Published in final edited form as:

Theor Popul Biol. 2007 September ; 72(2): 179–185. doi:10.1016/j.tpb.2006.05.006.

Prediction of multi-locus inbreeding coefficients and relation to linkage disequilibrium in random mating populations

William G. Hill and

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK

Bruce S. Weir

Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA

Abstract

An algorithm to predict the level of identity by descent simultaneously at multiple loci is presented, which can in principle be extended to any number of loci. The model assumes a random mating population, with random association of haplotypes. The relationship is shown between coefficients of multi-locus identity or non-identity by descent and moments of multi-locus linkage disequilibrium. Thus these moments can be computed from the multilocus identity or, using algorithms derived previously to predict the disequilibria moments, vice-versa. The results can be applied to predict multi-locus identity in, for example, gene mapping.

Keywords

Identity by descent; Multiple loci; Linkage disequilibrium; Random drift; Random mating

INTRODUCTION

The probability of identity by descent simultaneously at two or more neutral loci is a generalization of Wright's inbreeding coefficient, F . Such probabilities are clearly functions of both the population size and the breeding structure, as is F , but they also depend on the linkage relationships between the loci. For example, in a closed random mating population without mutation, the probability of double identity is approximately equal to F^2 for unlinked loci, but increases to F for a completely linked pair. The multi-locus identity is a useful parameter in predicting the joint ancestry of pairs of loci, for example in mapping studies, in inferences about historic population structure from current data (Hayes *et al.*, 2003), and also in computing variances and covariances of quantitative traits in finite populations. Whereas contributions to variance in the absence of epistasis depend only on two-locus identities or disequilibria, with epistasis multi-locus terms may be involved.

Weir and Cockerham (1969, 1974) derived algorithms for the computation of identity by descent at two loci. Recently Hernández-Sánchez *et al.* (2004) have provided approximate methods for predicting identity at three and four (and, in principle, more) loci. They use the

Corresponding author: Prof W. G. Hill, Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK, Phone: +44-131-650 5705, Fax: +44-131-650 6564, e-mail: w.g.hill@ed.ac.uk.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

results of Weir and Cockerham for two loci, and then predict by regression the conditional probability of identity at the third (and fourth) locus conditional on identity at the pair. Here we provide algorithms for computing exact values of multi-locus inbreeding for random mating populations with random selfing.

Just as there is a linear relationship between F (as identity by descent) and heterozygosity within populations, so there is between the moments for linkage disequilibrium and multi-locus identity. It is therefore possible to compute multi-locus identities from linkage disequilibrium and vice versa. Weir and Cockerham (1974) show the relationship for a pair of loci. Algorithms for the computation of multi-locus linkage disequilibrium have been given previously (Hill, 1974; Hill and Weir, 1988). We show here how these relate to multi-locus identity coefficients. For populations undergoing random mating and random selfing as assumed here, haplotype frequencies are a sufficient descriptor of the population. There is also assumed to be no selection at the loci of interest, or at loci linked to them.

DEFINITIONS

Identity and Non-identity Coefficients

To define identity coefficients, we need to consider the number of haplotypes sampled and we adopt a notation that is more easily extended to multiple loci than that of Weir and Cockerham (1969). Let $F_{m,h;\text{hap}}$ denote the probability of identity by descent at each of m loci, $A, B, \dots$ in a sample of h haplotypes ($2 \leq h \leq 2m$), where ‘hap’ denotes the haplotype structure of the sample. Thus $F_{1,2:A,A} = F_A$ and $F_{1,2:B,B} = F_B$ are the probabilities of identity of two alleles sampled at locus A and locus B respectively, and are the usual single-locus inbreeding coefficients. In the absence of mutation, or with equal mutation rates at each locus, $F_A = F_B = \dots$. Similarly $F_{2,2:AB,AB}$ denotes the probability of identity at loci A and B in a pair of haplotypes, and is the two-locus inbreeding coefficient F_{AB} (denoted F_{11} by Weir and Cockerham, 1969, for two haplotypes within one individual). The magnitude of this two-locus identity in successive generations also depends on the probability of identity of genes sampled from three and four haplotypes at two loci, $F_{2,3:A,B,AB}$ and $F_{2,4:A,A,B,B}$, respectively. Weir and Cockerham (1974 and elsewhere) denote two, three and four haplotype non-identities at two loci as Θ , Γ and Δ .

Similarly, let $X_{m,h;\text{hap}}$ be the equivalent probability of non-identity by descent at each of the m loci. Thus $X_{1,2:A,A} = 1 - F_{1,2:A,A} = 1 - F_A$ and, for example, $X_{2,2:AB,AB} = 1 - F_A - F_B + F_{2,2:AB,AB}$ and $X_{2,3:A,B,AB} = 1 - F_A - F_B + F_{2,3:A,B,AB}$, as can be seen by writing the four two-locus probabilities of identity or non-identity for a pair of loci in a 2×2 table with single-locus probabilities as marginal totals. Let the vector $\mathbf{x}$ define the non-identity coefficients,

$$\mathbf{x} = \begin{bmatrix} X_{2,2:AB,AB} \\ X_{2,3:A,B,AB} \\ X_{2,4:A,A,B,B} \end{bmatrix}. \quad (1)$$

For three loci $F_{3,2:ABC,ABC}$ is the probability of identity at each of three loci for alleles on two haplotypes, with

$$X_{3,2:ABC,ABC} = 1 - F_A - F_B - F_C + F_{2,2:AB,AB} + F_{2,2:AC,AC} + F_{2,2:BC,BC} - F_{3,2:ABC,ABC}. \quad (2)$$

Relationships between the 16 different three-locus identity and non-identity coefficients are shown in Appendix 1. These follow from writing the eight joint identity/non-identity

probabilities in a $2 \times 2 \times 2$ table with appropriate two-locus marginals. The number of coefficients rises rapidly with the number of loci, e.g. 139 for four loci (not listed!).

Linkage Disequilibrium Coefficients and Moments

It is sufficient to assume in the present analysis of disequilibrium calculations that there are two alleles at each locus. At locus A , for example, let the frequency of allele A be p_A and the frequency of the alternative allele, a , be $p_a = 1 - p_A$. To simplify later formulae for moments of disequilibria, let $\pi_A = p_A(1 - p_A) = p_A p_a$ and $\tau_A = (1 - 2p_A)/2 = (p_a - p_A)/2$. To simplify relations between identities and disequilibria, this last quantity is one-half of the τ used by Hill and Weir (1988). The frequency of the two-locus haplotype AB , for example, is p_{AB} , and the corresponding disequilibrium coefficient is $D_{AB} = p_{AB} - p_A p_B$. Note that $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$. For three loci $D_{ABC} = p_{ABC} - p_A p_B p_C - p_B p_A C - p_C p_A B + 2p_A p_B p_C$, and, for example, $D_{ABC} = -D_{ABc} = D_{Abc} = -D_{abc}$. For four loci $D_{ABCD} = p_{ABCD} - \sum_4 p_A p_B p_C D - \sum_3 p_A p_B p_C D + 2\sum_6 p_A p_B p_C D - 6p_A p_B p_C p_D$, where, for example, $\sum_4 p_A p_B p_C D$ denotes the sum of the four terms found by permuting A, B, C, D : $\sum_4 p_A p_B p_C D = p_A p_B p_C D + p_B p_A C D + p_C p_A B D + p_D p_A B C$; and $D_{ABCD} = -D_{ABcD} = -D_{ABcd} = -D_{Abcd} = D_{abcd}$.

Expected changes in variances and covariances of linkage disequilibria involve changes in interrelated moments (Hill and Robertson, 1968; Hill, 1974; Hill and Weir, 1988). The quantities used here, defined by the vector $\mathbf{v}$, facilitate comparisons between moments and the measures of identity. For two loci

$$\mathbf{v} = \begin{bmatrix} E(D_{AB}^2) \\ E(\tau_A \tau_B D_{AB}) \\ E(\pi_A \pi_B) \end{bmatrix} = \begin{bmatrix} E(D_{AB} D_{ab}) \\ E(p_A p_B D_{ab} + p_A p_b D_{aB} + p_a p_B D_{Ab} + p_a p_b D_{AB}) \\ E(p_A p_a p_B p_b) \end{bmatrix}. \quad (3)$$

Expectation in this context refers to the average over all finite populations that descend from a specified initial population. The elements of $\mathbf{v}$ for three loci are given in Table 1. Note that there is the same number (e.g. 16 for three loci) of disequilibrium moments and non-identity terms, i.e. $\mathbf{x}$ and $\mathbf{v}$ have the same dimension.

CORRESPONDENCE BETWEEN IDENTITY AND DISEQUILIBRIUM COEFFICIENTS

There is a linear relationship between moments and identity coefficients for a closed random mating finite population (Weir and Cockerham, 1974). It is simpler in the present model because here we assume that genotype frequencies equal products of haplotype frequencies, i.e. a haploid model with frequencies of each single locus genotype in Hardy-Weinberg proportions. Although not required for computation of the identities, we assume that the population was initially in linkage equilibrium, i.e. $\mathbf{v}_0' = [0, 0, 1](\pi_A \pi_B)_0$, and note that $E(\pi_A \pi_B)_t$ is the probability that, at generation t , from a sample of four haplotypes, the alleles sampled at locus A from two haplotypes are different and those at locus B sampled from two other haplotypes are also different. Hence $E(\pi_A \pi_B)_t = X_{2,4:A,B,A,B,t}(\pi_A \pi_B)_0$. (To simplify notation, subsequently in this section generation t is implied in the formulae.) Further, using the fact that for a population initially in linkage equilibrium, $E(D_{AB}) = E(p_a D_{AB}) = 0$ (Hill, 1974) so $E(\tau_A \tau_B D_{AB}) = E(p_a p_b p_{AB} - p_A p_a p_B p_b)$, and noting that $E(D_{AB}^2) = E(D_{AB} D_{ab}) = E[(p_{AB} - p_A p_B)(p_{ab} - p_a p_b)]$, we obtain the relationships:

$$\begin{aligned} E(D_{AB}^2) &= (X_{2,2:AB,AB} - 2X_{2,3:A,B,AB} + X_{2,4:A,B,A,B})(\pi_A \pi_B)_0 \\ E(\tau_A \tau_B D_{AB}) &= (X_{2,3:A,B,AB} - X_{2,4:A,B,A,B})(\pi_A \pi_B)_0 \\ E(\pi_A \pi_B) &= X_{2,4:A,B,A,B}(\pi_A \pi_B)_0. \end{aligned}$$

In matrix terms

$$\mathbf{v} = \mathbf{Q}\mathbf{x}(\pi_A \pi_B)_0 \text{ and } \mathbf{x} = \mathbf{Q}^{-1}\mathbf{v}[(\pi_A \pi_B)_0]^{-1}, \quad (4)$$

where

$$\mathbf{Q} = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{Q}^{-1} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \quad (5)$$

From (3) and (5), it can be shown that

$$\mathbf{Q}^{-1}\mathbf{v} = \frac{1}{4} \begin{bmatrix} E[2(p_{AB}p_{ab} + p_{Ab}p_{aB})] \\ E[p_{AB}p_a p_b + p_{Ab}p_b p_B + p_{aB}p_A p_b + p_{ab}p_A p_B] \\ E[4p_A p_a p_B p_b] \end{bmatrix}$$

The first element of $4\mathbf{Q}^{-1}\mathbf{v}$ is the expected frequency of double heterozygotes, i.e. of genotypes at which both loci on two haplotypes are non-identical in state and thus also non-identical by descent. Similarly the second and third elements are the expected frequencies of doubly heterozygous genotypes (i.e. non-identical in state and by descent) formed from, respectively, three haplotypes, taking two loci at one and one each from the others, and four haplotypes, taking one allele from each. Equivalent relationships apply with more loci.

The elements of $\mathbf{Q}$ relating non-identity coefficients $\mathbf{x}$ and moments $\mathbf{v}$ for three loci are given in Table 1. These follow from the definitions of the disequilibrium components, and for clarity the elements of each row are scaled such that the corresponding diagonal elements are all +1. The coefficients (κ) are +1 or -1, with the latter whenever an odd number of loci have to be represented by the alternative allele in the disequilibria, e.g. $D_{ABC}^2 = -D_{ABC}D_{abc}$ and $D_{AB}D_{AC}D_{BC} = -D_{AB}D_{aC}D_{bC}$, but $\tau_C D_{AB}D_{ABC} = \tau_C D_{AB}D_{abC}$. For example, as $D_{ABC}^2 = -D_{ABC}D_{abc}$ and $D_{ABC} = p_{ABC} - \Sigma_3 p_{APBC} + 2p_{APBP_C}$,

$$\begin{aligned} E(D_{ABC}^2) &= -[X_{3,2:ABC,ABC} - 2\Sigma_3 X_{3,3:A,BC,ABC} + 4X_{3,4:A,B,C,ABC} + \Sigma_3 X_{3,4:A,A,BC,BC} \\ &\quad + 2\Sigma_3 X_{3,4:B,C,AB,AC} - 4\Sigma_3 X_{3,5:A,A,B,C,BC} + 4X_{3,6:A,A,B,B,C,C}](\pi_A \pi_B \pi_C)_0. \end{aligned} \quad (6)$$

TRANSITION EQUATIONS FOR RECOMBINATION AND DRIFT

The population is assumed to be random mating including random selfing, with no selection and effective population size N diploids. The recombination fraction between pairs of loci is given by, for example, r_{AB} for loci A and B . The term r_{ABC} denotes the probability of any recombination between the three loci, i.e. $(1 - r_{ABC})$ is the probability the haplotype ABC is

transmitted intact, such that, in the absence of interference and if A, B and C are so ordered on the chromosome, $(1 - r_{ABC}) = (1 - r_{AB})(1 - r_{BC})$. A simple haploid model is assumed for the random sampling of gametes after recombination, such that, for example, $E(D_{AB,t+1}) = [1 - 1/(2N)](1 - r_{AB})E(D_{AB,t})$ where t denotes generation. The expectation process E is over all finite populations that descend from a specified initial population. As Weir and Cockerham (1974) point out, this is an approximation for a model in which recombination occurs only between haplotypes carried by an individual, but otherwise the analysis becomes impracticable for more than two loci and is an increasingly better approximation for increasing population size and decreasing recombination rate.

We obtain the recurrence relationships both for non-identity coefficients and for moments of disequilibria in order to show the processes involved although, as shown above, one set can be expressed in terms of the other. Details of the relevant matrices are given for two loci; results are summarised for three loci and the matrix equations are general.

Identity Coefficients

Assuming random mating and random selfing, with recombination preceding sampling, the expected non-identities among the gametes at generation $t+1$ in terms of those at t are

$$\mathbf{x}_{t+1} = \mathbf{T}\mathbf{x}_t = \mathbf{L}\mathbf{U}\mathbf{x}_t, \quad (7)$$

where $\mathbf{U}$ denotes transition probabilities due to recombination, $\mathbf{L}$ those due to sampling and $\mathbf{T} = \mathbf{L}\mathbf{U}$ the complete transition. Thus if all alleles are non-identical in the base or reference population,

$$\mathbf{x}_0 = \mathbf{1} \quad \text{and} \quad \mathbf{x}_t = \mathbf{T}^t \mathbf{1}, \quad (8)$$

where $\mathbf{1}$ is the unit vector.

For two loci,

$$\mathbf{U} = \begin{bmatrix} (1 - r_{AB})^2 & 2r_{AB}(1 - r_{AB}) & r_{AB}^2 \\ 0 & 1 - r_{AB} & r_{AB} \\ 0 & 0 & 1 \end{bmatrix} \quad (9)$$

For example, the first row of $\mathbf{U}$ defines the probabilities that two pairs of non-identical alleles on two haplotypes, $x_1 = X_{2,2;AB,AB}$, can have arisen from no recombination on a pair of non-identical haplotypes (U_{11}), no recombination on one and a recombination generating one haplotype on the other (U_{12}), or recombination on four haplotypes to create a non-identical pair (U_{13}).

The subsequent sampling of haplotypes with only non-identical alleles is given by

$$\mathbf{L} = (1 - \frac{1}{2N}) \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2N} & 1 - \frac{1}{N} & 0 \\ \frac{1}{2N^2} & \frac{2}{N}(1 - \frac{1}{N}) & (1 - \frac{1}{N})(1 - \frac{3}{2N}) \end{bmatrix} \quad (10)$$

(see also Weir and Cockerham, 1974, p. 349, setting recombination fractions to zero, $\lambda = 1$ in Ω_{EC} in their notation). The diagonal elements of $\mathbf{L}$, which correspond to sampling of h haplotypes in offspring from h different haplotypes in parents are simply $1[1-1/(2N)] \dots [1-(h-1)/(2N)]$, where $h = 2, 3$ and 4 for rows 1, 2 and 3 of $\mathbf{L}$. For sub-diagonal elements, where h' offspring haplotypes are sampled from h ($h < h'$) parental haplotypes, the coefficients are $k[1/(2N)^{(h'-h)}][1-1/(2N)] \dots [1-(h-1)/(2N)]$, where the number k of possible sampling sequences depends on the haplotypic designations of parents and offspring. We illustrate by example as no closed formula has been obtained, but an algorithm is given in Appendix 2.

Consider the element L_{32} , where four offspring haplotypes defined by alleles A, A, B, B are sampled from three parental haplotypes defined by (i) AB , (ii) A , (iii) B . The four possible sequences are: 1: $A(i), A(ii), B(i), B(iii)$ (shorthand for A from (i), A from (ii), B from (i), B from (iii)); 2: $A(i), A(ii), B(iii), B(i)$; 3: $A(ii), A(i), B(i), B(iii)$; and 4: $A(ii), A(i), B(iii), B(i)$. Thus $h = 3, h' = 4, k = 4$ and $L_{32} = 4[1/(2N)][1-1/(2N)][1-2/(2N)]$. Note that when two parental haplotypes are the same, e.g. in element L_{31} , ordering of these is irrelevant. For example, if the parental haplotypes are defined by (i) AB and (ii) AB , there are only two relevant orderings (i.e. $k = 2$), because $A(i), A(ii), B(i), B(ii) \equiv A(ii), A(i), B(ii), B(i)$ and $A(i), A(ii), B(ii), B(i) \equiv A(ii), A(i), B(i), B(ii)$.

Moments of Disequilibria

For two loci, the moments needed to describe transitions in squared linkage disequilibrium are given by the vector $\mathbf{v}$, for which a recurrence relation was given by Hill and Robertson (1968), assuming as above that recombination preceded sampling (but they used $(1 - 2p_A)(1 - 2p_B)D_{AB} = 4\tau_A\tau_B D_{AB}$ so coefficients differ):

$$\mathbf{v}_{t+1} = \mathbf{M}\mathbf{v}_t = \mathbf{K}\mathbf{R}\mathbf{v}_t \quad (11)$$

where

$$\mathbf{R} = \begin{bmatrix} (1 - r_{AB})^2 & 0 & 0 \\ 0 & 1 - r_{AB} & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (12)$$

denotes transitions due solely to recombination, and

$$\mathbf{K} = (1 - \frac{1}{2N}) \begin{bmatrix} \frac{1}{4N^2} + (1 - \frac{1}{2N})^2 & \frac{2}{N}(1 - \frac{1}{2N}) & \frac{1}{2N} \\ \frac{1}{2N}(1 - \frac{1}{N}) & (1 - \frac{1}{N})^2 & 0 \\ \frac{1}{2N^2} & \frac{2}{N}(1 - \frac{1}{2N}) & 1 - \frac{1}{2N} \end{bmatrix} \quad (13)$$

those due solely to sampling.

Using the relationship $\mathbf{v}_t = \mathbf{Q}\mathbf{x}_t$ (see 4), the recurrence relations for $\mathbf{v}$ and $\mathbf{x}$ can therefore also be written as

$$\mathbf{x}_t = \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}\mathbf{x}_{t-1} \quad \text{and} \quad \mathbf{v}_t = \mathbf{Q}\mathbf{T}'\mathbf{Q}^{-1}\mathbf{v}_{t-1}, \quad (14)$$

$$\mathbf{x}_t = \mathbf{T}' \mathbf{x}_0 = \mathbf{Q}^{-1} \mathbf{M}' \mathbf{Q} \mathbf{x}_0 \quad \text{and} \quad \mathbf{v}_t = \mathbf{M}' \mathbf{v}_0 = \mathbf{Q} \mathbf{T}' \mathbf{Q}^{-1} \mathbf{v}_0. \quad (15)$$

In the absence of recombination (i.e. r_{AB}, r_{AC} , etc. $\rightarrow 0$), $\mathbf{U} \rightarrow \mathbf{I}$ and $\mathbf{R} \rightarrow \mathbf{I}$ (eq. 9, 12), and hence $\mathbf{M} \rightarrow \mathbf{Q} \mathbf{L} \mathbf{Q}^{-1}$. In the absence of drift (i.e., $1/N \rightarrow 0$), $\mathbf{L} \rightarrow \mathbf{I}$ and $\mathbf{K} \rightarrow \mathbf{I}$ (eq. 10, 13) so that $\mathbf{U} \rightarrow \mathbf{Q}^{-1} \mathbf{R} \mathbf{Q}$. As $\mathbf{R}$ is diagonal, the relationship $\mathbf{U} = \mathbf{Q}^{-1} \mathbf{R} \mathbf{Q}$ is a convenient route for constructing $\mathbf{U}$ in the multi-locus case; indeed the most simply obtained matrices are $\mathbf{R}$, $\mathbf{Q}$ and $\mathbf{L}$.

Analysis for Three (or More) Loci

The elements of the diagonal matrix $\mathbf{R}$, defining changes in disequilibrium terms due to recombination, are also given in Table 1. Their definitions are obvious. The elements of $\mathbf{L}$ given in Table 2 are obtained by extending the arguments used to derive $\mathbf{L}$ for two loci (Eq. 10 and subsequent text, and Appendix.)

The matrix $\mathbf{U}$ for three loci, which defines recombination among identity coefficients, can be derived directly by considering how recombination among haplotypes can provide the relevant reduced set. For example, for element (12,16) the haplotypes for $X_{3,4:A,B,AC,BC}$ can arise from six haplotypes corresponding to $X_{3,6:A,A,B,B,C,C}$ by recombination both between A - - and - - C and between - B - and - - C haplotypes, with probability $r_{AC}r_{BC}$. It is simpler, however, to compute $\mathbf{U} = \mathbf{Q}^{-1} \mathbf{R} \mathbf{Q}$.

Mutation

Consider the infinite alleles model of Kimura and Crow (1964), such that mutation leads to a new allele, and also assume that alleles that are not identical in state as a consequence of mutation are not identical by descent, even though their ancestry coalesces. Hence the probability of multiple-locus heterozygosity and identity asymptote at values dependent on population size, mutation rates and recombination fractions. Let u_A denote the mutation rate at locus A , and assume it is sufficiently small that all terms such as u_A^2 , u_A/N , $u_A r_{AB}$ and similarly for other loci can be ignored. Then, in the present notation,

$$X_{1,2:A,A,t+1} = [1 - 1/(2N)]X_{1,2:A,A,t} + 2u_A(1 - X_{1,1:A,A,t}), \quad (16)$$

which is simply $1 - F_{A,t+1} = [1 - 1/(2N)(1 - F_{A,t}) + 2u_A F_{A,t}]$, a version of the standard equation of Kimura and Crow (1964). Similarly, allowing for mutation among haplotypes which are identical at one locus and not at the other, for two loci

$$\mathbf{x}_{t+1} = \mathbf{T} \mathbf{x}_t + 2u_A (X_{1,1:B,B,t} \mathbf{1} - \mathbf{x}_t) + 2u_B (X_{1,1:A,A,t} \mathbf{1} - \mathbf{x}_t) \quad (17)$$

where $\mathbf{1}$ is the unit vector. For three loci, an example (element 2) of the relevant expressions, letting $\mathbf{t}_2'$ denote the second row of $\mathbf{T}$, is:

$$X_{3,3:A,BC,ABC,t+1} = \mathbf{t}_2' \mathbf{x}_t + 2u_A (X_{2,2:BC,BC,t} - X_{3,3:A,BC,ABC,t}) + 2u_B (X_{2,3:A,C,AC,t} - X_{3,3:A,BC,ABC,t}) + 2u_C (X_{2,3:A,B,AB,t} - X_{3,3:A,BC,ABC,t}). \quad (18)$$

DISCUSSION

The main aim of this analysis was to derive and show clearly the relationship between multi-locus identities and moments of linkage disequilibrium and hence how to compute each most simply. The analysis here is restricted solely to the haploid model, but in a random mating population it becomes an increasingly accurate predictor for the diploid case for small recombination fractions and large population size (see Weir and Cockerham, 1974, for two-locus examples). The assumption of a random mating population is more critical than just in the use of the haploid model. In the absence of random mating, the changes in identity can not be described solely in terms of two-gene measures at each locus, and it is necessary to include more terms in the transition calculations (Weir and Cockerham, 1974). Further, the correspondence between identities and moments of linkage disequilibria also become more complicated.

Predictions of multi-locus identity are needed in, for example, prediction of the origins of haplotypes in pedigrees and of associating them with particular traits or disease conditions. While the method developed here is feasible for four loci it becomes unwieldy with more. Hernandez-Sanchez *et al.* (2004) showed how to predict probabilities of identity at three and four loci from those on two loci by a regression analysis. They used the exact results for two loci of Weir and Cockerham (1974) to obtain F_{AB} , F_{AC} and F_{BC} (e.g. $F_{AB} = F_{2,2:AB,AB}$) using the appropriate recombination fractions, and from these the covariances and regression coefficients of identity, e.g. $\text{cov}(F_A, F_C) = F_{AC} - F_A F_C$ and $\beta_{C,A} = \text{cov}(F_A, F_C) / \text{Var}(F_A) = (F_{AC} - F_A F_C) / [F_A(1 - F_A)]$. With these they computed the expected probability of identity at locus C given identity at A and B from a partial regression equation, and thus the three locus identity F_{ABC} . Hernandez-Sanchez *et al.* extended their regression method to predict identity at four loci in a two-step process from that predicted for three loci. The procedure gave good predictions as checked by simulation for three and four locus identity, although predictions were better for three loci and when the fitted loci were between the two reference loci. The results given here for three or four loci could be extended by standard multiple regression methods to make predictions for five or more loci, and thereby obtain accurate predictions of multi-locus identity for substantial lengths of the chromosome. They cannot however be used directly for other mating systems, whereas two-locus equations for different systems, including avoidance of selfing, are given by Weir and Cockerham (1974) and can be incorporated in the regression method. In a subsequent paper we shall consider other ways to predict multi-locus from two-locus identity.

Whilst it is usual to describe association between loci in terms of linkage disequilibria or derived statistics such as the correlation of gene frequencies, it can, for example, also be described in terms of observed multi-locus heterozygosities. For example, the ratio of non-identities $X_{2,2:AB,AB} / X_A X_B$ is also equal to the ratio of expected heterozygosity at a pair of loci relative to that of independent loci. The numbers of heterozygous loci have, for example, been used by Brown *et al.* (1980) to describe multi-locus association and by Weir *et al.* (2004), albeit less directly, as the number of multi-locus genotypes to identify regions of differential recombination rates in the human genome. Further, an important use of data on genetic markers is to compute relationships among pairs of individuals in non-pedigreed populations. This has been based on independent loci, but increasingly as we get more information on closely linked markers, relationships among linked loci will provide further information (Chapman and Thompson, 2002; Hu, 2005).

Analyses of population structure from samples of loci are based on the assumption that the genetic distance between them or from their ancestral population (F) is the same throughout the genome, apart from local sampling. Analyses of potential selective sweeps of human HapMap data show, however, that there are substantial differences among regions of the

genome in the levels of population dispersion (Weir et al., 2005), which should be quantified in multi-locus terms. The relationships between them enable multi-locus identity to be inferred from moments of linkage disequilibria which have known sampling properties.

Acknowledgments

This work was supported in part by NIH Grant GM 45344. We are grateful to Jules Hernández-Sánchez for helpful comments on an earlier draft.

References

- Brown AHD, Feldman MW, Nevo E. Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 1980;96:523–536. [PubMed: 17249067]
- Chapman NH, Thompson EA. The effect of population history on the lengths of ancestral chromosome segments. *Genetics* 2002;162:449–458. [PubMed: 12242253]
- Hernández-Sánchez J, Haley CS, Woolliams JA. On the prediction of simultaneous inbreeding coefficients at multiple loci. *Genet Res* 2004;83:113–120. [PubMed: 15219156]
- Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* 2003;13:635–643. [PubMed: 12654718]
- Hill WG. Disequilibrium among several linked neutral genes in finite population. II Variances and covariances of disequilibria. *Theor Pop Biol* 1974;6:184–198.
- Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theor Appl Genet* 1968;38:226–231.
- Hill WG, Weir BS. Variances and covariances of squared linkage disequilibria in finite populations. *Theor Pop Biol* 1988;33:54–78. [PubMed: 3376052]
- Hu XS. Estimating the correlation of pairwise relatedness along chromosomes. *Heredity* 2005;94:338–346. [PubMed: 15354191]
- Kimura M, Crow JF. The number of alleles that can be maintained in a finite population. *Genetics* 1964;49:725–738. [PubMed: 14156929]
- Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. *Genome Res* 2005;15:1468–1476. [PubMed: 16251456]
- Weir BS, Cockerham CC. Group inbreeding with two linked loci. *Genetics* 1969;63:711–742. [PubMed: 5399257]
- Weir BS, Cockerham CC. Behavior of pairs of loci in finite monoecious populations. *Theor Pop Biol* 1974;6:323–354. [PubMed: 4460261]
- Weir BS, Hill WG, Cardon LR. Allelic association patterns for a dense SNP map. *Genetic Epidemiology* 2004;24:1–9.

Appendix

APPENDIX 1: Definitions of Vector x of Non-identity Coefficients for Three Loci, and Correspondence to Identity Coefficients

To shorten the equations, let $G = 1 - F_A - F_B - F_C$

$$\begin{aligned}
x_1 &= X_{3,2:ABC,ABC} = G + F_{2,2:AB,AB} + F_{2,2:AC,AC} + F_{2,2:BC,BC} - F_{3,2:ABC,ABC} \\
x_2 &= X_{3,3:A,BC,ABC} = G + F_{2,3:A,B,AB} + F_{2,3:A,C,AC} + F_{2,2:BC,BC} - F_{3,3:A,BC,ABC} \\
x_3 &= X_{3,3:B,AC,ABC} = G + F_{2,3:A,B,AB} + F_{2,3:B,C,BC} + F_{2,2:AC,AC} - F_{3,3:B,AC,ABC} \\
x_4 &= X_{3,3:C,AB,ABC} = G + F_{2,3:A,C,AC} + F_{2,3:B,C,BC} + F_{2,2:AB,AB} - F_{3,3:C,AB,ABC} \\
x_5 &= X_{3,3:AB,AC,BC} = G + F_{2,3:A,B,AB} + F_{2,3:A,C,AC} + F_{2,3:B,C,BC} - F_{3,3:AB,AC,BC} \\
x_6 &= X_{3,4:A,B,C,ABC} = G + F_{2,3:A,B,AB} + F_{2,3:A,C,AC} + F_{2,3:B,C,BC} - F_{3,4:A,B,C,ABC} \\
x_7 &= X_{3,4:A,A,BC,BC} = G + F_{2,4:A,A,B,B} + F_{2,4:A,A,C,C} + F_{2,2:BC,BC} - F_{3,4:A,A,BC,BC} \\
x_8 &= X_{3,4:B,B,AC,AC} = G + F_{2,4:A,A,B,B} + F_{2,4:B,B,C,C} + F_{2,2:AC,AC} - F_{3,4:B,B,AC,AC} \\
x_9 &= X_{3,4:C,C,AB,AB} = G + F_{2,4:A,A,C,C} + F_{2,4:B,B,C,C} + F_{2,2:AB,AB} - F_{3,4:C,C,AB,AB} \\
x_{10} &= X_{3,4:B,C,AB,AC} = G + F_{2,4:B,B,C,C} + F_{2,3:A,B,AB} + F_{2,3:A,C,AC} - F_{3,4:B,C,AB,AC} \\
x_{11} &= X_{3,4:A,C,AB,BC} = G + F_{2,4:A,A,C,C} + F_{2,3:A,B,AB} + F_{2,3:B,C,BC} - F_{3,4:A,C,AB,BC} \\
x_{12} &= X_{3,4:A,B,AC,BC} = G + F_{2,4:A,A,B,B} + F_{2,3:A,C,AC} + F_{2,3:B,C,BC} - F_{3,4:A,B,AC,BC} \\
x_{13} &= X_{3,5:A,A,B,C,BC} = G + F_{2,4:A,A,B,B} + F_{2,4:A,A,C,C} + F_{2,3:B,C,BC} - F_{3,5:A,A,B,C,BC} \\
x_{14} &= X_{3,5:B,B,A,C,AC} = G + F_{2,4:A,A,B,B} + F_{2,4:B,B,C,C} + F_{2,3:A,C,AC} - F_{3,5:B,B,A,C,AC} \\
x_{15} &= X_{3,5:C,C,A,B,AB} = G + F_{2,4:A,A,C,C} + F_{2,4:B,B,C,C} + F_{2,3:A,B,AB} - F_{3,5:C,C,A,B,AB} \\
x_{16} &= X_{3,6:A,A,B,B,C,C} = G + F_{2,4:A,A,B,B} + F_{2,4:A,A,C,C} + F_{2,4:B,B,C,C} - F_{3,6:A,A,B,B,C,C}
\end{aligned}$$

APPENDIX 2: Algorithm to Compute Coefficients of Elements of $\mathbf{L}$

- i. Define vector $\mathbf{y}$ of dimension h as the haplotypes sampled in the parental array, with elements expressed in binary form, e.g. as follows for two loci: $AB = 11, A = 10, B = 01$. For example, if the parental haplotypes are $AB, AB, \mathbf{y}' = (11, 11)$.
- ii. Similarly define vector $\mathbf{z}$ of dimension h' as the haplotypes sampled in the progeny array. For example, if the progeny haplotypes are $A, B, AB, \mathbf{z}' = (10, 01, 11)$.
- iii. Construct matrix Λ , dimension h' by h , with elements which define whether the progeny haplotype can be obtained from sampling the parental haplotype:
If z_i & $y_j = z_i$, where & denotes the logical “and” operation, then the element is $\lambda_{ij} = z_i$; otherwise $\lambda_{ij} = 0$.
- iv. (iv) Consider all possible matrices Δ , dimension h' by h , with elements $\delta_{ij} = 0$ or 1 and $\sum_{j=1}^h \delta_{ij} = 1$ for all i , i.e. one element in each row takes the value 1, the rest 0, with the non-zero element defining the parental haplotype as source of the progeny haplotype.
- v. If $\sum_{i=1}^{h'} \delta_{ij} \lambda_{ij} = y_j$ for all j , this implies that the sampling leads to non-identical haplotypes. Count the number, K , of all matrices Δ that satisfy this criterion.
Count number of pairs, c , of identical haplotypes among the parents; for example (see text following Equation 4) if $\mathbf{y}' = (11, 11)$, then $c = 1$. Compute coefficient k in $\mathbf{L}$ as $K/2^c$.

Table 1

Non-identity coefficients and moments for three loci, elements of the upper triangular matrix **Q** defining their relationship, and diagonal elements of **R**, defining non-recombination* among moments. Rows of **Q** should be multiplied by the coefficient κ .

Moment	κ	Elements of Q																Non-identity	Element of R
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
1	D_{ABC}^2	-1	1	-2	-2	0	4	1	1	1	2	2	-4	-4	-4	4		$X_{3,2}ABC,ABC$	$(1-r_{ABC})^2$
2	$\tau_A D_{BC} D_{ABC}$	1	1	0	0	-1	-1	0	0	0	-1	-1	3	1	1	-2		$X_{3,3}A,BC,ABC$	$(1-r_{BC})(1-r_{ABC})$
3	$\tau_B D_{AC} D_{ABC}$	1	1	0	0	-1	0	-1	0	-1	0	-1	1	3	1	-2		$X_{3,3}B,AC,ABC$	$(1-r_{AC})(1-r_{ABC})$
4	$\tau_C D_{AB} D_{ABC}$	1	1	0	-1	0	0	-1	0	-1	-1	0	1	1	3	-2		$X_{3,3}C,AB,ABC$	$(1-r_{AB})(1-r_{ABC})$
5	$D_{AB} D_{AC} D_{BC}$	-1			1	0	0	0	0	-1	-1	-1	1	1	1	-1		$X_{3,3}AB,AC,BC$	$(1-r_{AB})(1-r_{AC})(1-r_{BC})$
6	$\tau_A \tau_B \tau_C D_{ABC}$	1				1	0	0	0	0	0	0	-1	-1	-1	2		$X_{3,4}A,B,C,ABC$	$1-r_{ABC}$
7	$\pi_A D_{BC}^2$	1					1	0	0	0	0	0	-2	0	0	1		$X_{3,4}AA,BC,BC$	$(1-r_{BC})^2$
8	$\pi_B D_{AC}^2$	1						1	0	0	0	0	0	-2	0	1		$X_{3,4}BB,AC,AC$	$(1-r_{AC})^2$
9	$\pi_C D_{AB}^2$	1							1	0	0	0	0	0	-2	1		$X_{3,4}CC,AB,AB$	$(1-r_{AB})^2$
10	$\tau_B \tau_C D_{AB} D_{AC}$	-1								1	0	0	0	-1	-1	1		$X_{3,4}BC,AB,AC$	$(1-r_{AB})(1-r_{AC})$
11	$\tau_A \tau_C D_{AB} D_{BC}$	-1									1	0	-1	0	-1	1		$X_{3,4}AC,AB,BC$	$(1-r_{AB})(1-r_{BC})$
12	$\tau_A \tau_B D_{AC} D_{BC}$	-1										1	-1	-1	0	1		$X_{3,4}AB,AC,BC$	$(1-r_{AC})(1-r_{BC})$
13	$\pi_A \tau_B \tau_C D_{BC}$	1											1	0	0	-1		$X_{3,5}AA,B,C,BC$	$1-r_{BC}$
14	$\pi_B \tau_A \tau_C D_{AC}$	1												1	0	-1		$X_{3,5}BB,A,C,AC$	$1-r_{AC}$
15	$\pi_C \tau_A \tau_B D_{AB}$	1														1	-1	$X_{3,5}C,C,AB,AB$	$1-r_{AB}$
16	$\pi_A \pi_B \pi_C$	1															1	$X_{3,6}AA,B,B,C,C$	1

* If loci A, B and C have that order on the chromosome and there is no interference $(1 - r_{AC}) = (1 - r_{AB})(1 - r_{BC}) + r_{AB}r_{BC}$ and $(1 - r_{ABC}) = (1 - r_{AB})(1 - r_{BC})$

Table 2

Elements of **L** for three loci. To reduce expressions, $v = 2N$. All elements should be multiplied by $\xi_1^2 = [1 - i/(2N)]$. Other terms are $\xi_i^2 = [1 - i/(2N)]$, $i = 2, \dots, 5$.

Non-identity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 $X_{3,2}ABCABC$	1															
2 $X_{3,3}ABACBC$	$1/v$	ξ_2^2														
3 $X_{3,3}BACABC$	$1/v$	0	ξ_2^2													
4 $X_{3,3}CABABC$	$1/v$	0	0	ξ_2^2												
5 $X_{3,3}ABACBC$	0	0	0	0	ξ_2^2											
6 $X_{3,4}ABACBC$	$1/v^2$	ξ_2/v	ξ_2/v	ξ_2/v	0	$\xi_2^2 v_3$										
7 $X_{3,4}ABCBBC$	$2/v^2$	$4\xi_2/v$	0	0	0	0	$\xi_2^2 v_3$									
8 $X_{3,4}BABACAC$	$2/v^2$	0	$4\xi_2/v$	0	0	0	0	$\xi_2^2 v_3$								
9 $X_{3,4}CCABAB$	$2/v^2$	0	0	$4\xi_2/v$	0	0	0	0	$\xi_2^2 v_3$							
10 $X_{3,4}BCABAC$	$1/v^2$	0	ξ_2/v	ξ_2/v	ξ_2/v	0	0	0	0	$\xi_2^2 v_3$						
11 $X_{3,4}ACABBC$	$1/v^2$	ξ_2/v	0	ξ_2/v	ξ_2/v	0	0	0	0	0	$\xi_2^2 v_3$					
12 $X_{3,4}ABACBC$	$1/v^2$	ξ_2/v	ξ_2/v	0	ξ_2/v	0	0	0	0	0	0	$\xi_2^2 v_3$				
13 $X_{3,5}AABCBBC$	$2/v^3$	$4\xi_2/v^2$	$2\xi_2/v^2$	$2\xi_2/v^2$	$2\xi_2/v^2$	$2\xi_2^2 v_3/v$	$\xi_2^2 v_3/v$	0	0	0	$2\xi_2^2 v_3/v$	$2\xi_2^2 v_3/v$	$\xi_2^2 v_3^2 v_4$			
14 $X_{3,5}BABACAC$	$2/v^3$	$2\xi_2/v^2$	$4\xi_2/v^2$	$2\xi_2/v^2$	$2\xi_2/v^2$	$2\xi_2^2 v_3/v$	0	$\xi_2^2 v_3/v$	0	$2\xi_2^2 v_3/v$	0	$2\xi_2^2 v_3/v$	0	$\xi_2^2 v_3^2 v_4$		
15 $X_{3,5}CCABAB$	$2/v^3$	$2\xi_2/v^2$	$2\xi_2/v^2$	$4\xi_2/v^2$	$2\xi_2/v^2$	$2\xi_2^2 v_3/v$	0	0	$\xi_2^2 v_3/v$	$2\xi_2^2 v_3/v$	$2\xi_2^2 v_3/v$	0	0	0	$\xi_2^2 v_3^2 v_4$	
16 $X_{3,6}AABBCBC$	$4/v^4$	$8\xi_2/v^3$	$8\xi_2/v^3$	$8\xi_2/v^3$	$8\xi_2^2 v_3/v^3$	$8\xi_2^2 v_3/v^2$	$2\xi_2^2 v_3/v^2$	$2\xi_2^2 v_3/v^2$	$2\xi_2^2 v_3/v^2$	$8\xi_2^2 v_3/v^2$	$8\xi_2^2 v_3/v^2$	$8\xi_2^2 v_3/v^2$	$4\xi_2^2 v_3^2 v_4/v$	$4\xi_2^2 v_3^2 v_4/v$	$4\xi_2^2 v_3^2 v_4/v$	$\xi_2^2 v_3^2 v_4^2 v_5$