i\) and \(B_j\), respectively, in the \(k\)th colony. Let \(\bar{x}_k\) and \(\bar{y}_k\) be their averages, so that \(\bar{x}_k = \sum_{i=1}^{n} e_{ij,k}/n\), \(\bar{y}_k = \sum_{j=1}^{n} y_{jk,k}/n\). In terms of these notations, various linkage disequilibrium coefficients may be defined.

Note that the conventional linkage disequilibrium coefficient may be regarded as the correlation of \(A_i\) and \(B_j\) of one gamete relative to that of two randomly chosen gametes from the population.

In the present case, various linkage disequilibrium coefficients may be written as follows. The disequilibrium of the \(k\)th colony is

\[
\delta_{ij,k} = e_{ij,k} - x_{ik}y_{jk,k}.
\]

The disequilibrium coefficient between the colonies or the correlation of \(A_i\) and \(B_j\) within a colony relative to that of the total population may be defined in various ways:

\[
d_{ij,k} = e_{ij,k} - \bar{x}_k\bar{y}_k
\]
\[
d_{ij,k} = x_{ik}y_{jk,k} - \bar{x}_k\bar{y}_k
\]

and

\[
d_{ij,k} = e_{ij,k} - \bar{y}_k.
\]

Finally, the disequilibrium coefficient of the entire population is

\[
D_{ij} = \bar{x}_i\bar{y}_j.
\]

Under the assumption of selective neutrality, the expected values of the above coefficients are all zero at equilibrium, and one needs to evaluate the variances of these coefficients. Under the infinite allele model, the variance of the disequilibrium coefficient is complicated and expressed by the sum over all alleles at the two loci (14). Population structure further complicates the formulation. In terms of the analogy with subdivision of inbreeding coefficients of Wright (12) (see also refs. 15 and 16), let us use the subscripts IS, ST, and IT as a way of comparing the correlation of \(A_i\) and \(B_j\). The process of the frequency change is stochastic, because of random genetic drift within a colony and of extinction/replacement of colonies, and the expectation is taken for the distribution. The symbol \(E\) denotes taking the expectation with respect to such a distribution of \(x_{ik}\), \(y_{jk}\), and \(e_{ij,k}\). The following formulas define various components.

\[
D_{IS}^2 = E\left\{\sum_{i,j} \delta_{ij,k}^2\right\} = E\left\{\sum_{i,j} (e_{ij,k} - x_{ik}y_{jk,k})^2\right\},
\]

in which \(\sum_{i,j}\) is the sum over all \(i\) and \(j\). \(D_{IS}^2\) is the expected variance of linkage disequilibrium within a colony. Because we assume that colonies are equivalent, the subscript \(k\) may be dropped after taking expectation, if one is interested in the equilibrium values.

\[
D_{ST}^2 = E\left\{\sum_{i,j} d_{ij,k}^2\right\} = E\left\{\sum_{i,j} (x_{ik}y_{jk,k} - \bar{x}_i\bar{y}_j)^2\right\},
\]

which is the variance of the correlation of genes of the two loci \((A_i, B_j)\) of different gametes of one colony relative to that of the total population. Other definitions are

\[
D_{IS}^2 = E\left\{\sum_{i,j} d_{ij,k}^2\right\} = E\left\{\sum_{i,j} (\bar{e}_{ij,k} - \bar{e}_{ij})^2\right\},
\]

which is the variance of the correlation of \(A_i\) and \(B_j\) of one gamete in a colony relative to that of the total population.

\[
D_{IT}^2 = E\left\{\sum_{i,j} d_{ij,k}^2\right\} = E\left\{\sum_{i,j} (e_{ij,k} - \bar{y}_k)^2\right\},
\]

which is the total variance of disequilibrium—i.e., the correlation of \(A_i\) and \(B_j\) of a gamete in a colony relative to that of different gametes of the total population.

Interestingly, the additivity principle, as in inbreeding coefficients, does not hold for the above components of linkage disequilibrium.

\[
D_{IT}^2 \neq D_{IS}^2 + D_{ST}^2.
\]

This is because covariance terms come into the formula. However, we have

\[
D_{IT}^2 = D_{IS}^2 + D_{ST}^2.
\]

These relationships will be examined in more detail later.

In order to obtain the above variance components, one needs all moments of \(e_{ij,k}, \bar{x}_k, \bar{y}_k, \ldots\), up to the fourth. Details of the derivation will be published elsewhere. Here, the difference of the formulas between the two-allele and infinite allele models is presented.

In the present case, the third moments such as \(E(\sum_{i,j} x_{ik}y_{jk,k})\) reduce to the second moments, \(\phi_1\) or \(\phi_2\), as noted by N. Takahata (personal communication) in his analyses on extranuclear genes, because expectation is taken over all \(i\) and \(j\). Therefore, they are not required in the present analyses. The fourth moments are expressed by the following formulas under the infinite allele model.

\[
f_1 = E\left\{\sum_{i,j} e_{ij,k}^2\right\}
\]
\[
f_2 = E\left\{\sum_{i,j} \bar{e}_{ij}^2\right\}
\]
\[
f_3 = E\left\{\sum_{i,j} \bar{e}_{ij,k} x_{ik}y_{jk,k}\right\}
\]
\[
f_4 = E\left\{\sum_{i,j} \bar{e}_{ij,k} (x_{ik}y_{jk,k})\bar{x}_i\bar{y}_j\right\}
\]

\[
f_5 = \frac{1}{2} E\left\{\sum_{i,j} \left[ (e_{ij,k}x_{ik}y_{jk,k})(\bar{y}_j + (e_{ij,k}y_{jk,k})\bar{x}_k) \right] \right\}
\]
Let f = (f1, f2, \ldots, f_{13}). The equation giving the changes in these moments in one generation, \Delta f = Af + b, requires only a slight modification from that for the two-allele model. The diagonal elements of the matrix A, a_{ii}, and the vector b of the corresponding equation in the two-allele model now take the following form under the infinite allele model, and no other changes are required. The term involving the mutation rate, \nu, of every \alpha_{ii} has to be \nu v in our model instead of \nu v as in the two-allele model.

\[
a_{ii} = -4\nu - h_{ii},
\]

where \( h_{ii} \) is the same function of \( 1/N \), \( \alpha = 2(\lambda + m)/n \), and m as in the two-allele model. The vector b becomes simpler in the infinite allele model than in the two-allele model and is expressed as follows:

\[
b = \left[ \begin{array}{cccc} \frac{1}{2N} & 0 & \phi_{1} & 0 \\ 0 & N' & 0 & \phi_{2} \\ \phi_{3} & 2N' & 0 & \phi_{4} \\ 0 & \phi_{5} & \phi_{6} & 0 \\ \phi_{7} & 0 & \phi_{8} & 0 \\ 0 & 0 & \phi_{9} & 0 \\ 0 & 0 & 0 & 0 
\end{array} \right].
\]

In terms of the fourth moments, the various components of linkage disequilibrium may be expressed by the following equations:

\[
\begin{align*}
D_{15}^2 &= f_1 + f_2 - 2f_3 & \text{[20]} \\
D_{ST}^2 &= f_6 + f_{13} - 2f_{12} & \text{[21]}
\end{align*}
\]

Table 1. Examples of variance components of linkage disequilibrium in finite subdivided populations at equilibrium

<table>
<thead>
<tr>
<th>m</th>
<th>c</th>
<th>( D_{15}^2 )</th>
<th>( D_{ST}^2 )</th>
<th>( D_{IT}^2 )</th>
<th>( D_{IS}^2 )</th>
<th>( \sigma_{15}^2 )</th>
<th>( \sigma_{ST}^2 )</th>
<th>( \phi_1 )</th>
<th>( \phi_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.00036</td>
<td>0.80314</td>
<td>0.74018</td>
<td>0.06344</td>
<td>0.80362</td>
<td>0.22614</td>
<td>0.13602</td>
<td>0.9615</td>
<td>0.3205</td>
</tr>
<tr>
<td>0.001</td>
<td>0.02205</td>
<td>0.53315</td>
<td>0.56357</td>
<td>0.01449</td>
<td>0.57806</td>
<td>0.31972</td>
<td>0.07241</td>
<td>0.7704</td>
<td>0.2734</td>
</tr>
<tr>
<td>0.005</td>
<td>0.06522</td>
<td>0.18121</td>
<td>0.28336</td>
<td>0.00407</td>
<td>0.26743</td>
<td>0.16513</td>
<td>0.06630</td>
<td>0.4581</td>
<td>0.1963</td>
</tr>
<tr>
<td>0.01</td>
<td>0.09993</td>
<td>0.08524</td>
<td>0.18285</td>
<td>0.00279</td>
<td>0.16644</td>
<td>0.13188</td>
<td>0.00399</td>
<td>0.3290</td>
<td>0.1645</td>
</tr>
<tr>
<td>0.05</td>
<td>0.03061</td>
<td>0.01213</td>
<td>0.43511</td>
<td>0.00201</td>
<td>0.04551</td>
<td>0.04395</td>
<td>0.00262</td>
<td>0.1656</td>
<td>0.1242</td>
</tr>
<tr>
<td>0.0</td>
<td>0.11038</td>
<td>0.08780</td>
<td>0.17794</td>
<td>0.05336</td>
<td>0.23129</td>
<td>0.24015</td>
<td>0.07630</td>
<td>0.3290</td>
<td>0.1645</td>
</tr>
<tr>
<td>0.001</td>
<td>0.08043</td>
<td>0.08642</td>
<td>0.18087</td>
<td>0.01234</td>
<td>0.19321</td>
<td>0.17643</td>
<td>0.01768</td>
<td>0.3290</td>
<td>0.1645</td>
</tr>
<tr>
<td>0.01</td>
<td>0.04725</td>
<td>0.08448</td>
<td>0.14703</td>
<td>0.00135</td>
<td>0.14838</td>
<td>0.10416</td>
<td>0.00194</td>
<td>0.3290</td>
<td>0.1645</td>
</tr>
<tr>
<td>0.025</td>
<td>0.02911</td>
<td>0.08394</td>
<td>0.12288</td>
<td>0.00064</td>
<td>0.12342</td>
<td>0.06843</td>
<td>0.00077</td>
<td>0.3290</td>
<td>0.1645</td>
</tr>
<tr>
<td>0.05</td>
<td>0.01775</td>
<td>0.08257</td>
<td>0.10714</td>
<td>0.00031</td>
<td>0.10745</td>
<td>0.03929</td>
<td>0.00044</td>
<td>0.3290</td>
<td>0.1645</td>
</tr>
</tbody>
</table>

Standardized variances (\( \sigma_{15}^2 \) and \( \sigma_{ST}^2 \)), the probability of gene identity at a locus within a colony (\( \phi_1 \)), and that in the total population (\( \phi_2 \)) are given. The effects of recombination rate \( c \) and migration rate \( m \) can be seen. Parameters are \( v = 10^{-}\lambda \), \( N = 100 \), \( n = 200 \), and \( \lambda = 10^{-2} \).
Fig. 1 shows the various variance components of disequilibrium as functions of Nm. It clearly shows that migration rate has a large effect on linkage disequilibrium. It is expected that when the rate of extinction/replacement of colonies (λ) becomes large, disequilibrium of the whole population may get fairly large. Details of such a prediction will be published elsewhere.

**DISCUSSION**

It is now clear that, when migration is limited, linkage disequilibrium becomes large if correlation of nonallelic genes within a colony is taken relative to that of the total population \(D_{ST}^2, D_{IS}^2, \) and \(D_{IT}^2\). In other words, limited migration is very effective in increasing differentiation of gamete types among the colonies by random genetic drift. It is often said that based on single-locus theory, when \(4Nm\) is larger than unity, differentiation is not pronounced among the colonies based on Wright’s analyses (22). The prediction is appropriate for single locus without mutation but is not so when the mutation rate is high. It deviates farther from the truth when two loci are considered, as the present results indicate.

From Tables 1 and 2, one predicts that the relationships \(D_{IS}^2 > D_{ST}^2\) and \(D_{IT}^2 > D_{IS}^2\) usually hold when migration is limited. On the other hand, if epistatic natural selection is responsible for linkage disequilibrium but not for local differentiation, one would predict \(D_{ST}^2 < D_{IS}^2\) and \(D_{IT}^2 < D_{IS}^2\) because gametes with favorable combinations of alleles would increase in every colony. Thus, the ratio \(D_{ST}^2/D_{IS}^2\) or \(D_{IT}^2/D_{IS}^2\) may be a measure for testing which of the factors (epistasis or population subdivision) is mainly responsible for the observed linkage disequilibrium.

For this purpose, data on the MHC of mouse rather than man would be more suitable. This is because rapid improvements in modern transportation may have drastically altered human migration patterns in the recent past. Thus, the present human population may be a mixture of previously separated colonies. In fact, the large disequilibrium observed in the English population (9) is likely to reflect such a history. The concept of supergenes or super-supergenes of Bodmer (9), which states that the complex is made of functionally interrelated groups of genes and is kept tightly linked because of functional requirement—i.e., epistatic natural selection—needs revision. It is more likely that the supergenes are held together because of their origin—i.e., duplication of a primordial gene.

As compared with mammalian species, *Drosophila* populations seem to have more migration among the colonies. With the exception of loci associated with inversion in chromosomes, most enzyme loci are in linkage equilibrium (23–26), and one would expect sufficient migration among the local colonies. In contrast with this, linkage disequilibrium is pronounced among enzyme loci between clones of wild populations of *Escherichia coli* (27), indicating that there is little recombination among the clones. In conclusion, epistatic natural selection seems to be much less important than some have thought, and most of the observed linkage disequilibria in animals may simply be a result of interplay between random drift and subdivided population structure.

Finally, it is worthy to mention that extremely high polymorphism at the MHC may be the result of gene conversion or double unequal crossing-over among the genes of different loci belonging to this supergene family. It has recently been reported that the genes of MHC constitute a multigene family because 30–40 crosshybridizing genomic clones are obtained, even if each marker locus is of the single copy (28). Furthermore, it now seems that, whenever homologous genes are tandemly arranged in a short chromosomal region, transfer of gene segments takes place between the loci in the course of evolution either by gene conversion or by double unequal crossing-over (29–32). Then one would expect that the transfer of gene segments takes place also among the genes of different loci belonging to the MHC and, as a consequence, concerted evolution of the supergene family occurs. This hypothesis explains an enigmatic observation that the gene identity among alleles is

![Fig. 1. Variance components of linkage disequilibrium as functions of Nm. Parameters are \(v = 10^{-6}, N = 100, n = 100, \lambda = 10^{-4}, \) and \(e = 0.01\).](image-url)
roughly 90% in terms of amino acid identity, whereas it is about 85% among genes of different loci (between HLA-A and HLA-B or H2-K and H2-D) (see ref. 33). In other words, gene homology between the different loci is only slightly lower than that among alleles. It is not difficult to choose a set of reasonable parameter values to give the above observed gene identity, based on population genetics theory of multigene families (34). The hypothesis does not contradict the present analyses because the transfer of gene segments between the different loci would have the same effect as mutations on linkage disequilibrium. Also, the hypothesis is a revised form of the original proposals of Bodmer (35) and Silver and Hood (36) in that each marker region such as H2-K or HLA-A contains a cluster of many loci (a multigene family) of which only one would be expressed.

I thank Dr. M. Kimura for his encouragement and stimulating discussions throughout the course of this work. I also thank Dr. J. F. Crow for his useful suggestions on the notation of disequilibrium components and on a possible test to discriminate between the effect of epistasis and that of limited migration. Thanks are also due to Dr. M. Kimura, Dr. N. Takahata, and Dr. K. Aoki for going over the manuscript and offering many valuable suggestions to improve the presentation. This paper is contribution no. 1400 from the National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan.