Decomposing Multilocus Linkage Disequilibrium

Root Gorelick[1] and Manfred D. Laubichler

Department of Biology, Arizona State University, Tempe, Arizona 85287-1501

Manuscript received November 26, 2003
Accepted for publication December 18, 2003

ABSTRACT

We present a mathematically precise formulation of total linkage disequilibrium between multiple loci as the deviation from probabilistic independence and provide explicit formulas for all higher-order terms of linkage disequilibrium, thereby combining J. Dausset *et al.*'s 1978 definition of linkage disequilibrium with H. Geiringer's 1944 approach. We recursively decompose higher-order linkage disequilibrium terms into lower-order ones. Our greatest simplification comes from defining linkage disequilibrium at a single locus as allele frequency at that locus. At each level, decomposition of linkage disequilibrium is mathematically equivalent to number theoretic compositions of positive integers; *i.e.*, we have converted a genetic decomposition into a mathematical decomposition.

A precise measurement of linkage disequilibrium is required for studying virtually any phenomenon in multilocus population genetics. This is especially true for explicit multilocus models that investigate the contributions of physiological epistasis to additive genetic variance (Cheverud and Routman 1995; Wagner *et al.* 1998; Wagner and Laubichler 2000). Linkage disequilibrium is usually defined as the deviation from probabilistic independence between alleles at two different loci. This deviation from independence can have different causes, such as a lack of independent segregation or recombination, or any number of other evolutionary forces. The presence of linkage disequilibrium (gametic disequilibrium) is thus an indication that either stochastic (*e.g.*, drift) or deterministic (*e.g.*, selection, gene flow) evolutionary forces have been acting on a population (Hedrick 2000; Ardlie *et al.* 2002).

The classical definition of linkage disequilibrium, D, follows the probability theory definition of deviation from independence. Independence of two events, B and C, means that $\Pr(BC) = \Pr(B) \cdot \Pr(C)$, where Pr is probability and BC is the joint distribution of B and C, so that the deviation from independence is measured as $D = \Pr(BC) - \Pr(B) \cdot \Pr(C)$. Changing notation slightly to let $A_{k(i)}$ designate the kth allele at the ith locus gives the linkage disequilibrium between the alleles at two loci, D_2, as $D_2 = \Pr(A_{k(1)}A_{k(2)}) - \Pr(A_{k(1)}) \cdot \Pr(A_{k(2)})$, where Pr represents probability and $A_{k(1)}A_{k(2)}$ represents the joint occurrence of $A_{k(1)}$ and $A_{k(2)}$ in a single haploid gamete. In most modern interpretations of probability theory, the primitive concept of "probability" is interpreted as a relative frequency; therefore, $\Pr(A_{k(1)})$ is the same as the frequency of allele k at locus 1.

[1]*Corresponding author:* Department of Biology, Arizona State University, P.O. Box 871501, Tempe, AZ 85287-1501.
E-mail: cycad@asu.edu

The quintessential examples of linkage disequilibrium are coadapted gene complexes, in which several loci are tightly linked because they provide a large selective advantage if they occur together. In these cases, linkage disequilibrium is maintained by selection. Although coadapted gene complexes are implicit in Wright's shifting-balance hypothesis (Wright 1931), have been used to explain outbreeding depression (Dobzhansky 1948; Lynch 1991), and are frequently cited as evolutionary hypotheses (Palopoli and Wu 1996; Rawson and Burton 2002), the linkage disequilibrium of these purported coadapted gene complexes is almost never quantified. This is particularly surprising given the well-cited article by Geiringer (1944), in which she provides most of the algorithm for computing higher-order linkage disequilibrium coefficients. In this article, we complete and simplify Geiringer's formulation and then show how the sums of products of those coefficients equal the definition of (total) linkage disequilibrium as the deviation from probabilistic independence given by Dausset *et al.* (1978).

Methodologically, we follow Geiringer's lead and decompose higher-order linkage disequilibrium into lower-order linkage disequilibrium terms. In other words, we take a top-down approach to defining multilocus linkage disequilibrium, rather than the bottom-up approach followed by virtually everyone since Geiringer (1944). Lewontin (1974) is typical of the bottom-up approach. There are very few other top-down decomposition approaches such as Bulmer's (1980) decomposition of multilocus epistasis or Wagner and Laubichler's (2000) character decomposition approach in population genetics.

In this article, we first define linkage disequilibrium at a single locus as the allele frequency at this locus, which greatly simplifies notation. Second, we extend the definition of linkage disequilibrium to multiple loci

by invoking compositions of positive integers. Our decomposition of multilocus linkage disequilibrium is entirely consistent with the standard definitions for two loci, as well as its previous extensions to three, four, and six loci (GEIRINGER 1944; BENNETT 1954; HASTINGS 1984). Third, we show how this definition is entirely consistent with the notion of linkage disequilibrium as the deviation from probabilistic independence.

DECOMPOSITION OF MULTILOCUS LINKAGE DISEQUILIBRIUM

Define the one-locus coefficient of linkage disequilibrium, D_1, as $D_1(A_{k(i)}) = \Pr(A_{k(i)})$. This definition may appear paradoxical, but it dramatically simplifies notation for the decomposition of multilocus linkage disequilibrium. In elementary algebra we have the analogous problem of defining the algebraic expression x^n when $n = 0$ (LAKOFF and NÚÑEZ 2000). Note that our definition of a locus encompasses protein-coding loci, quantitative trait loci, and even single nucleotides.

Following HASTINGS (1984), the formulas for two- and three-locus multilocus linkage disequilibrium, in which $D_1(A_{k(i)})$ was substituted for $\Pr(A_{k(i)})$, are defined as

$$D_2 = \Pr(A_{k(1)}A_{k(2)}) - D_1(A_{k(1)}) \cdot D_1(A_{k(2)})$$

$$D_3 = \Pr(A_{k(1)}A_{k(2)}A_{k(3)}) - D_1(A_{k(1)}) \cdot D_1(A_{k(2)}) \cdot D_1(A_{k(3)}) - D_1(A_{k(1)}) \cdot D_2(A_{k(2)}A_{k(3)}) - D_1(A_{k(2)}) \cdot D_2(A_{k(1)}A_{k(3)}) - D_1(A_{k(3)}) \cdot D_2(A_{k(1)}A_{k(2)}).$$

Let D_n be the coefficient of linkage disequilibrium between n loci. Then the pattern here is that $D_n = \Pr(A_{k(1)}A_{k(2)} \ldots A_{k(n)})$ minus all possible products of lower-order linkage disequilibrium coefficients, such that each term has all of its subscripts adding up to n. The key to writing down an explicit formula for D_n is that the phrase "all possibilities of the subscripts adding up to n" refers to partitions of the positive integer n (ANDREWS 1976). A partition π of a positive integer n is a set of positive integers that adds up to n; *i.e.*, $\pi = \{n_1, n_2, \ldots, n_m\}$ such that $\Sigma_{i=1}^{m} n_i = n$. The set of all partitions of n is designated $p(n)$; *e.g.*, $p(5) = \{\{5\}, \{4, 1\}, \{3, 2\}, \{2, 2, 1\}, \{3, 1, 1\}, \{2, 1, 1, 1\}, \{1, 1, 1, 1, 1\}\}$. To define multilocus linkage disequilibrium, we have to add over all partitions, excluding the trivial partition $\pi = \{n\}$, and permute over all alleles for a given number of loci. However, the order of elements of the partition matters, and hence we construct the number-theoretic compositions c of the positive integer n (ANDREWS 1976). For example, all of the compositions for the partition $\pi = \{2, 2, 1\}$ are the ordered triples (2, 2, 1), (2, 1, 2), and (1, 2, 2). Using these mathematical notions we can generalize the two- and three-locus cases to define linkage disequilibrium between n loci as

$$D_n(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \Pr(A_{k(1)}A_{k(2)} \ldots A_{k(n)}) - \sum_{\substack{\text{all compositions } c \text{ of } n \\ \text{except } c=(n)}} \left[\prod_{n_i \in c} D_{n_i}(\ldots)\right], \quad (1a)$$

where $n_i \in c$ means that $n_i \in c$ is a scalar component of the vector c. Equivalently,

$$D_n(A_{k(1)}A_{k(2)} \ldots A_{k(n)}) = \Pr(A_{k(1)}A_{k(2)} \ldots A_{k(n)}) - \sum_{\substack{\sum_{i=1}^{m} n_i = n \\ 1 \le n_i < n \\ 1 \le m \le n}} \left[\prod_{i=1}^{m} D_{n_i}(\ldots)\right]. \quad (1b)$$

The only way to decompose n into a single positive integer is $c = (n)$. Therefore, we can also write the highest-order coefficient of linkage disequilibrium as $D_n(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \Sigma_{c=(n)} [\Pi_{n_i \in c} D_{n_i}(\ldots)]$, where the summation has only a single term and the product has only a single factor. Therefore, Equation 1a yields

$$\Pr(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \sum_{\text{all compositions } c \text{ of } n} \left[\prod_{n_i \in c} D_{n_i}(\ldots)\right], \quad (2)$$

which we use below.

Equation 1 has never been written explicitly for general multilocus linkage disequilibrium, even though special cases have been given by GEIRINGER (1944), BENNETT (1954), and HASTINGS (1984). The only explicit definition previously given for multilocus linkage disequilibrium is due to DAUSSET *et al.* (1978),

$$\mathbf{D}_n(A_{k(1)}A_{k(2)} \ldots A_{k(n)}) = \Pr(A_{k(1)}A_{k(2)}A_{k(3)} \ldots A_{k(n)}) - \prod_{i=1}^{n} D_1(A_{k(i)}), \quad (3)$$

which we call total linkage disequilibrium, $\mathbf{D}_n$, where we have again replaced $\Pr(A_{k(i)})$ with $D_1(A_{k(i)})$. We refer to $\mathbf{D}_n$ as total linkage disequilibrium because, as we show below, all of the nonboldface linkage disequilibrium coefficients $D_1, D_2, D_3, \ldots, D_n$ can be independent from one another and contribute to $\mathbf{D}_n$. Equation 3 has a simple heuristic interpretation: $\mathbf{D}_n(A_{k(1)} \ldots A_{k(n)})$ measures how far the haploid genotype at all n loci deviates from probabilistic independence.

We are now ready to derive the relationship between D_n and $\mathbf{D}_n$. In Equation 3, substitute $\Sigma_{\text{all compositions } c \text{ of } n} [\Pi_{n_i \in c} D_{n_i}(\ldots)]$ for $\Pr(A_{k(1)}, \ldots A_{k(n)})$ (see Equation 2), yielding $\mathbf{D}_n(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \Sigma_{\text{all compositions } c \text{ of } n} [\Pi_{n_i \in c} D_{n_i}(\ldots)] - \Pi_{i=1}^{n} D_1(A_{k(i)})$. The last term in this equation is simply the value of $\Pi_{n_i \in c} D_{n_i}(\ldots)$ for the composition $c = (1, 1, 1, \ldots, 1)$, *i.e.*,

$$n = \underbrace{1 + 1 + \ldots + 1}_{n \text{ times}}.$$

Therefore, Equation 3 becomes

$$\mathbf{D}_n(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \sum_{\substack{\text{all compositions } c \text{ of } n \\ \text{except } c=(1,1,1,\ldots,1)}} \left[\prod_{n_i \in c} D_{n_i}(\ldots)\right] \quad (4a)$$

or, equivalently,

$$\mathbf{D}_n(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \sum_{\substack{\sum_{i=1}^{m} n_i = n \\ 1 \le n_i < n \\ 1 \le m < n}} \left[\prod_{i=1}^{m} D_{n_i}(\ldots)\right]. \quad (4b)$$

Equation 4 provides the crucial link between deviations from independence ($\mathbf{D}_n$) and the linkage disequilibrium coefficients D_n computed by GEIRINGER (1944) and her intellectual successors by decomposing $\mathbf{D}_n$ into the terms D_{n_i}, where $\Sigma n_i = n$.

DISCUSSION

We have converted the genetics problem of decomposing linkage disequilibrium into the mathematical problem of decomposing positive integers into their additive parts, all while maintaining the convenient heuristic definition of total linkage disequilibrium as the deviation from independence. Unlike GEIRINGER (1944), we can write down an explicit formula for multilocus linkage disequilibrium because we invoke partitions of integers and define $D_1(A) = \Pr(A)$, thereby merging her notion of linkage disequilibrium with those of DAUSSET *et al.* (1978).

One immediate consequence of our decomposition approach is that the single highest-order coefficient of linkage disequilibrium, D_n, cannot be examined in isolation. Because $D_n(A_{k(1)}, A_{k(2)}, \ldots A_{k(n)}) = \Sigma_{\text{except } c=(1,1,1,\ldots,1)}^{\text{all compositions } c \text{ of } n} [\Pi_{n_i \in c} D_{n_i}(\ldots)]$, we need to examine all lower-order linkage disequilibrium coefficients, $D_{n_i}(\ldots)$ with $n_i < n$. All of the subscripted linkage disequilibrium coefficients $D_1, D_2, D_3, \ldots, D_n$ can be independent from one another and all contribute to $\mathbf{D}_n$, which we therefore call total linkage disequilibrium.

Multilocus definitions of linkage disequilibrium have not been used very often in empirical studies because of the large number of inputs and linkage disequilibrium coefficients that must be analyzed ($2^n - 1$). Currently, even third-order linkage disequilibrium is seldom measured (THOMSON and BAUR 1984). However, explicit terms for multilocus linkage disequilibrium are of theoretical importance.

One important theoretical application is the analysis of multilocus epistasis. CHEVERUD and ROUTMAN (1995) developed a two-locus model of physiological epistasis that has been further refined by WAGNER *et al.* (1998). To analyze the evolutionary consequences of epistasis in these models, one has to first define linkage disequilibrium for a subset of the loci. Thus, to extend models of physiological epistasis to multiple loci, we must first define linkage disequilibrium for that subset of loci, which we have just done. Models of multilocus epistasis will be crucial in debates over what factors maintain coadapted gene complexes, increase additive genetic variance, and foster speciation (GOODNIGHT 1988, 1995; WADE and GOODNIGHT 1998).

We thank Phil Hedrick, Tom Dowling, and two anonymous reviewers for their helpful comments.

LITERATURE CITED

ANDREWS, G. E., 1976 *The Theory of Partitions.* Addison-Wesley, Reading, MA.

ARDLIE, K. G., L. KRUGLYAK and M. SEIELSTAD, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

BENNETT, J. H., 1954 On the theory of random mating. Ann. Eugen. **18:** 311–317.

BULMER, M. G., 1980 *The Mathematical Theory of Quantitative Genetics.* Clarendon Press, Oxford.

CHEVERUD, J. M., and E. J. ROUTMAN, 1995 Epistasis and its contribution to genetic variance components. Genetics **139:** 1455–1461.

DAUSSET, J., L. LEGRAND, V. LEPAGE, L. CONTÚ, A. MARCELLI-BARGE *et al.*, 1978 A haplotype study of HLA complex with special reference to HLA-DR series and to Bf, C2 and glyoxalase I polymorphisms. Tissue Antigens **12:** 297–307.

DOBZHANSKY, T., 1948 Genetics of natural populations. XVIII. Experiments on chromosomes of *Drosophila pseudoobscura* from different geographical regions. Genetics **33:** 797–808.

GEIRINGER, H., 1944 On the probability theory of linkage in Mendelian heredity. Ann. Math. Stat. **15:** 25–57.

GOODNIGHT, C. J., 1988 Epistasis and the effect of founder events on the additive genetic variance. Evolution **42:** 441–454.

GOODNIGHT, C. J., 1995 Epistasis and the increase in additive genetic variance: implications for phase 1 of Wright's shifting balance process. Evolution **49:** 502–511.

HASTINGS, A., 1984 Linkage disequilibrium, selection and recombination at three loci. Genetics **106:** 153–164.

HEDRICK, P. W., 2000 *Genetics of Populations,* Ed. 2. Jones & Bartlett, Sudbury, MA.

LAKOFF, G., and R. E. NÚÑEZ, 2000 *Where Mathematics Comes From.* Basic Books, New York.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

LYNCH, M., 1991 The genetic interpretation of inbreeding depression and outbreeding depression. Evolution **45:** 622–629.

PALOPOLI, M. F., and C.-I WU, 1996 Rapid evolution of a coadapted gene complex: evidence from the segregation distorter (SD) system of meiotic drive in *Drosophila melanogaster.* Genetics **143:** 1675–1688.

RAWSON, P. D., and R. S. BURTON, 2002 Functional coadaptation between cytochrome c and cytochrome c oxidase within allopatric populations of a marine copepod. Proc. Natl. Acad. Sci. USA **99:** 12955–12958.

THOMSON, G., and M. P. BAUR, 1984 Third order linkage disequilibrium. Tissue Antigens **24:** 250–255.

WADE, M. J., and C. J. GOODNIGHT, 1998 Perspective: the theories of Fisher and Wright in the context of metapopulations: when nature does many small experiments. Evolution **52:** 1537–1553.

WAGNER, G. P., and M. D. LAUBICHLER, 2000 Character identification in evolutionary biology: the role of the organism. Theory Biosci. **119:** 20–40.

WAGNER, G. P., M. D. LAUBICHLER and H. BAGHERI-CHAICHIAN, 1998 Genetic measurement theory of epistatic effects. Genetica **102/103:** 569–580.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Communicating editor: M. W. FELDMAN