Prediction of multilocus identity-by-descent

DOI: 10.1534/genetics.107.074344

Prediction of Multilocus Identity-by-Descent

William G. Hill[1] and Jules Hernández-Sánchez

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom

Manuscript received April 10, 2007
Accepted for publication May 14, 2007

ABSTRACT

Previous studies have enabled exact prediction of probabilities of identity-by-descent (IBD) in random-mating populations for a few loci (up to four or so), with extension to more using approximate regression methods. Here we present a precise predictor of multiple-locus IBD using simple formulas based on exact results for two loci. In particular, the probability of non-IBD X_{ABC} at each of ordered loci *A*, *B*, and *C* can be well approximated by $X_{ABC} = X_{AB}X_{BC}/X_B$ and generalizes to $X_{123\ldots k} = X_{12}X_{23\ldots}X_{k-1,k}/X^{k-2}$, where X is the probability of non-IBD at each locus. Predictions from this chain rule are very precise with population bottlenecks and migration, but are rather poorer in the presence of mutation. From these coefficients, the probabilities of multilocus IBD and non-IBD can also be computed for genomic regions as functions of population size, time, and map distances. An approximate but simple recurrence formula is also developed, which generally is less accurate than the chain rule but is more robust with mutation. Used together with the chain rule it leads to explicit equations for non-IBD in a region. The results can be applied to detection of quantitative trait loci (QTL) by computing the probability of IBD at candidate loci in terms of identity-by-state at neighboring markers.

IN a recent article formulas for computing probabilities of identity-by-descent (IBD) at multiple loci in random-mating populations were obtained (Hill and Weir 2007) by extending methods of Weir and Cockerham (1969, 1974) for a haploid model. Recurrence equations were presented for multilocus non-IBD, from which IBD can be computed; but the number of terms involved quickly becomes impracticably large to compute. For example, prediction of nonidentity at three loci requires recurrence equations for a total of 16 non-IBD measures defined for loci sampled on two, three, four, five, and six different haplotypes. For four loci the number of measures rises to 139 (Hill and Weir 2007). Hernández-Sánchez *et al.* (2004) have developed approximations based on multiple regression to compute IBD at multiple loci from that at two loci, but the formulas become increasingly less tractable and accurate as the number of loci increases.

Here we develop a straightforward method (the chain rule) for predicting probabilities of multilocus non-IBD, and thus IBD, which uses exact results only on two-locus non-IBD probabilities. Assuming a known population history, this predictor can be very precise for many loci and can enable IBD for a whole chromosome region to be computed. We also develop simple approximate recurrence equations that are generally less precise, except in the presence of mutation.

An application of multiple-locus extensions of Wright's inbreeding coefficient is in gene or quantitative trait loci (QTL) mapping on the basis of the association between phenotypic similarity of individuals and shared IBD at a particular genomic region (Meuwissen *et al.* 2002; Hernández-Sánchez *et al.* 2006). The magnitude of IBD at a QTL is computed from the identity-by-state (IBS) of neighboring marker loci, but to do so it is necessary to know the extent of joint IBD across the QTL and markers relative to some reference population.

METHODS

Background: *Definitions:* Let *A*, *B*, and *C* be three loci located in that order on a chromosome, and denote by F_A, F_{AB}, and F_{ABC} probabilities of IBD at locus *A*, loci *A* and *B*, and loci *A*, *B*, and *C*, respectively. Similarly, let X_A, X_{AB}, X_{ABC} denote the probabilities of non-IBD at the corresponding loci; *i.e.*, X_{AB} is the probability that neither *A* nor *B* is IBD. These quantities refer to the case where identity is examined at all loci on a pair of haplotypes. There are other measures when considering more than two haplotypes. For example, two IBD loci can also be sampled in three and four different haplotypes (Weir and Cockerham 1974).

The IBD and non-IBD probabilities are related at any generation by, for example,

[1]*Corresponding author:* Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, W. Mains Rd., Edinburgh, EH9 3JT, United Kingdom. E-mail: w.g.hill@ed.ac.uk

$$F_A = 1 - X_A, \tag{1a}$$

$$F_{AB} = 1 - X_A - X_B + X_{AB} \tag{1b}$$

$$F_{ABC} = 1 - X_A - X_B - X_C + X_{AB} + X_{AC} + X_{BC} - X_{ABC} \tag{1c}$$

(Hill and Weir 2007). In general, k-locus IBD and non-IBD measures are related as

$$F_{1\ldots k} = 1 - \sum_{i=1}^{k} X_i + \sum_{i<j}^{k} X_{ij} - \ \ldots\ + (-1)^k X_{1\ldots k}, \tag{1d}$$

where $F_{1\ldots k}$ and $X_{1\ldots k}$ denote, respectively, the probabilities of IBD and non-IBD for all ordered loci 1 to k on two haplotypes. Equation 1d is an example of the inclusion–exclusion principle. The multilocus measures used here, which are extensions of the Wright–Malécot definitions of inbreeding such that pairs of genes at each of two loci may be IBD even though identity at each locus traces back to different ancestors, differ from the "chromosome segment homozygosity" defined by Hayes *et al.* (2003), which defines identity of haplotypes back to a common ancestral haplotype *without* intervening recombination.

The following parameters are also used and assumptions made. All genes in the founder population (generation $t = 0$) are assumed to be non-IBD at all loci; *i.e.*, $X_{A(0)} = X_{AB(0)} = \ldots = 1$. The effective population size is N diploids ($2N$ genes) and is constant over generations. There is random mating (with or without selfing, as specified) and there is no selection at or near the identified loci. The recombination fraction between loci A and B is r_{AB} and there is no crossover interference. The map length of a region of chromosome is denoted l (in morgans). The rate of mutation at each locus is u, where any mutant gene is assumed to be non-IBD to all existing genes at that locus in the population (*i.e.*, infinite-alleles model), and the rate of migration is m, where migrant haplotypes come from an infinitely large and unrelated population, such that in the generation following migration, genotypes comprising one or two migrant haplotypes are non-IBD at all loci. Also we define $R_{AB} = 4Nr_{AB}$, $L = 4Nl$, $U = 4Nu$, and $M = 4Nm$.

Exact method: By extending methods of Weir and Cockerham (1974), Hill and Weir (2007) give an exact way to predict probabilities of multilocus non-IBD, and from that IBD, by transition matrix iteration over generations, assuming a haploid model. Although the method is feasible for four loci it rapidly becomes unwieldy with more, so we review and consider alternative methods to predict identity for multiple loci from results for fewer loci, *e.g.*, F_{ABC} from F_{AB} and F_{BC}.

Regression method: Hernández-Sánchez *et al.* (2004) proposed a regression analysis to predict probabilities of identity at three and four loci from those on two loci given by Weir and Cockerham (1974). For example, F_{AB}, F_{AC}, and F_{BC} are computed each generation, and from these the regression coefficients of identity at locus B given identity at A are calculated; for example, $\beta_{B.A} = \mathrm{Cov}(F_A, F_B)/\mathrm{Var}(F_A) = (F_{AB} - F_A F_B)/[F_A(1 - F_A)]$. Consequently the conditional probability $F_{B|AC}$ of identity at locus B given identity at A and C is predicted from a partial regression equation including terms in $\beta_{B.A}$ and $\beta_{B.C}$, and thus the three-locus identity $F_{ABC} = F_{B|AC}F_{AC}$ (Hernández-Sánchez *et al.* 2004, Equation 3). On the basis of this three-locus prediction, but still using exact results for only two loci, Hernández-Sánchez *et al.* extended the regression method to predict identity at four loci in a two-step process. The method gave good predictions for three- and four-locus identity obtained by simulation, for example, for three- and four-locus inbreeding coefficients in random-mating diploid populations for values of $R = 4$ between adjacent loci (*e.g.*, $N = 10$, $r = 0.1$) and 8 ($N = 20$, $r = 0.1$). Predictions were poorer for four loci or if the conditional identities were predicted for loci outside (C from A and B) rather than between the two reference loci (B from A and C). Their method could be extended by standard multiple-regression methods to make more precise predictions for five or more loci using the results given by Hill and Weir (2007) for three or four loci, but computation of the partial regression coefficients rapidly becomes unwieldy as the number of loci increases.

Conditional (chain-rule) method for multilocus non-IBD: *Principle:* The regression method of Hernández-Sánchez *et al.* (2004) does not utilize the ordering of the loci on the chromosome directly, *i.e.*, the fact that for loci ordered $A, B, C, \ldots$, a recombination between A and B usually also implies a recombination between A and C. This suggests alternative methods for predicting the multilocus (non)identities by utilizing such information. Therefore a "natural" predictor of the three-locus nonidentity is to approximate the joint probability $X_{ABC} = X_{AB}X_{C|AB}$ by $\hat{X}_{ABC} = X_{AB}X_{C|B}$, where $X_{C|B} = X_{BC}/X_B$ is the conditional probability of nonidentity at locus C given nonidentity at the adjacent locus B. This implies that knowledge of IBD probability at the more distant A adds no further information and gives the predictor

$$\hat{X}_{ABC} = X_{AB}X_{BC}/X_B. \tag{2}$$

In the absence of mutation it turns out that (2) is remarkably precise, as shown by examples in Table 1 in which predictions of X_{ABC} are compared to exact values (Hill and Weir 2007), with most predictions deviating $<0.1\%$ in absolute terms and 1% in relative terms. These are better than those based on the regression method of Hernández-Sánchez *et al.* (2004), particularly at higher values of R. For example, for $N = t = 100$, the predictions from the regression method are 0.6058 (*i.e.*, exact), 0.5796, 0.5159, and 0.3802 (an absolute deviation of almost 1% in X_{ABC}) for $R_{AB} = R_{BC} = 0$, $\frac{1}{4}$, 1, and 4, respectively (*cf.* Table 1). It is important to note that, unlike in the regression method,

TABLE 1

Exact (*E*) and predicted (*P*, from Equation 2) values of three-locus nonidentity X_{ABC} with a haploid model and no mutation or migration for $N = 100$ and a range of $R_{AB} = 4Nr_{AB}$ and $R_{BC} = 4Nr_{BC}$

R_{AB}, R_{BC}:	0, 0:	$\frac{1}{4}$, $\frac{1}{4}$		$\frac{1}{4}$, 1		1, 1		1, 4		4, 4	
t	E, P	E	P	E	P	E	P	E	P	E	P
25	0.8822	0.8792	0.8792	0.8748	0.8748	0.8705	0.8704	0.8554	0.8553	0.8406	0.8404
50	0.7783	0.7683	0.7683	0.7545	0.7545	0.7410	0.7409	0.6998	0.6996	0.6611	0.6606
100	0.6058	0.5794	0.5794	0.5457	0.5456	0.5141	0.5138	0.4375	0.4369	0.3727	0.3715
150	0.4715	0.4322	0.4321	0.3856	0.3854	0.3443	0.3438	0.2616	0.2608	0.1992	0.1979
200	0.3670	0.3202	0.3201	0.2689	0.2687	0.2261	0.2255	0.1533	0.1526	0.1044	0.1033
300	0.2223	0.1740	0.1739	0.1281	0.1279	0.0945	0.0940	0.0513	0.0509	0.0281	0.0276
400	0.1347	0.0940	0.0939	0.0603	0.0602	0.0388	0.0386	0.0170	0.0168	0.0075	0.0073

Predictions are exact if $R_{AB} = 0$ or $R_{BC} = 0$.

the ordering of the loci is important for the chain rule; for example, $X_{AB}X_{AC}/X_A$ is a very poor predictor of X_{ABC}.

In view of the high predictive value of Equation 2, unsurprisingly the natural extension to four loci

$$\hat{X}_{ABCD} = X_{AB}X_{BC}X_{CD}/(X_BX_C)$$

is also a good predictor (results not shown). For k loci, this "chain-rule" predictor of multilocus nonidentity $\hat{X}_{12\ldots k}$, from adjacent two-locus $X_{i,i+1}$ and one-locus nonidentities $X \equiv X_i$, which are assumed to be the same at each locus, is

$$\hat{X}_{12\ldots k} = \prod_{i=1}^{k-1} X_{i,i+1}/X^{k-2} \tag{3}$$

and for equally spaced markers

$$\hat{X}_{12\ldots k} = X(X_{1,2}/X)^{k-1}. \tag{4}$$

Examples of predictions of multilocus nonidentity computed from Equation 4 are compared with results obtained by stochastic simulation using Wright–Fisher sampling in Figure 1, where it is seen that there is excellent correspondence for these examples in which there is a population of constant size with no mutation or migration. The method can be used for any mating system, *e.g.*, a haploid (Table 1) or a diploid with selfing included (Figure 1), for nonconstant population size, and in the presence of migration or mutation. As we show subsequently, of these only mutation causes significant errors.

Regional non-IBD: Using Equation 4, the probability $X(l)$ that *all* sites in a region of length l morgans are non-IBD can be predicted by dividing it into very many, say $s = k - 1$, small equally sized segments and taking the limit

$$\hat{X}(l) = X \lim_{s\to\infty}\{[X_{(l/s)}/X]^s\}, \tag{5}$$

where $X_{(l/s)}$ denotes the probability of joint non-IBD of a pair of markers l/s map units apart. This probability approaches that for loci with recombination fraction $r = l/s$ as $s \to \infty$. Equation 5 can therefore be written as

$$\begin{aligned}\hat{X}(l) &= X \lim_{r\to 0}\left\{\left[\frac{X_{(r)}}{X}\right]^{l/r}\right\} \\ &= X \lim_{r\to 0}\left\{\left[1 + \left(\frac{r}{X}\right)\left(\frac{X_{(r)} - X}{r}\right)\right]^{l/r}\right\}.\end{aligned}$$

The limits are

$$\lim_{r\to 0} X_{(r)} = X \quad \text{and} \quad \lim_{r\to 0}\left(\frac{X_{(r)} - X}{r}\right) = \frac{dX_{(r)}}{dr}|_{r=0},$$

the derivative of the two-locus non-IBD probability with respect to the recombination fraction r between the loci evaluated at $r = 0$, and it is convenient to define

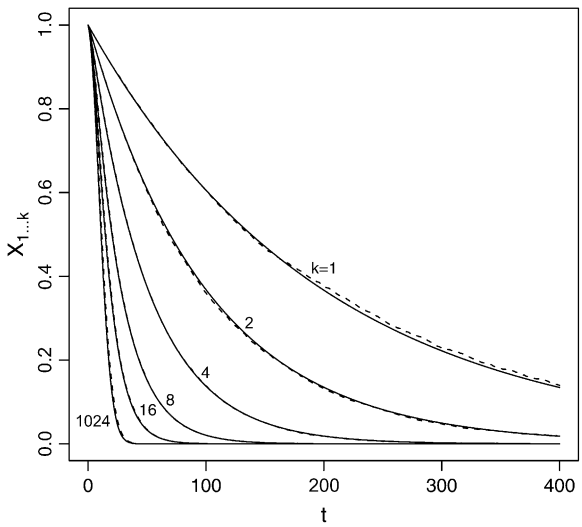

FIGURE 1.—Probabilities of non-IBD at multiple loci ($X_{1\ldots k}$) over generations (t). Predictions using Equation 4 are shown as solid lines and computer simulations (1000 replicates) as dashed lines for a monoecious diploid population. $N = 100$, with 100 cM between outside markers, with equal spacing of markers within the region.

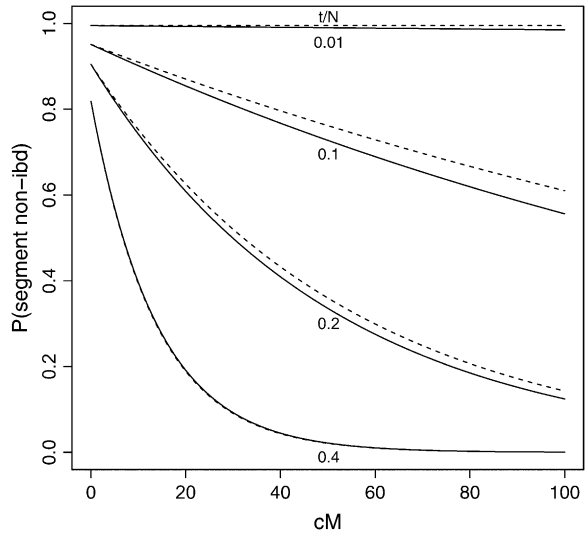

FIGURE 2.—Regional non-IBD for an increasing genome length (in centimorgans) and specific generations/population size (t/N) in a monoecious diploid population. The solid lines were obtained using Equation 7, where the derivative was obtained numerically, and dashed lines using Equation 15, where a closed formula for the derivative was obtained after approximating transition matrices with the linearization in Equation 13.

$$\gamma = \frac{dX_{(r)}}{Xdr}|_{r=0} \equiv \frac{d\log X_{(r)}}{dr}|_{r=0}. \tag{6}$$

Using the definition of the exponential function, Equation 5 reduces to

$$\hat{X}(l) = X \lim_{r\to 0}(1 + r\gamma)^{l/r} = X\exp(l\gamma). \tag{7}$$

Equation 7 can also be derived from the incremental change in $X(l)$ as l is increased by an infinitesimally small amount and integrating the resultant "growth" equation.

The derivatives in Equation 6 (which are negative) can be evaluated numerically at any generation by iteration of the transition matrix for a small value of r and computing γ as $[X_{(r)}/X - 1]/r$. To ensure there are no errors due to rounding or inclusion of higher-order terms, consistency can be checked using a range of values of r (we found consistency for r between 10^{-4} and 10^{-7}). Equation 7 can also be expressed in terms of $L = 4Nl$ if the derivative is similarly rescaled. Examples are given in Figure 2. In these examples mutation is assumed to be absent. Indeed, to include mutation it would be necessary to define a mutation rate per unit map length as a continuous function, and in view of the limited accuracy of the chain rule in the presence of mutation, we do not consider this extension to the analysis.

Multilocus IBD: F_{ABC}, F_{ABCD}, etc., can be predicted from Equations 1–3 directly. For example, from Equations 1c and 2

$$\hat{F}_{ABC} = 1 - X_A - X_B - X_C + X_{AB} + X_{AC} + X_{BC} - X_{AB}X_{BC}/X_B. \tag{8}$$

A similar simple conditional argument to that used to obtain Equation 2 would lead to a different prediction $\tilde{F}_{ABC} = F_{AB}F_{BC}/F_B$. This prediction equation for $\tilde{F}_{ABC}$ does not hold because the conditional probability $F_{BC|AB}$ does not equal $F_{BC|B}$ as the regions AB and BC may be IBD for different founder haplotypes. In contrast, replacing non-IBD for IBD coefficients using Equations 1a and 1b and rearranging Equation 8 gives

$$\hat{F}_{ABC} = F_{AC} - (F_A - F_{AB})(F_C - F_{BC})/(1 - F_B). \tag{9}$$

Thus for the chain rule in terms of IBD, the term on the right of Equation 9 is the overall probability of identity at A and C less situations in which there is nonidentity at B but identity at A and C.

Prediction of k-locus IBD from non-IBD using Equation 1d involves $2^k - 1$ terms, and becomes computationally impractical for evaluating IBD over multiple sites (*e.g.*, 6 hr of computation for $k = 30$ with an ~1 Mflop computer). There is, however, a very efficient algorithm for adding successive loci in the chain. Note that

$$F_{AB} - F = X_{AB} - X$$

$$F_{ABC} - F_{AB} = -X + X_{AC} + X_{BC} - X_{ABC}$$

from Equations 1b and 1c, and from Equation 2

$$\begin{aligned}\hat{F}_{ABC} - F_{AB} &= X_{AC} - X + X_{BC} - X_{AB}X_{BC}/X \\ &= F_{AC} - F - (X_{BC}/X)(F_{AB} - F).\end{aligned}$$

Similarly

$$\hat{F}_{ABCD} - \hat{F}_{ABC} = F_{AD} - F - (X_{BD}/X)(F_{AB} - F) - (X_{CD}/X)(\hat{F}_{ABC} - F_{AB}).$$

Let $\Delta_1 \equiv F$, $\Delta_2 \equiv F_{12} - F$, $\Delta_3 \equiv \hat{F}_{123} - F_{12}$, and, in general, $\Delta_i \equiv \hat{F}_{1\ldots i} - \hat{F}_{1\ldots i-1}$. Then

$$\Delta_k = X_{1k} - X - \frac{1}{X}\sum_{i=2}^{k-1}(X_{ik}\Delta_i) \tag{10}$$

for $k > 2$, and

$$\hat{F}_{1\ldots k} = \sum_{i=1}^{k}\Delta_i. \tag{11}$$

Equation 10, in which one locus is added at each iteration to compute the change in multilocus IBD, involves k terms when the kth locus is added and thus a total of $\frac{1}{2}k(k-1)$ in all. This contrasts with the $2^k - 1$ needed in Equation 1d, such that the computation is feasible up to thousands of loci (*e.g.*, 10 sec computation for $k =$ 2000 with the same computer). To predict the probability of IBD on a region assuming equal recombination

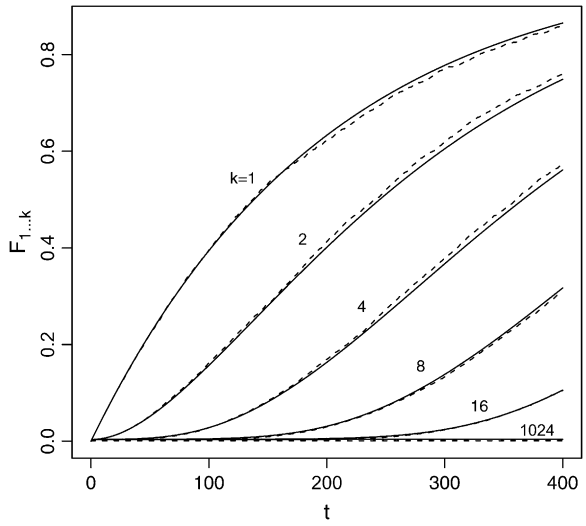

FIGURE 3.—Multiple-locus IBD probabilities ($F_{1\ldots k}$) over generations (t). Solid lines were obtained with Equation 11 and dashed lines with computer simulations averaging 10,000 replicates in a monoecious diploid population. $N = 100$, with 100 cM between outside markers with equal spacing.

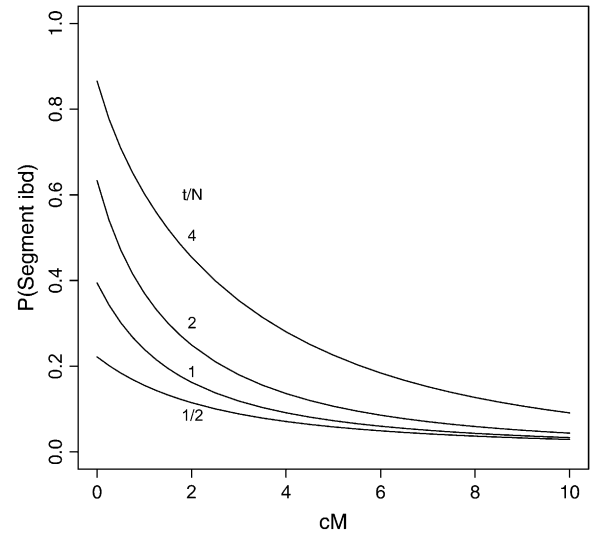

FIGURE 4.—Regional IBD for an increasing genome length (in centimorgans) and specific t/N using Equation 11 with $k = 1000$ loci in a monoecious diploid population. Note that as larger regions are less saturated with loci than shorter ones, *e.g.*, 1000 loci within 1 cM *vs.* 1000 loci within 10 cM, predictions are likely to be more accurate for short than for long regions.

fractions between consecutive loci, it requires the evaluation only of $k-1$ values of X_{ik}, $i = 1, \ldots, k$, and it is also possible to predict regional IBD simply by estimating IBD for a very large number of sites.

A comparison between predictions of multilocus IBD from simulation and use of Equation 11 is given in Figure 3 for a population of constant size in the absence of mutation or migration. In view of the excellent predictions of non-IBD shown in Figure 1, for example, the fit of IBD is to be expected. Results for regional IBD are given for a wider range of parameters in Figure 4. Figures 3 and 4 also show how slowly the multilocus IBD increases with generation if many loci are considered, which implies that there can be small regions of the genome non-IBD even when most nearby sites are IBD.

Mutation, migration, and population bottlenecks: The chain-rule predictions of multilocus non-IBD probabilities, and of those from IBD, can be undertaken for any random-mating system (*e.g.*, in haploid and monoecious or dioecius diploid populations with/without avoidance of selfing) by using an appropriate transition matrix to compute the two-locus non-IBD (WEIR and COCKERHAM 1974).

Changes in population size, for example due to bottlenecks, are easily accounted for in the chain rule by using the appropriate value of N. Migration, under the continent-to-island model, increases the probability of non-IBD. This can be accounted for by replacing $\mathbf{x}_t$ by $\mathbf{x}_t + \mathbf{m}_t$ in Equation 7 of HILL and WEIR (2007) in the following vector [assuming for simplicity that the migration rate m is small so terms of $O(m^2)$ can be ignored],

$$\mathbf{m}_t = \begin{bmatrix} (1-2m)x_1 + 2m \\ (1-3m)x_2 + m + 2mx_1 \\ (1-4m)x_3 + 4mx_1 \end{bmatrix},$$

where x_i refers to the ith component of vector $\mathbf{x}_t$ of two-, three-, and four-haplotype coefficients of non-IBD. Continent-to-island and also multiple small-island models results in Figure 5 show an excellent level of prediction of simulated values of F_{ABC} using the chain rule, which also implies that the chain rule would apply within a nonrandom mating population, for example, incorporating avoidance of mating of relatives.

Mutation is the only evolutionary force considered in this study for which the chain rule gave poor predictions (Figure 5). Although the departure is small with realistic u ($<10^{-5}$) and few loci in small populations, it worsens as mutation rate (U) increases and as linkage becomes very tight as do predicted regional non-IBD and IBD probabilities (results not shown). A simple explanation of why mutation breaks the chain rule is that the adjacent locus does not contain all the information about the non-IBD status at a given locus (with mutation $X_{A|BC} > X_{A|B}$ and without $X_{A|BC} = X_{A|B}$). In the presence of mutation, information about the IBD status at locus C is useful in predicting the status at A because B may be non-IBD due to mutation and, except for this mutation, the chromosome region including A and C would be IBD. The chain rule assumes a first-order Markov chain that is violated in the presence of mutation because mutations occur independently of position (so that an IBD locus can be next to a mutant locus). In contrast, migration affects the whole string of loci, so a subset contains all the information (which will subsequently suffer recombination in the standard fashion). A formal analysis demonstrating the bias due to mutation on the chain rule for the case of completely linked loci is in the next section.

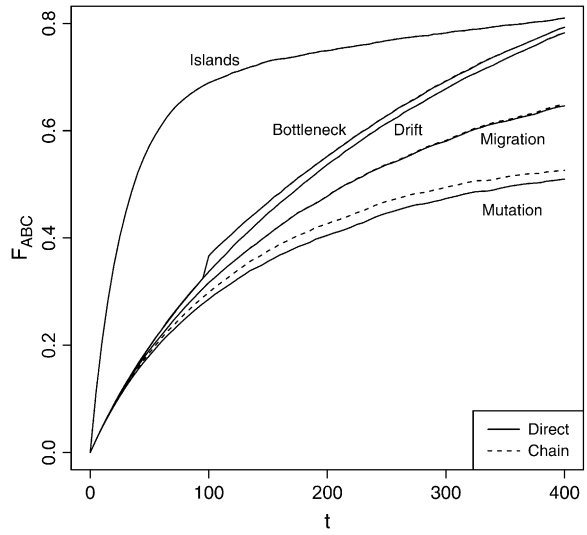

FIGURE 5.—Three-loci IBD probability (F_{ABC}) over generations (t) directly observed (solid line) and predicted using Equation 9 (dashed line), using computer simulation with 10,000 replicates in monoecious diploid populations. "Drift" denotes a haploid population with $N = 100$, $r_{AB} = 0.000625$, and $r_{BC} = 0.0025$. "Bottleneck" denotes the drift population with a bottleneck at $t = 100$ when $N = 10$ ($N = 100$ otherwise). "Migration" denotes the drift population receiving migrants from an infinitely large continent at a rate of $m = 0.000625$ haplotypes/generation. "Islands" denotes a five-islands migration model with total migration rate $m = 0.005$ and $N = 20$ per island. "Mutation" denotes the drift population with mutation rate of $u = 0.000625$/gene/generation.

Simple recurrence relations: *Principle:* The recurrence equations for non-IBD at two loci depend on terms in two-, three-, and four-haplotype probabilities in previous generations (WEIR and COCKERHAM 1974; HILL and WEIR 2007), although some may have very small coefficients in the recurrence equations. Numerical examples (not shown), however, indicate that these three- and four-haplotype identities are of similar magnitude to each other over quite a wide range of parameters, as are corresponding terms for three or more loci. Thus, if genes at one of the pair of loci A and B are sampled from different haplotypes, the probability of (non-)IBD depends little on whether the other A and B genes are sampled from one or two more haplotypes. In addition, if the two loci are not very tightly linked, the probability of two-locus (non-)IBD for genes sampled on four different haplotypes is slightly greater than $X_A X_B$, *i.e.*, the joint probability for two independent loci. Hence approximate recurrence predictions of non-IBD for two linked loci can be obtained solely by considering the probabilities on a pair of haplotypes and at individual loci. Similar arguments apply for more loci. Thus for two loci, this prediction of the two-locus non-IBD, X^*_{AB}, satisfies

$$X^*_{AB,t+1} \sim (1 - r_{AB})^2[1 - 1/(2N)X^*_{AB,t}] + [1 - (1 - r_{AB})^2][(1 - 1/(2N))]^2 X_{A,t}X_{B,t}. \quad (12)$$

If r_{AB} is small and N is large, (12) reduces to

$$X^*_{AB,t+1} \sim [1 - 2r_{AB} - 1/(2N)]X^*_{AB,t} + 2r_{AB}X_{A,t}X_{B,t}, \quad (13)$$

where $X_{A,t} = X_{B,t} = [1 - 1/(2N)]^t \sim \exp(-t/2N)$. The first term in Equations 12 and 13 denotes sampling two different and nonrecombined haplotypes that are non-IBD at both loci and the second denotes the sampling of recombinant gametes that are non-IBD at both loci. Equation 13 extends naturally to more loci, allowing for recombination between A and B and between B and C, and ignoring the chance of double recombinants. For example,

$$X^*_{ABC,t+1} \sim [1 - 2r_{AB} - 2r_{BC} - 1/(2N)]X^*_{ABC,t} + 2r_{AB}X_{A,t}X^*_{BC,t} + 2r_{BC}X^*_{AB,t}X_{C,t}. \quad (14)$$

The two-locus terms in Equation 14 can be predicted from Equation 13.

These are simple rather than necessarily precise predictors, but Equation 12 is exact if linkage is complete ($r_{AB} = 0$) or if loci are essentially independent ($R_{AB} \to \infty$). Evaluations using Equations 13 and 14 compared to exact methods (HILL and WEIR 2007) are illustrated in Figure 6. The method is seen to give reasonably good predictions for much of the range of $R\,(0, \frac{1}{4}, 1, 4, 16)$ and $t/N\,(0, 0.01, \ldots, 4)$. This is probably because the second term in Equations 12 and 13 makes a small contribution when r is very small, but $X_{A,t}X_{B,t}$ departs most from the actual probability when both loci are segregating; and when r is large, it makes a larger contribution but is a better approximation of $X_{A,t}X_{B,t}$. Other examples (not shown) indicate that the approximation behaves relatively poorly in small populations (say $N < 10$) than large (say $N > 50$) for the same value of $R = 4Nr$, which is expected since relative probabilities of random sampling from three rather than four haplotypes are more likely when N is small. Similar results can be obtained using Equation 14 or alternatively by joint use of Equations 12 and 13 for loci pairs AB and BC together with the three-locus chain prediction (Equation 2). It can also be shown that Equations 12, 13, and 14 are consistent: *i.e.*, replacing $X^*_{ABC,t}$ by $X^*_{AB,t}X^*_{BC,t}/X_{B,t}$ at t and $t + 1$ satisfies Equation 14 if terms of $O(<1/N)$ are excluded.

Regional non-IBD: Formulas for the two-locus non-IBD after integration with respect to time are derived in APPENDIX A in the case of no mutation or migration (Equation A1). This equation can then be used with the chain rule to obtain multilocus non-IBD and, as it can be differentiated explicitly (Equation A2), can be used with Equation 7, to obtain a remarkably simple formula for regional non-IBD (Equation A3),

$$\hat{X}(l) \sim X \exp[L(1 - e^{-t/2N} - t/2N)] \quad (15)$$

or, if $\Delta F = 1/2N$, then $\hat{X}(l) = X \exp[L(F - t\Delta F)]$, where $L = 4Nl$. Results in Figure 2 show that Equation 15 gives reasonably satisfactory predictions of regional IBD.

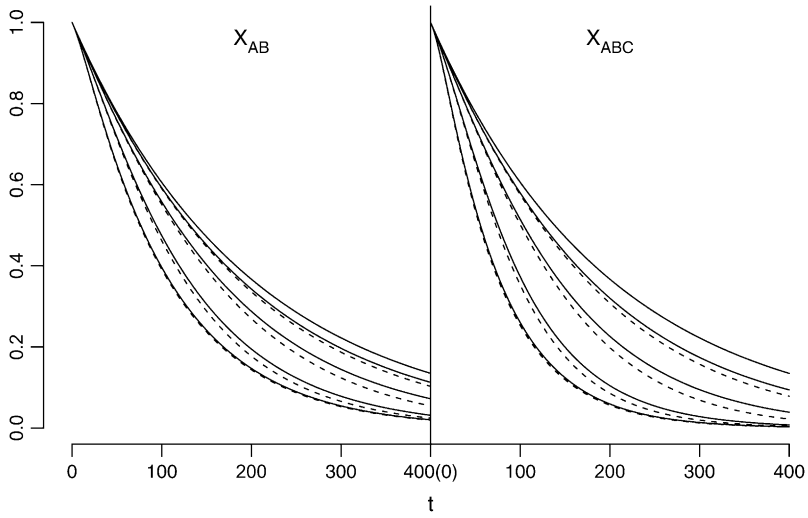

FIGURE 6.—Approximation (dashed lines) and exact (solid lines) two and three loci non-IBD (X_{AB}, X_{ABC}) over generations (t) for $R_{AB} = R_{BC} = 0$ (top lines), $\frac{1}{4}$, 1, 4, and 16 (bottom lines) with a haploid model for $N = 100$. [Note that the first $t = 400$ corresponds to the end of X_{AB} and to the beginning ($t = 0$) of X_{ABC}.]

Bottlenecks, mutation, and migration: These recurrence formulas (Equations 12–14) extend straightforwardly to include bottlenecks in population size by changing N accordingly. Assume for simplicity that mutation rates (u) are the same at each locus and migration is at rate m haplotypes from a completely unrelated and large population (*i.e.*, continent-to-island model). (A more complete migration analysis, for example using a finite-island model, is more complicated (VITALIS and COUVET 2001) and beyond the scope of this article.) From KIMURA and CROW (1964), the recurrence relation for a single locus is (ignoring higher-order terms)

$$X^*_{A,t+1} = (1 - 2u - 2m)X_{A,t} + 2u + 2m.$$

For two loci the two-locus non-IBD arises if there is no mutation on two nonrecombinant haplotypes or a mutation at locus B at haplotypes on which A is non-IBD and vice versa. In the migration model used, an immigrant haplotype is non-IBD at all loci. Hence extending Equation 13 and similarly Equation 14 leads to

$$\begin{aligned} X^*_{AB,t+1} \sim{}& [1 - 2r_{AB} - 4u - 2m - 1/(2N)]X_{AB,t} \\ &+ 2r_{AB}X_{A,t}X_{B,t} + 2u(X_A + X_B) + 2m \end{aligned} \quad (16)$$

$$\begin{aligned} X^*_{ABC,t+1} = {}& [1 - 2r_{AB} - 2r_{BC} - 6u - 2m - 1/(2N)]X_{ABC,t} \\ &+ 2r_{AB}X_{A,t}X_{BC,t} + 2r_{BC}X_{AB,t}X_{C,t} \\ &+ 2u(X_{AB,t} + X_{AC,t} + X_{BC,t}) + 2m. \end{aligned} \quad (17)$$

With mutation and migration included, asymptotic expectations as $t \to \infty$ are given in APPENDIX B. With complete linkage (so results are exact) and no migration the asymptotic value of the k-locus non-IBD probability based on iterating (17) reduces to $k!U^k$ for small values of U (from Equation A5). In contrast, it reduces to $2^{k-1}U^{2k-1}$ by using Equation 16 to obtain the two-locus non-IBD and then applying the chain rule. This illustrates the breakdown of the chain rule with mutation, whereas with migration and no mutation or recombination, the k-locus non-IBD asymptotes at $M/(M + 1)$ for any number of loci, satisfying the chain rule.

RESULTS AND DISCUSSION

The probability of IBD simultaneously at two or more neutral loci is a generalization of Wright's inbreeding coefficient, F. Such probabilities are clearly functions of the population size, time, and the breeding structure, as is F, but they also depend on the degree of linkage between loci. For example, in a closed random-mating population without mutation, the probability of double IBD is approximately equal to F^2 for unlinked loci, but increases to F for a completely linked pair. The multilocus IBD is a useful parameter in predicting the joint ancestry of multiple loci, for example, in mapping studies (MEUWISSEN *et al.* 2002), 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 (WEIR and COCKERHAM 1977; BARTON and TURELLI 2004). Whereas contributions to variance in the absence of epistasis depend only on two-locus identities or disequilibria, with epistasis, multilocus terms may be involved.

Although in principle a method exists for predicting multilocus IBD (HILL and WEIR 2007), it is unwieldy for more than four loci and applies only for a haploid model. In contrast, the chain-rule method proposed here, which utilizes the independence of crossing-over events to compute multilocus non-IBD, is computationally simple for an unlimited number of loci and applies for diploid as well as haploid models assuming random mating. It is not, however, applicable exactly in the presence of mutation. The approximate method proposed previously by HERNÁNDEZ-SÁNCHEZ *et al.* (2004) generally gives poorer predictions and becomes unwieldy to apply for more than five or so loci.

The second method proposed in this article, which is based on ignoring some of the descent measures defined by Weir and Cockerham (1974) for two loci and Hill and Weir (2007) for more, gives less precise predictions because of the simplifications made, but is straightforward to apply and leads to closed formulas at intermediate generations and for regional non-IBD. In addition, it can be applied when there is much mutation, for it generally performs better than the chain rule for any degree of recombination when mutation rates are moderate or high ($U > 0.25$) (results not shown). As the chain rule is in any case easier to apply for multiple loci, there seems little benefit in using the simple method other than to cope with mutation.

The relation between multilocus non-IBD and moments of multilocus linkage disequilibria is shown by Weir and Cockerham (1974) and Hill and Weir (2007). These require all the relevant descent measures; for two loci, for example, the expected linkage disequilibrium, $E(D^2)$ is a function of nonidentity of genes sampled from two haplotypes (*i.e.*, X_{AB}), three haplotypes, and four haplotypes. Thus neither of the linear methods developed here involving only sampling from two haplotypes can be used to predict such moments of disequilibria.

A potential application of this theory is fine mapping of QTL, where the data comprise phenotypes for the trait and genotypes at nearby marker loci, such that probabilities of IBD at the QTL can be computed for any individuals (Meuwissen and Goddard 2001). Using the equations developed here to calculate multilocus (non-)IBD, the probability of IBD at putative QTL can be computed for any pair of individuals in the population, conditional on their genotypes or IBS at marker loci. For example, for marker A and QTL B, P(IBS A, IBD B)/[P(IBS A, IBD B) + P(IBS A, non-IBD B)], in which $P(\text{IBS } A, \text{IBD } B) = F_{AB} + (F_B - F_{AB})(1 - H_A)$, where H_A is its heterozygosity in the founder population. Assuming a model of random QTL effects, the covariance due to the QTL between individuals i and j is $\text{cov}_{\text{QTL}}(i,j) = \sigma^2_{\text{QTL}} \frac{1}{2} \sum_{k,l=1}^{2} \text{IBD}(ik, jl)$, where k and l denote QTL alleles. Therefore, the variance contributed by a putative QTL (σ^2_{QTL}) at any position can be estimated using predicted IBD among all alleles in a sample. Likewise, the regression models proposed by Hernández-Sánchez *et al.* (2006) to predict IBD at the QTL given IBS at linked markers can now be more easily extended to include multiple markers together using this multilocus theory. These calculations require assumptions of population history and marker allele frequencies or heterozygosity at its foundation. In this application, at least in the livestock context, population sizes are not likely to be so large that mutation rates at marker loci, particularly SNPs, will be sufficient to lead to appreciable inaccuracies of prediction because of breakdown of the chain rule. More importantly, the robustness of the rule to migration or population introgression seems a far more important feature.

Regional IBD has also been used in gene mapping. For example, Goldgar (1990) predicted regional IBD among sibling pairs and Guo (1995) extended the method to accommodate any pair of relatives within a simple pedigree. Henceforth, gene mapping consisted of correlating phenotypic similarity with regional IBD. Regional IBD is conceptually linked to Fisher's (1953) junction theory. As junctions were defined as recombination events delimiting different IBD regions, there must be a link between the number of junctions and the regional IBD obtained in this work (*e.g.*, MacLeod *et al.* 2005).

Finally, predicting IBD from IBS requires, as do Meuwissen and Goddard (2000), information on population history, and robustness to historical assumptions is an issue needing research.

We are grateful to Mike Goddard, Bruce Weir, Xu-Sheng Zhang, and two referees for suggestions, comments, and advice. This work was supported in part by grants from the Biotechnology and Biological Sciences Research Council to W.G.H. and to Sara Knott.

LITERATURE CITED

Barton, N. H., and M. Turelli, 2004 Effects of genetic drift on variance components under a general model of epistasis. Evolution **58:** 2111–2132.

Fisher, R. A., 1953 A fuller theory of "junctions" in inbreeding. Heredity **8:** 187–197.

Goldgar, D. E., 1990 Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47:** 957–967.

Guo, S.-W., 1995 Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am. J. Hum. Genet. **56:** 1468–1476.

Hayes, B. J., P. M. Visscher, H. McPartlan and M. E. Goddard, 2003 Novel multi-locus measure of linkage disequilibrium to estimate past effective population size. Genome Res. **13:** 635–643.

Hernández-Sánchez, J., C. S. Haley and J. A. Woolliams, 2004 On the prediction of simultaneous inbreeding coefficients at multiple loci. Genet. Res. **83:** 113–120.

Hernández-Sánchez, J., C. S. Haley and J. A. Woolliams, 2006 Prediction of IBD based on population history for fine gene mapping. Genet. Sel. Evol. **38:** 231–252.

Hill, W. G., and B. S. Weir, 2007 Prediction of multi-locus inbreeding coefficients and relation to linkage disequilibrium in random mating populations. Theor. Popul. Biol. (in press).

Kimura, M., and J. F. Crow, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

Korn, G. A., and T. M. Korn, 1968 *Mathematical Handbook for Scientists and Engineers*, McGraw-Hill, New York.

MacLeod, A. K., C. S. Haley, J. A. Woolliams and P. Stam, 2005 Marker densities and the mapping of ancestral junctions. Genet. Res. **85:** 69–79.

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics **155:** 421–430.

Meuwissen, T. H. E., and M. E. Goddard, 2001 Prediction of identity by descent probabilities from marker haplotypes. Genet. Sel. Evol. **33:** 605–634.

Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics **161:** 373–379.

Vitalis, R., and D. Couvet, 2001 Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. Genet. Res. **77:** 67–81.

Weir, B. S., and C. C. Cockerham, 1969 Group inbreeding with two linked loci. Genetics **63:** 711–742.

Weir, B. S., and C. C. Cockerham, 1974 Behavior of pairs of loci in finite monoecious populations. Theor. Popul. Biol. **6:** 323–354.
Weir, B. S., and C. C. Cockerham, 1977 Two-locus theory in quantitative genetics, pp.247–269 in *Proceedings of the International Conference on Quantitative Genetics*, edited by E. Pollak, O. Kempthorne and T.B. Bailey, Jr. Iowa State University Press, Ames, IA.

Communicating editor: J. Wakeley

APPENDIX A : EXPLICIT APPROXIMATION FOR SEGMENTAL NON-IBD

From the iterative approximation (Equation 13) assuming no mutation,

$$X^*_{AB,t+1} - X^*_{AB,t} = -[2r_{AB} + 1/(2N)]X^*_{AB,t} + 2r_{AB}X_{A,t}X_{B,t}.$$

Replacing this difference equation by a differential equation, and noting that $(1 - 1/(2N)]^t \sim e^{-t/2N}$,

$$\frac{dX^*_{AB}}{dt} = -(2r_{AB} + \frac{1}{2N})X_{AB} + 2r_{AB}e^{-t/N}.$$

The equation $dy/dx + a(x)y = f(x)$ has solution $y = [\int f(x)b(x)dx]/[\int a(x)dx] + C$ (Korn and Korn 1968). Hence, after rearrangement and integration with respect to t, and noting that $X_{AB} = 1$ if $t = 0$,

$$X^*_{AB,t} = \frac{1}{4Nr_{AB} - 1}[4Nr_{AB}e^{-t/N} - e^{-(4Nr_{AB}+1)t/2N}] = \frac{1}{R_{AB} - 1}[R_{AB}e^{-t/N} - e^{-(R_{AB}+1)t/2N}] \tag{A1}$$

for $r_{AB} \neq 1/4N$, and $X^*_{AB,t} = e^{-t/N}[1 + t/2N]$ if $r_{AB} = 1/4N$. To utilize the chain rule to compute non-IBD for genomic segments, we require the derivative at generation t:

$$\frac{dX^*_{AB}}{dR_{AB}} = -\frac{1}{(R_{AB} - 1)^2}[R_{AB}e^{-t/N} - e^{-(R_{AB}+1)t/2N}] + \frac{1}{R_{AB} - 1}[e^{-t/N} + \frac{t}{2N}e^{-(R_{AB}+1)t/2N}].$$

Evaluating the derivative at $R_{AB} = 0$ (see Equation 4) and dividing by $X_A = e^{-t/2N}$,

$$\frac{d \log X^*_{AB}}{dR_{AB}}|_{R_{AB}=0} = 1 - e^{-t/2N} - \frac{t}{2N} \tag{A2}$$

and the derivative wrt r_{AB} is $4N$ times larger. Hence, an approximation, for a region of length $L = 4Nl$ is

$$\begin{aligned}\hat{X}_{A(d)Z} &= X \exp(l \frac{d \log X_{A(z)B}}{dz}) \\ &= X \exp[L(1 - e^{-t/2N} - t/2N)] = \exp[-t/2N + L(1 - e^{-t/2N} - t/2N)].\end{aligned} \tag{A3}$$

APPENDIX B: ASYMPTOTIC VALUES FROM THE SIMPLE APPROXIMATION

With mutation and migration included, we consider just asymptotic expectations, denoted $\tilde{X}_A$, $\tilde{X}^*_{AB}, \ldots$, assuming other parameters to be constant and $t \to \infty$. Equating values in successive generations following Kimura and Crow (1964), for a single locus $\tilde{X}_A = (U + M)/(U + M + 1)$, and for two loci from Equation 16,

$$\tilde{X}^*_{AB} \sim \left[\frac{1}{2U + M + R_{AB} + 1}\right]\left[\frac{(U + M)^2 R_{AB}}{(U + M + 1)^2} + \frac{2U(U + M)}{U + M + 1} + M\right]. \tag{B1}$$

With complete linkage and no migration ($M = R_{AB} = 0$), (B1) is exact and reduces to $\tilde{X}^*_{AB} = [U/(U + 1)][2U/(2U + 1)]$. For k loci with the same assumptions, by using Equation 17 it can be shown that

$$\tilde{X}^*_{1\ldots k} = \prod_{i=1}^{k}\left(\frac{iU}{iU + 1}\right), \tag{B2}$$

which reduces to $k!\,U^k$ for small values of U.