Homozygosity and Linkage Disequilibrium

Chiara Sabatti*,[1] and Neil Risch†

*Department of Human Genetics and Statistics, University of California, Los Angeles, California 90095-7088 and †Department of Genetics, Stanford University, Stanford, California 94305-5120

Manuscript received October 9, 2001
Accepted for publication January 14, 2002

ABSTRACT

We illustrate how homozygosity of haplotypes can be used to measure the level of disequilibrium between two or more markers. An excess of either homozygosity or heterozygosity signals a departure from the gametic phase equilibrium: We describe the specific form of dependence that is associated with high (low) homozygosity and derive various linkage disequilibrium measures. They feature a clear biological interpretation, can be used to construct tests, and are standardized to allow comparison across loci and populations. They are particularly advantageous to measure linkage disequilibrium between highly polymorphic markers.

TESTING for the presence of linkage disequilibrium (LD) and measuring its value are two important instruments of statistical genetics that have recently received a great deal of attention. The first studies of LD were mainly in the context of population genetics; for example, disequilibrium between markers was used to assess the age of various populations. In the last decade, instead, measures of LD have also been rediscovered as a tool for disease mapping, so that investigation has focused on measuring the disequilibrium between an unknown disease gene and a known set of markers. Indeed, the presence of linkage disequilibrium between a disease gene and a given set of markers identifies the chromosomal region spanned by the markers as a candidate location of the disease gene. Moreover, the pattern of variation of LD values along a stretch of DNA also carries information: It can be used to pinpoint the most likely location of a disease gene within a region or to reconstruct the modality of recombination. The hope of exploiting the relation between the amount of linkage disequilibrium and the recombination fraction between two loci has motivated, on the one hand, the development of a series of statistical methodologies (Hastbacka *et al.* 1992; Kaplan *et al.* 1995; Terwilliger 1995; Devlin *et al.* 1996; Xiong and Guo 1997; Graham and Thompson 1998; Lazzeroni 1998; McPeek and Strahs 1999; Service *et al.* 1999; Lam *et al.* 2000; Morris *et al.* 2000; Liu *et al.* 2001) and, on the other hand, the design of genome screens where a high number of densely spaced markers are typed in a population-like sample, to be analyzed with linkage disequilibrium techniques (Collins *et al.* 1997; Kruglyak 1998; Lonjou *et al.* 1999; Wright *et al.* 1999). As more extensive and systematic data are collected (Kidd *et al.* 1998; Huttley *et al.* 1999; Reich *et al.* 2001; Stephens *et al.* 2001), it has become apparent that levels of disequilibrium vary greatly between genomic regions and across populations. To design and to interpret LD genome screens one needs to refer to a "map" of the background levels of disequilibrium that can be expected in a given region of the genome and in a given population. To construct such a map, the researchers' attention has been directed, once again, to measure in the most effective manner the levels of disequilibrium between close-by markers. The literature on these measures is quite rich (see Devlin and Risch 1995 and Wier 1996 for reviews), but there are still open problems. In particular, there is no generally satisfactory measure of disequilibrium between two markers that have more than two alleles or between more than two markers. And yet, most of the data being collected in LD genome screens are of this form. In this work we analyze how it is possible to address this specific question using haplotype homozygosity (the probability of selecting two identical haplotypes at random from the population).

Among the numerous suggestions that are documented in the literature for testing and measuring LD, various references to homozygosity can be found. Sved (1968), Avery and Hill (1979), and Brown *et al.* (1980) proposed to use the variance of homozygosity; Ohta (1980) suggested a measure of disequilibrium that is based on the homozygosity of two loci and is analyzed by Hedrick (1987) in his review article. Morton and Simpson (1983) define kinship between loci as a homozygosity index and use it to reconstruct distances. Even though the cited literature illustrates the presence of a connection between variation in homozygosity and linkage disequilibrium, the nature of this connection has never been precisely analyzed and hence the reliability of homozygosity to test and measure disequilibrium remains

[1]*Corresponding author:* Department of Human Genetics, UCLA School of Medicine, 695 Charles E. Young Dr. S., Los Angeles, CA 90095-7088. E-mail: csabatti@mednet.ucla.edu

unclear. It is our goal to show what property of the population frequencies of the haplotypes defined by two markers is captured by homozygosity. While the focus of this article is on the definition of measures of disequilibrium calculated from the true population distribution, we also briefly consider the associated inferential problems. In particular, we show how Markov chain Monte Carlo algorithms can be used to conduct permutation tests and to measure disequilibrium on the basis of sample haplotypes.

RELATIONS BETWEEN LINKAGE DISEQUILIBRIUM AND HOMOZYGOSITY

Notation and definition of homozygosity: In the following we consider two markers A and B, respectively, with r and c possible alleles, $A_1, A_2, \ldots, A_r$ and $B_1, B_2, \ldots, B_c$. The population frequencies of the above alleles and of the haplotypes defined by these two markers are described in

$$\{\pi_{ij}\} = \begin{array}{c|cccc|c} & B_1 & B_2 & \cdots & B_c & \\ \hline A_1 & \pi_{11} & \pi_{12} & \cdots & \pi_{1c} & \pi_{1\cdot} \\ A_2 & \pi_{21} & \pi_{22} & \cdots & \pi_{2c} & \pi_{2\cdot} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ A_r & \pi_{r1} & \pi_{r2} & \cdots & \pi_{rc} & \pi_{r\cdot} \\ \hline & \pi_{\cdot 1} & \pi_{\cdot 2} & \cdots & \pi_{\cdot c} & 1 \end{array} \qquad (1)$$

where π_{ij} is the population frequency of the haplotype (A_i, B_j); $\pi_{i\cdot}$ is the population frequency of allele A_i; and $\pi_{\cdot j}$ is the population frequency of allele B_j. We indicate with $H_A(\{\pi_{ij}\}) = \sum_{i=1}^{r} \pi_{i\cdot}^2$ and $H_B(\{\pi_{ij}\}) = \sum_{j=1}^{c} \pi_{\cdot j}^2$ the homozygosities of the two markers and with $H_{AB}(\{\pi_{ij}\}) = \sum_{i,j} \pi_{ij}^2$ the haplotype homozygosity (the probability of selecting two identical haplotypes at random from the population). When there is no room for confusion, we omit the argument $(\{\pi_{ij}\})$ in the formula above. Typically, the population frequencies above are unknown and one estimates them from a random sample of haplotypes. However, for the time being we assume $\{\pi_{ij}\}$ to be known and investigate the relation between $\sum_{i,j} \pi_{ij}^2$ and linkage disequilibrium.

We note that homozygosity has been previously used to identify the location of disease genes with the strategy that goes under the name of homozygosity mapping (Smith 1953; Lander and Botstein 1987). In such cases, however, the data came from inbred families, while the measures we consider here are appropriate for a random or case-control sample from the entire population of haplotypes—indeed, related haplotypes should be excluded from the analysis.

Linkage disequilibrium: Loci A and B are said to be in gametic phase equilibrium (GPE) if $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all i, j (if the qualitative random variables A and B are independent). Linkage disequilibrium is defined as a departure from GPE. This broad definition of disequilibrium as association between A and B poses some problems. A deviation from GPE could be due to a number of population genetic phenomena such as stratification, admixture, or genetic drift. It is often impossible, on the basis of tables such as (1) alone, to determine the origin of the disequilibrium. Moreover, there is not a precise notion of distance from independence that allows one to order a set of tables. We show how homozygosity measures a specific direction of the departure from independence. The utility of this particular measure, then, depends on its genetic interpretability and its connection with the specific problem at hand.

The value of haplotype homozygosity under maximal dependency and equilibrium: The existence of a connection between haplotype homozygosity and linkage disequilibrium is easily established. The homozygosity of any given marker is higher when fewer alleles are present with a significant frequency. Indeed, in statistics, heterozygosity is known as the Gini index of diversity (see, for example, Bhargava and Uppuluri 1977a,b). Similarly, when the contingency table (1) has few cells different from zero, the value of the haplotype homozygosity is high. If we do not fix the values of the marginal distributions, this happens in the case of maximum disequilibrium: Each allele at one marker is found in combination with one and only one allele at the other marker; that is, only one cell both per row and per column is different from zero. High homozygosity is, thus, associated with high disequilibrium.

On the other hand, under linkage equilibrium, the multiplicative property of $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ translates into $H_{AB} = H_A H_B$ and the haplotype homozygosity is equal to the product of the marker homozygosities. However, this equation does not hold only in the case of linkage equilibrium. A brief consideration of a 2×2 table clarifies the issue. In a 2×2 contingency table, let $\pi_{1\cdot} = p$, $\pi_{\cdot 1} = q$, and $D = \pi_{11} - pq$. Then, we can reexpress $\{\pi_{ij}\}$ in the following form that emphasizes the existing linear constraints and the departure from independence,

$$\{\pi_{ij}\} = \begin{array}{c|cc|c} & B_1 & B_2 & \\ \hline A_1 & pq + D & p(1-q) - D & p \\ A_2 & (1-p)q - D & (1-p)(1-q) + D & 1-p \\ \hline & q & 1-q & \end{array}, \qquad (2)$$

with $\max(-pq, -(1-p)(1-q)) \le D \le \min(p(1-q), q(1-p))$. The homozygosity associated with this table is

$$\begin{aligned} H_{AB} &= (pq + D)^2 + ((1-q)p - D)^2 \\ &\quad + ((1-p)q - D)^2 + ((1-p)(1-q) + D)^2 \\ &= 4D^2 + 2D(2p-1)(2q-1) + H_A H_B. \end{aligned} \qquad (3)$$

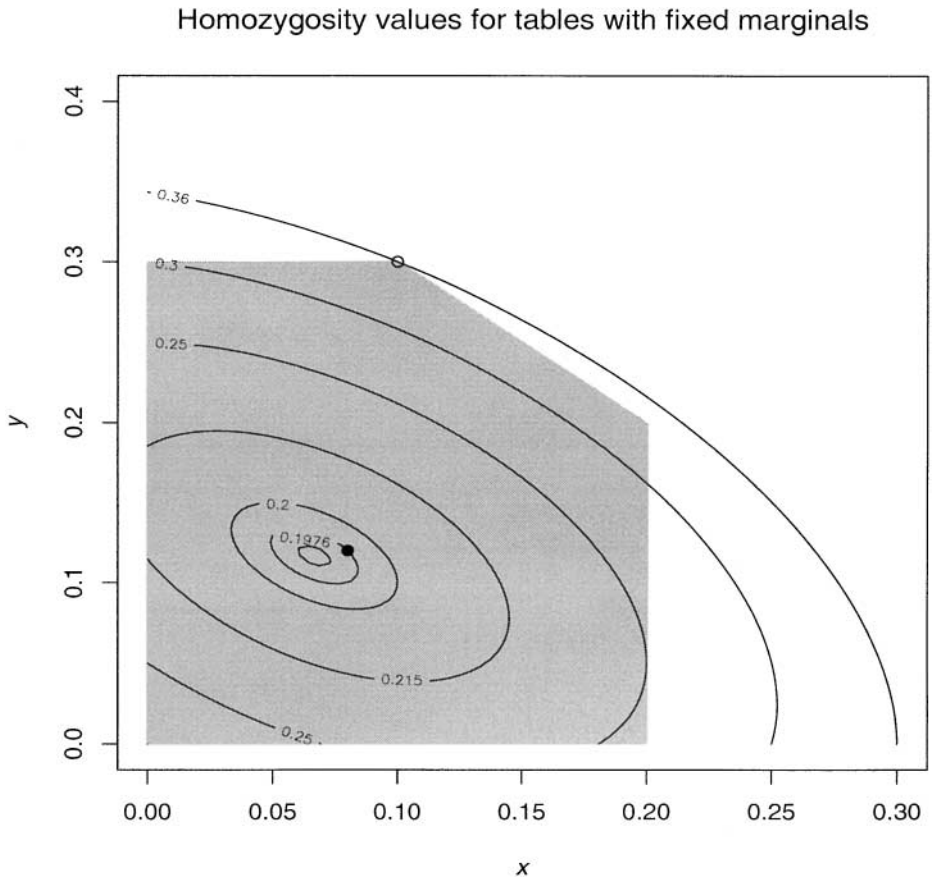

FIGURE 1.—Homozygosity values for the class of haplotype distributions described in (4). The shaded area represents the admissible tables, in the space of $x = \pi_{11}$ and $y = \pi_{21}$. The solid circle identifies the table corresponding to gametic phase equilibrium. The open circle signals the table with highest haplotype homozygosity. The ellipses are level sets of homozygosity. It is apparent that there is a set of tables that share the same homozygosity value as the independence one and that there are tables with higher heterozygosity than the independence one.

From (3), it is clear that $H_{AB} = H_A H_B$ when $D = 0$, but also when $D = (1 - 2p)(2q - 1)/2$. Indeed, the haplotype homozygosity can be smaller than the product of the marker homozygosity. An expression similar to the above can be obtained for tables of any dimension. Indeed, letting $D_{ij} = \pi_{ij} - \pi_{i\cdot}\,\pi_{\cdot j}$, one obtains

$$H_{AB} = \sum_{ij} D_{ij}^2 + 2 \sum_{ij} D_{ij} \pi_{i\cdot} \pi_{\cdot j} + H_A H_B$$

(see OHTA 1980). By extending the results in (3), one notes that for multiallelic markers the haplotype homozygosity H_{AB} can be equal to $H_A H_B$ for an unlimited number of tables. Figure 1 illustrates the relation between haplotype homozygosity and linkage disequilibrium described up to this point. We consider one biallelic marker (with allele frequencies 0.4 and 0.6) and a marker with three alleles (frequencies 0.2, 0.3, and 0.5). The space of all possible tables with these marginals can be parametrized as a function of two parameters $x = \pi_{11}$ and $y = \pi_{21}$:

$$\{\pi_{ij}\} = \begin{array}{|c|c|c|c|} \hline & B_1 & B_2 & \\ \hline A_1 & x & 0.2 - x & 0.2 \\ \hline A_2 & y & 0.3 - y & 0.3 \\ \hline A_3 & 0.4 - x - y & 0.1 + x + y & 0.5 \\ \hline & 0.4 & 0.6 & \\ \hline \end{array} . \qquad (4)$$

Requiring that $0 \le \pi_{ij} \le 1$ for all i and j is equivalent to requiring that $0 \le x \le 0.2$, $0 \le y \le 0.3$, and $x + y \le 0.4$. The space of all the possible tables that satisfy these constraints is represented in Figure 1 by the shaded area. The table of linkage equilibrium, corresponding to the values $x = 0.08$, $y = 0.12$, is represented with a solid circle. For each table in the space, we can calculate the homozygosity value. Contour levels of homozygosity as a function of x and y are depicted in Figure 1; it can be seen that there is a set of tables that have the same homozygosity level as the $\{\pi_{ij}\}$ corresponding to linkage equilibrium. It is also evident that there exist tables with homozygosity levels lower than $H_A H_B$. Another table emphasized in Figure 1 is the one that leads to the highest chi-square statistic, which in this case is also the one with highest homozygosity (corresponding to the values $x = 0.1$, $y = 0.3$). The fact that $H = H_{AB} - H_A H_B = 0$ not only in the case of independence clearly signals that H is not, strictly speaking, a measure of dependence, but rather of one particular form of association that can be zero even if the table $\{\pi_{ij}\}$ shows dependence. This is a common characteristic of measures of association. For example, the correlation coefficient between two random variables is zero not only in the case of independence, but whenever there is no linear association. Yet, it is often used as a measure of dependence, but with due caution. We clarify below the form of association measured by homozygosity.

Connection between homozygosity and recombination fraction: Much of the current interest in linkage disequilibrium between markers is due to the fact that its evolution over time can be related to the recombination fraction between the loci. Consider a simplified model where each individual has one chromosome and chromosomes of the next generation $(t + 1)$ are obtained by either sampling one from the present generation (t) and not recombining it or sampling two and recombining them. Then, for each i, j it is easily seen that

$$D_{ij}^{t+1} = \pi_{ij}^{t+1} - \pi_{i\cdot}\pi_{\cdot j} = (1 - \theta)D_{ij}^{t} = (1 - \theta)^{t+1}D_{ij}^{0}, \tag{5}$$

where θ is the recombination fraction between the two loci. This dynamic assures that $\pi_{ij} \to \pi_{i\cdot}\pi_{\cdot j}$ as $t \to \infty$. An immediate consequence is that $H^t \to 0$ as $t \to \infty$. It is of interest to monitor the behavior of this convergence. By the same reasoning used above,

$$\begin{aligned} H^{t+1} &= H_{AB}^{t+1} - H_A H_B \\ &= (1 - \theta)^2 \sum_{ij} (\pi_{ij}^t)^2 + 2(1 - \theta)\theta \sum_{ij} \pi_{i\cdot}\pi_{\cdot j}\pi_{ij}^t \\ &\quad + \theta^2 \sum_{ij} \pi_{i\cdot}^2\pi_{\cdot j}^2 - \sum_{ij} \pi_{i\cdot}^2\pi_{\cdot j}^2 \\ &= (1 - \theta)^2 H^t + 2\theta(1 - \theta) \sum_{ij} \pi_{i\cdot}\pi_{\cdot j}D_{ij}^t. \end{aligned}$$

From the last expression it is evident that $H^t = 0$ is not a sufficient condition for stability: Unless $\sum_{ij}\pi_{i\cdot}\pi_{\cdot j}D_{ij}^t = 0$, equilibrium is not reached. Then, even if there are numerous tables such that $H(\pi) = 0$, only the table corresponding to independence represents an equilibrium for the system. The differential equation describing the behavior of H^t can be further simplified recalling (5) and defining $\overline{D}^0 = \sum_{ij}\pi_{i\cdot}\pi_{\cdot j}D_{ij}^0$:

$$H^{t+1} = (1 - \theta)^2 H^t + 2\theta(1 - \theta)^{T+1}\overline{D}_0. \tag{6}$$

Then, by recursion we get

$$\begin{aligned} H^{t+1} &= (1 - \theta)^{2(t+1)}H^0 + 2\theta\overline{D}^0 \sum_{j=1}^{t+1} (1 - \theta)^{t+j} \\ &= (1 - \theta)^{2(t+1)}H^0 + 2\overline{D}^0(1 - \theta)^{t+1}(1 - (1 - \theta)^{t+1}). \end{aligned} \tag{7}$$

The evolution of $\{\pi_{ij}\}$ and H^t for a given value of θ ($\theta = 0.01$) is illustrated in Figure 2 for two different starting disequilibrium situations: $(x_0, y_0) = (0, 0.1)$ (open circle) and $(x_0', y_0') = (0.15, 0.15)$ (open square). On the left, the evolution in the space of all possible tables is emphasized: Arrows indicate the convergence path from the two initial points to the linkage equilibrium situation. It can be seen that one of the paths crosses the locus of tables with $H = 0$ once before reaching equilibrium. On the right, the values of H^t for the two systems are plotted as a function of the number of generations t: In one case the homozygosity is monotonically decreasing toward the equilibrium value, while in the other it assumes values slightly smaller than $H_A H_B$ before converging to equilibrium. Equation 7 specifies the relation between evolution of haplotype homozygosity over time and recombination fraction θ between the considered markers. Figure 3 illustrates the values of the excess of homozygosity over the equilibrium one as a function of recombination fraction for a population that is 100 generations old and two distinct initial haplotype frequencies, corresponding to the two table values (x_0, y_0) and (x_0', y_0') defined above. It is clear that for some values of H^0 and $\overline{D}^0$ the relation between homozygosity and recombination fraction is not monotonic. An obvious implication is that H should be used with caution for mapping purposes. However, Figures 2 and 3 do demonstrate monotonic behavior of H^t with both t and θ when H^t is restricted to positive values.

MEASURING DISEQUILIBRIUM WITH HOMOZYGOSITY

The preceding section illustrated how haplotype homozygosity captures a particular form of departure from equilibrium. In this section we make precise the nature of this dependence and give operative definitions of measures of disequilibrium on the basis of homozygosity. The key idea is that haplotype homozygosity measures agreement between markers; it indicates how likely it is that, sampling two haplotypes at random from a population, if they are identical at one marker they are also identical at the other one, or, vice versa, if they are different at one marker, they are also different at the other one. To make this more precise, it is useful to introduce the notion of agreement between partitions.

Agreement between the partition of haplotypes by two markers: Let S be the set of all the existing population haplotypes defined by markers A and B. Any subdivision of S into subsets S_i such that each haplotype in S belongs to exactly one of the subsets S_i is called a *partition* of S. Each of the two markers A and B identifies a partition of S by putting in the same subset haplotypes with the same allele. For example, for a population with eight haplotypes, suppose that the set of the haplotypes S is

$$\begin{aligned} h_1 &= (A_1, B_2) \\ h_2 &= (A_1, B_3) \\ h_3 &= (A_2, B_1) \\ h_4 &= (A_2, B_1) \\ h_5 &= (A_2, B_2) \\ h_6 &= (A_1, B_3) \\ h_7 &= (A_3, B_3) \\ h_8 &= (A_4, B_4). \end{aligned}$$

The partition of the haplotypes according to the first marker is $\{h_1, h_2, h_6\}$, $\{h_3, h_4, h_5\}$, $\{h_7\}$, $\{h_8\}$, while the

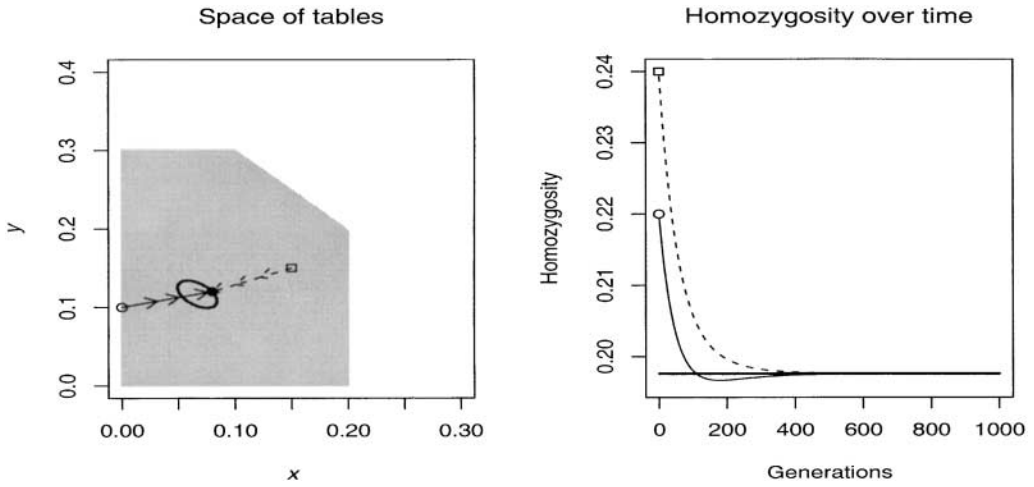

FIGURE 2.—Convergence over time to linkage equilibrium. On the left, the space of tables is as described in (4). Two disequilibrium situations are considered and identified by an open circle and an open square. The solid circle indicates the table corresponding to linkage equilibrium. The lines with arrows indicate the path to equilibrium in successive generations for the considered tables. The ellipse identifies the set of tables that have the same homozygosity value as the equilibrium one. On the right, the values of homozygosity for the two populations are depicted as a function of generations; the solid line corresponds to the evolution of the table identified by an open circle on the left and the dashed line to the evolution of the table identified with an open square. The boldface solid line identifies the equilibrium homozygosity value.

partition of the haplotypes according to the second marker is $\{h_3, h_4\}$, $\{h_1, h_5\}$, $\{h_2, h_6, h_7\}$, $\{h_8\}$. Every possible partition can be represented by a matrix with as many rows and columns as the number of haplotypes in S. For example, for a population with eight haplotypes, we can represent the partitions according to the loci A and B given above by two matrices $\mathcal{A}$ and $\mathcal{B}$. The element α_{lm} of $\mathcal{A}$ is going to be equal to 1 if haplotypes l and m are in the same group in the partition defined by A and zero otherwise. The definition of $\mathcal{B}$ is similar. Again, in our example, the matrices $\mathcal{A}$ and $\mathcal{B}$ would be

$$\mathcal{A} = \begin{array}{c|cccccccc} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 \\ h_1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ h_2 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ h_3 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ h_4 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ h_5 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ h_6 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ h_7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ h_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array},$$

$$\mathcal{B} = \begin{array}{c|cccccccc} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 \\ h_1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ h_2 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ h_3 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ h_4 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ h_5 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ h_6 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ h_7 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ h_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}.$$

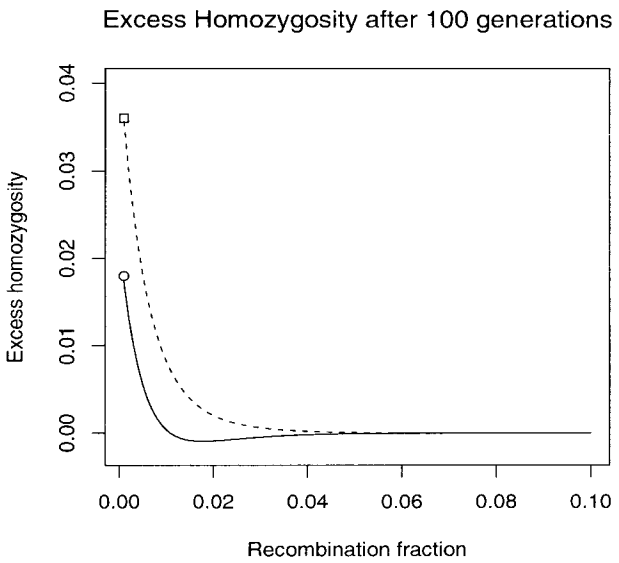

FIGURE 3.—Relation between homozygosity and recombination fraction. Letting $t = 100$ and considering the two disequilibrium situations depicted in Figure 2, the formula (7) leads to this graph where the excess of homozygosity over the equilibrium situation (on the y-axis) is depicted as a function of recombination fraction (on the x-axis).

Let us now consider the agreement between these partitions. The agreement would be perfect if each allele at marker A corresponded to one, and only one, allele at marker B: Two haplotypes are in the same group according to B if and only if they are in the same group according to A. On the contrary, the agreement is lowest if whenever A puts two haplotypes in the same group, B separates them. Between these two extremes there is the agreement that one gets just by chance. A simple

way to measure the agreement is to consider $\mathcal{A}$ and $\mathcal{B}$ as vectors (for example, reading the numbers left to right and top to bottom) and calculate the covariance between them (see HUBERT and BAKER 1978). Let N be the number of haplotypes in S; then

$$\text{Agr} = \sum_{l,m} a_{lm}b_{lm}/N^2 - \sum_{l,m} a_{lm}/N^2 \sum_{l,m} b_{lm}/N^2.$$

To see how this is related to H, note that we can describe the set S of haplotypes with a contingency table $\{\pi_{ij}\}$. In our example,

	B_1	B_2	B_3	B_4	
A_1	0	1/8	2/8	0	3/8
A_2	2/8	1/8	0	0	3/8
A_3	0	0	1/8	0	1/8
A_4	0	0	0	1/8	1/8
	2/8	2/8	3/8	1/8	

Now, it can be verified that

$$\sum_{l,m} a_{lm}b_{lm} = \sum_{i,j} \pi_{ij}^2 N^2$$
$$\sum_{l,m} a_{lm} = \sum_i \pi_{i\cdot}^2 N^2$$
$$\sum_{l,m} b_{l,m} = \sum_j \pi_{\cdot j}^2 N^2.$$

Hence H = Agr; that is, the difference between the haplotype homozygosity and the product of the marginal homozygosities measures the agreement between the partitions defined by the markers: A positive value of H (excess homozygosity) indicates more agreement than that expected by chance; a negative value of H (excess heterozygosity) indicates less agreement. Either of these excesses signifies a departure from gametic phase equilibrium (independence). Indeed, a founder effect can generate both a positive and negative value of H. Suppose, for example, that you have a population of 100 chromosomes and a disease-causing mutation appears on one of them, close to a biallelic locus. If the chromosome that experienced the mutation had at the nearby locus an allele with population frequency ≤ 0.5, there will be excess homozygosity for the disease locus-marker locus haplotype [according to formula (3) and D^2 being small, $H_{AB} - H_A H_B \approx 2D(2p-1)(2q-1) > 0$, provided $q < \frac{1}{2}$ when $p = 0.01$]. For a marker allele frequency >0.5, there will be excess heterozygosity.

References to the literature on agreement between partitions can be obtained from HUBERT and BAKER (1978) and FOWLKES and MALLOWS (1983). Incidentally we note that BLOCH and KRAEMER (1989) proposed a translation of measures of agreement into measures of dependence, which is entirely different from the present one.

Measures of disequilibrium based on *H*: Having clarified the nature of dependence captured by H, we now set to define a measure based on H that allows comparisons across tables. Generally speaking, it is useful to standardize H to obtain an index that has absolute value <1 and is equal to ± 1 in case of maximal dependence and 0 in correspondence of independence. In defining maximal dependence, recall that the degree of linkage disequilibrium between markers should be independent from the allele frequencies of each of the markers considered separately. This implies that H should be standardized using the extreme values it can take on for the given marginal distributions $\pi_{i\cdot}$, $\pi_{\cdot j}$. That is, for a table $\{\pi_{ij}\}$, we are interested in the index of dependence,

$$H'(\{\pi_{ij}\}) = \begin{cases} \dfrac{H(\{\pi_{ij}\})}{\max_{\{\tau_{ij} \mid \tau_{i\cdot}=\pi_{i\cdot}, \tau_{\cdot j}=\pi_{\cdot j}\}} H(\tau)} & \text{if } H(\{\pi_{ij}\}) \geq 0 \\ \dfrac{H(\{\pi_{ij}\})}{\min_{\{\tau_{ij} \mid \tau_{i\cdot}=\pi_{i\cdot}, \tau_{\cdot j}=\pi_{\cdot j}\}} H(\{\tau_{ij}\})} & \text{if } H(\{\pi_{ij}\}) < 0. \end{cases} \quad (8)$$

Unfortunately, the maximization involved in the definition of H' does not have a closed form solution for all c and r. In the simple case of a 2×2 table, this constrained quadratic maximization is, however, easy to solve. Recalling the parameterization of a 2×2 table given in (2), one quickly realizes that the problem is quadratic in D and that the solution is on the boundaries. The table corresponding to the maximal homozygosities will have the following value of π_{11}:

$$\pi_{11} = \begin{cases} \min(p, q) & \text{if } p > 1-p \text{ and } q > 1-q \\ p - \min(p, 1-q) & \text{if } p > 1-p \text{ and } q \leq 1-q \\ q - \min(1-p, q) & \text{if } p < 1-p \text{ and } q > 1-q \\ p + q - 1 + \min(1-p, 1-q) & \text{if } p < 1-p \text{ and } q < 1-q \end{cases}.$$

The minimal homozygosity is achieved for

$$\pi_{11} = \begin{cases} -\dfrac{p+q}{2} & \text{if } \max(-pq, -(1-p)(1-q)) \\ & \quad \leq -\dfrac{(2p-1)(2q-1)}{4} \leq \min(p(1-q), q(1-p)) \\ \max(0, p+q-1) & \text{if } -\dfrac{(2p-1)(2q-1)}{4} < \max(-pq, -(1-p)(1-q)) \\ \min(p, q) & \text{if } \min(p(1-q), q(1-p)) < -\dfrac{(2p-1)(2q-1)}{4}. \end{cases}$$

This allows us to define an index H' that takes on value 1 in correspondence of maximal homozygosity and -1 in correspondence of maximal heterozygosity. To illustrate briefly the meaning of H' and its difference with a traditional measure of disequilibrium, let us consider the following tables with identical margins:

$\{\pi_{ij}^a\}$ =

	B_1	B_2	
A_1	0.2	0	0.2
A_2	0.7	0.1	0.8
	0.9	0.1	

, $\{\pi_{ij}^b\}$ =

	B_1	B_2	
A_1	0.1	0.1	0.2
A_2	0.8	0	0.8
	0.9	0.1	

.

For 2×2 tables, LEWONTIN (1964) has popularized a measure, D', that is a standardization of the value $D = \pi_{11} - \pi_{1\cdot}\pi_{\cdot 1}$, so that D' is always <1 in absolute value, is equal to 0 in case of independence, and has positive

sign when the association is along the main diagonal of the contingency table (A_1 with B_1, A_2 with B_2). Recalling the parametrization of 2×2 tables given in (2), the definition of the measure D' is as follows

$$\mathrm{D}' = \begin{cases} \dfrac{D}{\min(p(1-q),\, q(1-p))} & \text{if } D \geq 0 \\ \dfrac{D}{\min(pq,\, (1-p)(1-q))} & \text{if } D < 0. \end{cases}$$

For the tables above, $D'(\{\pi_{ij}^{a}\}) = 1$ and $D'(\{\pi_{ij}^{b}\}) = -1$, while $H'(\{\pi_{ij}^{a}\}) = -1$ and $H'(\{\pi_{ij}^{b}\}) = 1$. The sign of D' depends on the order of the rows and columns of the tables; when there is not a natural order for the outcomes of variables A and B, this seems a rather arbitrary decision. In contrast to this, the sign of H' indicates excessive homozygosity or heterozygosity and is independent from row or column order.

For generic $c > 2$ and $r > 2$, in the absence of an exact solution of the maximization in (8), one can bound the denominator in the definition of H', obtaining an index that will always have absolute value <1 and may attain value 1 only for some particular marginal distributions. There are multiple ways of obtaining such bounds, by considering the following table:

	Homoz. at B	Heter. at B	
Homoz. at A	H_{AB}	$H_A - H_{AB}$	H_A
Heter. at A	$H_B - H_{AB}$	$1 - H_A - H_B + H_{AB}$	$1 - H_A$.
	H_B	$1 - H_B$	

(9)

Using the same reasoning that is behind the construction of the common measure D', we can define

$$H^* = \begin{cases} \dfrac{H}{\min(H_A(1-H_B),\, H_B(1-H_A))} & \text{if } H \geq 0 \\ \dfrac{H}{\min(H_A H_B,\, (1-H_A)(1-H_B))} & \text{if } H < 0. \end{cases}$$

If one wants to use the same standardization for positive and negative values of H, one can use $\overline{H}^* = H/[\max(\min(H_A H_B,\ (1-H_A)(1-H_B)),\ \min(H_A (1-H_B),\, H_B(1-H_A)))]$. Furthermore, by equating, at each marker, homozygosity with the numeric value 1 and heterozygosity with the numeric 0, one can get an index that is the analog of the correlation coefficient:

$$HR = \frac{H}{\sqrt{H_A(1-H_A)\, H_B(1-H_B)}},$$

which will attain the maximal values 1 and -1 for an even more restricted set of marginals.

Note that once we decide to restrict our attention to table (9), any measure of dependence defined on it will give an indication of how much H_{AB} differs from $H_A H_B$. In particular, one may choose to look at the odds ratio

$$\Omega = \frac{H_{AB}(1 - H_A - H_B + H_{AB})}{(H_A - H_{AB})(H_B - H_{AB})}$$

and at its standardized version $(\Omega - 1)/(\Omega + 1)$. We decided to focus on the rescaling of H for ease of interpretation.

The notion of homozygosity can be applied to haplotypes that contain more than one marker. Consider, for example, the case of three loci. Then, H naturally generalizes to $\sum_{ijk} \pi_{ijk}^2 - \sum_i \pi_{i\cdot\cdot}^2 \sum_j \pi_{\cdot j\cdot}^2 \sum_k \pi_{\cdot\cdot k}^2$. The maximization (minimization) of H given the marginal distributions, however, is computationally even more demanding. Nonetheless, it is possible to define an index that is appropriate when the haplotype homozygosity is greater than the product of the individual marker homozygosities ($H > 0$):

$$H^m = \frac{\sum_{ijk} \pi_{ijk}^2 - \sum_i \pi_{i\cdot}\pi_{\cdot j\cdot}^2 \pi_{\cdot\cdot k}^2}{\min\{\sum_i \pi_{i\cdot}^2, \sum_j \pi_{\cdot j\cdot}^2, \sum_k \pi_{\cdot\cdot k}^2\} - \sum_i \pi_{i\cdot}^2 \sum_j \pi_{\cdot j\cdot}^2 \sum_k \pi_{\cdot\cdot k}^2}.$$

This index is based on the observation that haplotype homozygosity necessarily has to be smaller than each marker homozygosity. We illustrate its application with one example.

SAMPLE-BASED MEASURE OF DEPENDENCE

Estimating *H* from sample frequencies: In contrast to what has been assumed thus far, the matrix $\{\pi_{ij}\}$ of the true haplotype frequencies for loci A and B is unknown. Linkage disequilibrium between the markers must then be estimated from a sample of haplotypes of size n, leading to the counts represented in the following matrix:

$$\{n_{ij}\} = \begin{array}{c|ccccc} & B_1 & B_2 & \dots & B_c & \\ \hline A_1 & n_{11} & n_{12} & \dots & n_{1c} & n_{1\cdot} \\ A_2 & n_{21} & n_{22} & \dots & n_{2c} & n_{2\cdot} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ A_r & n_{r1} & n_{r2} & \dots & n_{rc} & n_{r\cdot} \\ \hline & n_{\cdot 1} & n_{\cdot 2} & \dots & n_{\cdot c} & n \end{array}.$$

The measures described in the previous section are applicable to analysis of sample data using the "plug-in" principle, that is, substituting for the theoretical quantities their sample analogs. Hence, instead of $\pi_{i\cdot}$, one uses $n_{i\cdot}/n$, etc. It is worth noting that homozygosity can be estimated from sample haplotypes in two ways, to which we refer as direct count and maximum likelihood. For marker A, homozygosity is estimated by direct count as

$$\hat{H}_A^{\text{count}} = \frac{\text{No. homoz. genotypes}}{\text{No. genotypes}}$$

or, assuming Hardy-Weinberg (HW) equilibrium, by maximum likelihood as

$$\hat{\pi}_{i\cdot} = n_{i\cdot}/n \quad \forall i$$
$$\hat{H}_A = \sum_i \hat{\pi}_{i\cdot}^2,$$

the latter method being more efficient when HW holds. Similarly, the joint homozygosity can be estimated in two ways:

$$\hat{H}_{AB}^{\text{mle}} = \sum_{ij} \hat{\pi}_{ij}^2 \tag{10}$$

$$\hat{H}_{AB}^{\text{count}} = \frac{\text{No. homoz. genotypes}}{\text{No. genotypes}}. \tag{11}$$

To evaluate the first of these estimators, one has to estimate π_{ij} by its sample counterpart n_{ij}/n; this is immediate whenever the phase of haplotypes is known. In such cases, $\hat{H}_{AB}^{\text{mle}}$ or its unbiased version $(n/(n-1)\hat{H}_{AB}^{\text{mle}} - 1/n)$ is preferable to $\hat{H}_{AB}^{\text{count}}$, as it will have a smaller variance; It is effectively the expected value of $\hat{H}_{AB}^{\text{count}}$ given the sufficient statistics for this model (see Lehman 1983). The expressions for the variances follow (see Bhargava and Uppuluri 1977b):

$$\text{Var}(\hat{H}_{AB}^{\text{mle}}) = \frac{2((2n-4)\sum_{ij}\pi_{ij}^3 + (3-2n)H_{AB}^2 + H_{AB})}{n(n-1)}$$

$$\text{Var}(H_{AB}^{\text{count}}) = \frac{2H_{AB}(1-H_{AB})}{n}.$$

However, when the phase of the genotypes is not available, the count estimator (11) becomes a handy alternative. Note that to ensure that the estimates of the indices take on values between −1 and 1, one should use the same estimation procedure for the haplotype and marker homozygosities.

Testing for linkage disequilibrium and sample size effects: We have outlined how the plug-in principle can be used to obtain measures of disequilibrium on the basis of H from sample data. However, analyzing a random sample, one has to evaluate the possibility that the observed counts—with their associated disequilibrium—are generated by a table $\{\pi_{ij}\}$ characterized by independence. In other words, prior to measuring disequilibrium, one should conduct a test to assess whether the hypothesis of GPE can or cannot be rejected. It is possible to use homozygosity to test for GPE; we do not intend to propose the following procedures as an alternative to the numerous tests already studied in the literature, but rather consider them for completeness. It is easy to construct asymptotic tests:

1. The statistic

$$TH_1 = \frac{\hat{H}_{AB}^{\text{mle}} - \hat{H}_A\hat{H}_B}{\sqrt{\text{Var}(\hat{H}_{AB}^{\text{mle}})}}$$

has, under independence, an approximate $N(0, 1)$ distribution for $n \to \infty$ and leads to a Gaussian test.

2. From the 2×2 table of observed haplotype homozygosity

	Homoz. at B	Heter. at B	
Homoz. at A	$\hat{H}_{AB}$	$\hat{H}_A - \hat{H}_{AB}$	$\hat{H}_A$
Heter. at A	$\hat{H}_B - \hat{H}_{AB}$	$1 - \hat{H}_A - \hat{H}_B + \hat{H}_{AB}$	$1 - \hat{H}_A$,
	$\hat{H}_B$	$1 - \hat{H}_B$	

one can obtain a χ^2 test—again assuming $n \to \infty$ and a sizeable number of observations per cell.

We do not present details of the power of these two tests, but do note that that their power can be zero for those alternatives to linkage equilibrium that give the same joint homozygosity as independence (see, for example, Figure 2). Hence, technically, the tests above are useful only if they result in a rejection of the null hypothesis. We also note that the second test is particularly practical in the case of unphased data: It can be evaluated directly on the sample data without requiring phasing.

The tests outlined above are based on asymptotic approximations; however, the assumption of $n \to \infty$ sometimes represents a serious limitation. This can be overcome with exact permutation tests that are based on the statistic $H(\{n_{ij}/n\})$. In this context, one is interested in considering all the possible tables m_{ij} with the same marginal counts as the observed $n_{i\cdot}$, $n_{\cdot j}$ and evaluating the probability of the set of these tables that leads to an excess of homozygosity greater than or equal to the observed $H(\{n_{ij}/n\})$. Figure 4 illustrates the space of all tables $\{m_{ij}\}$ with $n = \Sigma_{ij}m_{ij} = 20$ and marginal relative frequencies as in (4). The table corresponding to independence and the one with highest homozygosity excess are identified. With regard to the probability with which each table is observed under independence, it is well known that $\{m_{ij}\}$ has a Fisher-Yates (FY) distribution. The probability

$$\Pr(|H(\{m_{ij}/n\})| \geq |H(\{n_{ij}/n\})|) \quad \text{where } m_{ij} \sim \text{FY}(n_{i\cdot}, n_{\cdot j}) \tag{12}$$

represents the achieved significance level (P value) of an exact permutation test. It is possible to evaluate (12) either by direct computation (as in the algorithm described in Mehta and Patel 1983) or with a Markov chain Monte Carlo (MCMC) procedure as described in Lazzeroni and Lange (1997). We draw attention in particular to the use of MCMC samples, as they represent the only method effectively applicable for multidimensional contingency tables with highly polymorphic markers. A MCMC is used to obtain a sample of contingency tables with distribution $\text{FY}(n_{i\cdot}, n_{\cdot j})$. The percentage of tables $\{m_{ij}^s\}$ in the sample such that $|H(\{m_{ij}^s/m\})| \geq |H(\{n_{ij}/n\})|$ is taken as an estimate of the exact P value (12). Lazzeroni and Lange (1997) describe how to obtain a sample $\{m_{ij}^s\}$ with the appropriate Fisher-Yates distribution. Diaconis and Sturmfels (1998) give an-

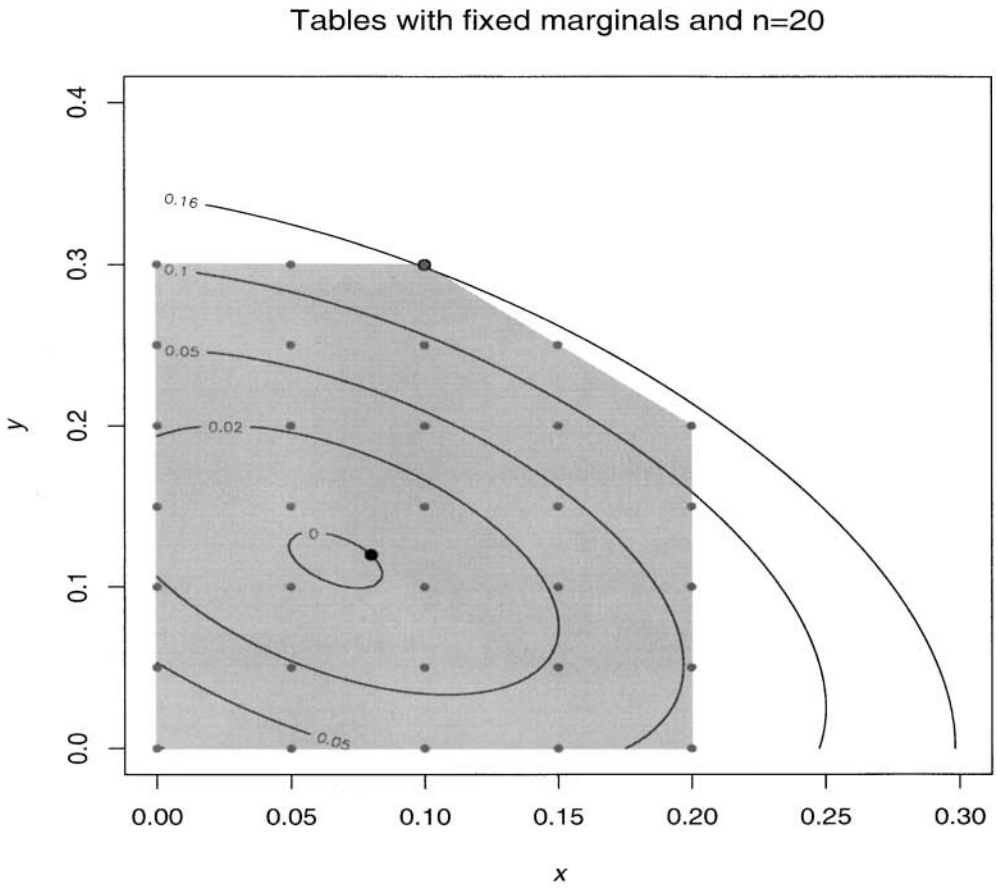

FIGURE 4.—Space of all possible $\{m_{ij}\}$ with marginal relative frequencies $m_{i\cdot}/m$ and $m_{\cdot j}/m$ as in (4) and total number of haplotypes $m = 20$. Tables are identified by bullets. The shaded area represents the space of all probability distributions $\{\pi_{ij}\}$ with the same marginals. The solid circle indicates the table corresponding to independence and the bullet with darker perimeter identifies the table with highest homozygosity. The ellipses are level sets of absolute deviation of the haplotype's homozygosity from its value under independence.

other MCMC algorithm that can be used for this purpose. The chain that these authors propose is, however, more directly applicable to the evaluation of another quantity that provides significant information on the amount of disequilibrium in the observed table. Recall that the maximization problem required in the definition of H' (8) does not have a closed-form solution. When dealing with haplotype counts, one can consider the following corresponding discrete problem:

$$\max_{\{m_{in}\}:m_{i\cdot}=n_{i\cdot},m_{\cdot j}=n_{\cdot j}} H(\{m_{ij}/m\}).$$

As n, c, r increase, this problem also becomes computationally difficult, but its solution can be approximated with a MCMC algorithm. In particular, the chain described by DIACONIS and STURMFELS (1998) leads to a sample of tables $\{m^s_{ij}\}$ with uniform distribution among the tables with fixed marginal counts $n_{i\cdot}$ and $n_{\cdot j}$ (that is, uniform on the space of tables described in Figure 4). A sample-dependent version of H' can then be evaluated as

$$H'_s(\{n_{ij}/n\}) = \begin{cases} \dfrac{H(\{n_{ij}/n\})}{\max\limits_{\{m^s_{ij}\}|s\in\text{sample}} H(\{m^s_{ij}/m\})} & \text{if } H(\{n_{ij}/n\}) \geq 0 \\ -\dfrac{H(\{n_{ij}/n\})}{\min\limits_{\{m^s_{ij}\}|s\in\text{sample}} H(\{m^s_{ij}/m\})} & \text{if } H(\{n_{ij}/n\}) < 0. \end{cases}$$

EXAMPLES

We now consider two datasets previously published in the literature for which measuring disequilibrium is particularly interesting; one because of implications regarding the presence of recombination in mitochondria and the second regarding the history of populations. The first dataset consists of biallelic markers: We evaluate the sample analog of H', substituting n_{ij}/n for π_{ij}. In the second dataset, four different markers are considered at the same time to obtain a "global" measure of disequilibrium. We evaluate an empirical version of H^m, where $\sum_{ijk}\pi^2_{ijk}$ is substituted by the direct count of haplotype homozygosity.

Example 1. Recombination in mitochondria: We consider here a dataset that has recently been used to provide evidence for the presence of recombination in mitochondria (AWADALLA *et al.* 1999). It is particularly interesting since the conclusion of the analysis depends critically on which measure of disequilibrium is used: It represents, then, a clear example of the need for reliable measures of disequilibrium. The data come from the analysis of (I) six sites (7025, 10,394, 12,308, 13,366, 15,606, 15,925) in 86 Swedish and Finnish individuals; (II) seven sites (1715, 5176, 7933, 8391, 10,394, 10,397, 13,262) in 167 Siberians; and (III) five sites (663, 5176, 10,394, 10,397, 13,262) in 153 Native Americans. Detailed description of these sites and samples can be

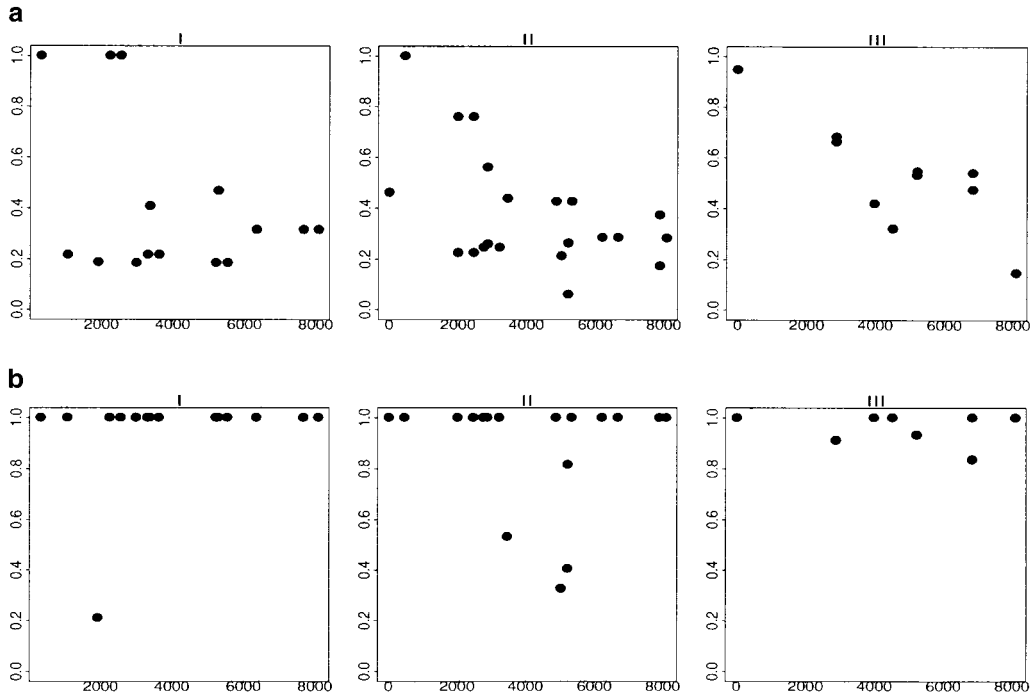

FIGURE 5.—The pattern of linkage disequilibrium values in the datasets considered by AWADALLA *et al.* (1999) using (a) R^2 and (b) $|D'|$. On the *x*-axis, distances between markers in number of basepairs are shown. On the *y*-axis, the measured disequilibrium values are shown.

found in the original articles cited by AWADALLA *et al.* (1999). Every possible pairing of the sites has been considered and the amount of disequilibrium measured between them has been plotted against their distance apart.

The measure of disequilibrium used in AWADALLA *et al.* (1999) is R^2: $(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})^2/\pi_{1.}\pi_{2.}\pi_{.1}\pi_{.2}$. Figure 5a reproduces the article's findings: The level of disequilibrium decreases as the distance between the markers increases, as to be expected in a system with recombination (we plotted $|R|$ rather than R^2 to ease the comparison with D'). Figure 5b illustrates the effect of using $|D'|$ rather then $|R|$ as a measure of disequilibrium: The mentioned effect completely disappears. The difference between R and D' relies substantially in the standardization of the measures: While in D' the measure is standardized so that the values -1 and 1 are achievable for any set of marginals, in R the extremes 1 and -1 are attainable in theory only. The graph in Figure 5b would seem to suggest that the effect noted by Awadalla *et al.* is due exclusively to the variation in marginal frequencies rather than to disequilibrium. However, there is a sample-size effect associated with D' that has to be considered in interpreting Figure 5b. As soon as one of the cells of a 2×2 contingency table is empty, the absolute value of D' is equal to one. When the marginal allele frequencies are such that the probability associated with that cell is very small under independence, and the sample size is small, there is a risk of evaluating as complete disequilibrium what is really quite close to independence. It is of interest, then, to analyze the datasets with other measures of disequilibrium that, differently from R^2, take into account marginal distributions but also, differently from D', do not inflate disequilibrium for small sample sizes. H' is the ideal candidate based on homozygosity; Figure 6 shows the results of H' to the datasets in question. It is clear that the effect observed by AWADALLA *et al.* (1999) disappears with an appropriate consideration of the marginal allele frequencies.

Example 2. Variation of disequilibrium across populations: According to the "out of Africa" hypothesis, there was a single migration of modern *Homo sapiens* out of Africa and an additional loss of variation as that initial non-African founder population grew and expanded to the East and later into the Americas. Estimating the values of linkage disequilibrium in various populations can help corroborate this hypothesis. To this purpose, four sites [three single nucleotide polymorphism (SNP) and one short tandem repeat polymorphism (STRP)] have been studied at the DRD2 locus on chromosome 11q by KIDD *et al.* (1998). The physical map for this region is SNP1–4.7 kb–SNP2–1.4 kb–STRP–19.3 kb–SNP4; thus a total of 25.4 kb is spanned by the four sites. Data from 28 populations covering five continents and 1324 subjects have been generated and analyzed to determine the overall pattern of disequilibrium in this chromosomal segment and how it varies across populations. We have reanalyzed the data using a global measure of disequilibrium defined above (H^m) on the basis of haplotype homozygosity for the four sites and obtained the results presented in Table 1. The table shows a clear pattern of increasing LD moving from

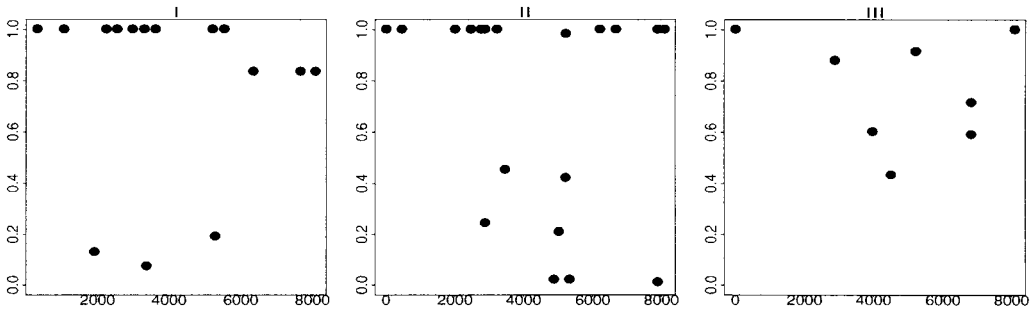

FIGURE 6.—The pattern of linkage disequilibrium values in the datasets considered by AWADALLA *et al.* (1999) using H'. On the *x*-axis, distances between markers in number of basepairs are shown. On the *y*-axis, the measured disequilibrium values are shown.

African to European/Western Asian to Eastern Asian and Amerinds, which is consistent with the out-of-Africa hypothesis. One can note an aberrantly high value of H^m for Ethiopians. Examination of the haplotype frequencies for this population reveals a pattern of nearly complete LD. Although we have included this as an African population, it is actually intermediate between Africans and Europeans/Western Asians, and could reasonably also be included in the latter group. Also, this population has the smallest sample size ($n = 32$), possibly leading to extreme variability. To address the significance of geographic origins, we have calculated the average variance within continent *vs.* variance between continent means. The within-continent variance is 0.0186

TABLE 1

Linkage disequilibrium values across populations at DRD2 (data from KIDD *et al.* 1998)

Continent	Ethnicity	H^m by ethnicity	Mean H^m	Median H^m	Variance of H^m
Africa			0.433	0.42	0.0209
	Sekele San	0.37			0.0058 (without Ethiopians)
	Central San	0.43			
	Northern Sotho	0.37			
	Tsonga	0.47			
	Biaka	0.42			
	Mbuti	0.25			
	Ethiopians	0.72			
Europe			0.606	0.63	0.0110
Western Asia	Yemenites	0.49			
	Druze	0.63			
	Adygei	0.74			
	Danes	0.66			
	Finns	0.51			
Eastern Asia			0.794	0.76	0.0104
	Han (San Francisco)	0.78			
	Han (Taiwan)	0.90			
	Koreans	0.74			
	Japanese	0.65			
	Ami	0.73			
	Atayal	0.74			
	Cambodians	0.96			
	Yakut	0.85			
Melanesia	Nasioi	0.67	0.67		
Americas			0.801	0.86	0.03214
	Cheyenne	0.86			
	Jemez Pueblo	0.92			
	Pima	0.99			
	Maya	0.49			
	Ticuna	0.69			
	Rondonia Surui	0.71			
	Karitiana	0.95			

(or 0.0149 leaving out the Ethiopians) *vs.* 0.0308 between continents. The ratio (between *vs.* within) is 1.66, or 2.07 omitting the Ethiopians.

DISCUSSION

We have discussed the use of haplotype homozygosity to measure linkage disequilibrium or, equivalently, of $\sum_{ij} \pi_{ij}^2$ to measure the amount of dependency in a contingency table $\{\pi\}$. The statistical literature contains references to this index from two different perspectives: as an index of agreement between partitions (see Hubert and Baker 1978) and as an index of diversity of the distribution $\{\pi\}$ (see Bhargava and Uppuluri 1977a). As we illustrated, both points of view provide a statistical interpretation of the relation between homozygosity and linkage disequilibrium. What remains to be discussed is the relevance for genetic purposes of the direction of disequilibrium measured by homozygosity; this will require further examination. We limit ourselves to consider the four properties that a measure of disequilibrium should have according to Hedrick (1987): (1) a simple biological interpretation (obviously satisfied for homozygosity), (2) an available statistical test (we showed how to construct it), (3) a direct relationship to evolutionary factors as recombination, and (4) a standardization that allows comparison across loci and populations (we illustrated the available standardizations and their limits). Point (3) is particularly relevant when one wants to use disequilibrium measures to identify the location of a disease gene (see the review by Devlin and Risch 1995). We have seen that, unfortunately, the relation between homozygosity and recombination fraction is not always direct, although it is so for excess homozygosity. The fact that homozygosity is defined independently of the number of alleles per locus makes H particularly suitable to measure LD between highly polymorphic markers. As the most recent LD-based genome screens have brought to general attention, it is important to collect information on the expected pattern of disequilibrium in different regions of the genome and in different populations. Kidd *et al.* (1998), Huttley *et al.* (1999), Reich *et al.* (2001), and Stephens *et al.* (2001) represent a step in this direction. The LD measures used in these works are either the P value of a test of hypothesis (a solution acceptable in their case, but not robust to sample size fluctuations) or D', which applies only to biallelic markers. The definition of robust measures of disequilibrium that can be used for successful comparison is crucial to this goal and we believe that measures based on homozygosity can play a significant role.

The measures we have defined on the basis of haplotype homozygosity are particularly suited to assessing linkage disequilibrium of multiple sites (such as SNPs) and multiallelic systems (such as STRPs), on the basis of randomly ascertained samples. The problem of localizing a disease gene among a group of closely linked markers usually entails nonrandom sampling, where disease allele-bearing chromosomes are oversampled (Devlin and Risch 1995). The measures we have described are not robust to such nonrandom sampling. For this particular application of linkage disequilibrium analysis, many different approaches, either analyzing one marker locus at a time (Hastbacka *et al.* 1992; Kaplan *et al.* 1995; Terwilliger 1995; Devlin *et al.* 1996; Xiong and Guo 1997; Graham and Thompson 1998; Lazzeroni 1998) or analyzing full multilocus haplotypes (McPeek and Strahs 1999; Service *et al.* 1999; Lam *et al.* 2000; Morris *et al.* 2000; Liu *et al.* 2001), have been described.

Chiara Sabatti was supported by the Nancy Pritzker Foundation. Neil Risch was supported in part by National Institutes of Health grant GM057672.

LITERATURE CITED

Avery, P., and W. Hill, 1979 Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. Genetics **91:** 817–844.

Awadalla, P., A. Eyre-Walker and J. Maynard-Smith, 1999 Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science **286:** 2524–2525.

Bhargava, T., and V. Uppuluri, 1977a An axiomatic derivation of the Gini's index of diversity with applications. Metron **33:** 41–53.

Bhargava, T., and V. Uppuluri, 1977b Sampling distribution of Gini's index of diversity. Appl. Math. Comput. **3:** 1–24.

Bloch, D., and H. Kraemer, 1989 2×2 kappa coefficients: measures of agreement or association. Biometrics **45:** 269–287.

Brown, A., M. Feldman and E. Nevo, 1980 Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics **96:** 523–536.

Collins, F., M. Guyer and A. Chakravarti, 1997 Variations on a theme: cataloging human DNA sequence variation. Science **278:** 1580–1581.

Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311–322.

Devlin, B., N. Risch and K. Roeder, 1996 Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics **29:** 311–316.

Diaconis, P., and B. Sturmfels, 1998 Algebraic algorithms for sampling from conditional distributions. Ann. Stat. **26:** 363–397.

Fowlkes, E., and C. Mallows, 1983 A method for comparing two hierarchical clusterings. J. Am. Statist. Assoc. **78:** 553–569.

Graham, J., and E. Thompson, 1998 Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am. J. Hum. Genet. **63:** 1517–1530.

Hastbacka, J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. **2:** 204–211.

Hedrick, P., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

Hubert, L., and F. Baker, 1978 Evaluating the conformity of sociometric measurements. Psychometrika **43:** 31–41.

Huttley, G., M. Smith, M. Carrington and S. O'Brien, 1999 A scan for linkage disequilibrium across the human genome. Genetics **152:** 1711–1722.

Kaplan, N., W. Hill and B. Weir, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. **56:** 18–32.

Kidd, K., B. Morar, C. M. Castiglione, H. Zhao and A. J. Pakstis, 1998 A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum. Genet. **103:** 211–227.

Kruglyak, L., 1998 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lam, J., K. Roeder and B. Devlin, 2000 Haplotype fine mapping by evolutionary trees. Am. J. Hum. Genet. **66:** 659–673.

Lander, E., and D. Botstein, 1987 Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science **236:** 1567–1570.

Lazzeroni, L., 1998 Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am. J. Hum. Genet. **62:** 159–170.

Lazzeroni, L., and K. Lange, 1997 Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. Ann. Stat. **25:** 138–168.

Lehman, E., 1983 *Theory of Point Estimation.* John Wiley & Sons, New York.

Lewontin, R., 1964 The interaction of selection and linkage. I. General considerations: heterotic models. Genetics **49:** 49–67.

Liu, J., C. Sabatti, J. Teng, B. Keats and N. Risch, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. **11:** 1716–1724.

Lonjou, C., A. Collins and N. Morton, 1999 Allelic association between marker loci. Proc. Natl. Acad. Sci. USA **96:** 1621–1626.

McPeek, M., and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am. J. Hum. Genet. **65:** 858–875.

Mehta, C., and N. Patel, 1983 A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **78:** 427–434.

Morris, A., J. Whittaker and D. Balding, 2000 Bayesian fine-scale mapping of disease loci by hidden Markov models. Am. J. Hum. Genet. **67:** 155–169.

Morton, N., and S. Simpson, 1983 Kinship mapping of multilocus systems. Hum. Genet. **64:** 103–104.

Ohta, T., 1980 Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. Genet. Res. **36:** 181–197.

Reich, D., M. Cargill, S. Bolk, J. Ireland, P. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Service, S., D. Temple-Lang, N. Freimer and L. Sandkuijl, 1999 Linkage disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am. J. Hum. Genet. **64:** 1728–1738.

Smith, C., 1953 The detection of linkage in human genetics. J. R. Stat. Soc. B **15:** 153–184.

Stephens, J., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

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

Terwilliger, J., 1995 A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am. J. Hum. Genet. **56:** 777–787.

Xiong, M., and S. Guo, 1997 Fine-scale genetic mapping based on linkage disequilibrium: theory and application. Am. J. Hum. Genet. **60:** 1513–1531.

Weir, B., 1996 *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

Wright, A., A. Carothers and M. Pirastu, 1999 Population choice in mapping genes for complex diseases. Nat. Genet. **23:** 397–404.

Communicating editor: G. A. Churchill