

PREDICTION AND APPLICATIONS OF MULTI-LOCUS INBREEDING

J. Hernández-Sánchez¹ and W.G. Hill¹

¹ Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK

INTRODUCTION

Inbreeding (F) is the probability that alleles at a locus are identical by descent (IBD) within individuals. F is part of a set of IBD coefficients that play a key role in computing (co)variances of quantitative traits (Lynch and Walsh 1998), inferring historic population structure (Hayes *et al.* 2003), and gene mapping (Hernández-Sánchez *et al.* 2006). Historical IBD probabilities enable fine gene mapping because they contain short-range linkage disequilibrium (LD) information. For example, Meuwissen and Goddard (2004) mapped the DGAT1 gene in dairy cattle to a 0.04 cM region after combining linkage with LD and multi-trait data. F depends on population size, generation and breeding structure. Weir and Cockerham (1974) extended F to any two loci A and B, F_{AB} , for haploid and diploid populations with various mating structures. F_{AB} depends on the same parameters as F and also on recombination rates between loci (c_{AB}). Hernández-Sánchez *et al.* (2004) developed accurate approximations for three and four loci (F_{ABC} and F_{ABCD}) using regression methods to predict conditional inbreeding at one locus given inbreeding at two other loci, e.g. $F_{ABC} = F_{B|AC}F_{AC}$ and from that $F_{ABCD} = F_{C|ABD}F_{ABD}$. Predictions were generally better when IBD at the middle locus, e.g. B, was predicted conditional on IBD at the outer pair, e.g. A and C. Hill and Weir (2006) developed transition matrices to predict exact multi-locus inbreeding in haploid populations, and showed how IBD coefficients depend on non-IBD coefficients, e.g. $X_{AB} = 1 - F_A - F_B + F_{AB}$, where X_{AB} is the probability of non-IBD at both A and B, and how non-IBD coefficients relate to moments of multi-locus LD. Here we develop a general approximation for non-IBD at each of several ordered loci from non-IBD at two loci, e.g. $X_{ABC} = X_{AB}X_{BC}/X_B$, that is slightly more accurate than that proposed by Hernández-Sánchez *et al.* (2004). Moreover, this approximation allows us to predict non-IBD and then F at any number of loci sequentially. It follows that, as the number of loci increases within a DNA segment of fixed length, i.e. as $c \rightarrow 0$ between consecutive loci, the multi-locus F asymptotes to the probability of IBD for that DNA segment.

MATERIAL AND METHODS

Computation of non-IBD coefficients. Following Weir and Cockerham (1974) it is necessary to define recurrence relations between generations for probabilities of non-IBD at each of the k loci sampled from 2, 3 ... $2k$ haplotypes. For example, consider the case of two loci in a diploid population with random mating and selfing, no selection or mutation and constant population size. The non-IBD coefficients at loci A and B are $\mathbf{x}' = [X_{AB,AB} \ X_{A,B,AB} \ X_{A,B,A,B}]$, where $X_{AB,AB}$ (also simply denoted X_{AB}) denotes alleles sampled in two haplotypes, $X_{A,B,AB}$ in three haplotypes, and $X_{A,B,A,B}$ in four haplotypes. Let $c = c_{AB}$, N the number of diploid individuals, and \mathbf{x}_t the vector of non-IBD probabilities at generation t . The recurrence is $\mathbf{x}_t = \mathbf{T}'\mathbf{x}_0$, where

$$\mathbf{T} = \begin{bmatrix} (1-c)^2 - \frac{(1-2c)}{2N} & \frac{2(N-1)c(1-c)}{N} & \frac{(N-1)c^2}{N} \\ \frac{(1-c)}{2N} - \frac{(1-2c)}{4N^2} & \frac{(N-1)[(N+1) + (1-2c)(N-2)]}{2N^2} & \frac{(N-1)(2N-3)c}{2N^2} \\ \frac{2N-1}{4N^3} & \frac{(N-1)(2N-1)}{N^3} & \frac{(N-1)(2N-1)(2N-3)}{4N^3} \end{bmatrix}$$

If all individuals are non-inbred and unrelated initially, then \mathbf{x}_0 is a unit vector. Weir and Cockerham (1974) showed how to compute \mathbf{T} for other mating systems with two loci. Hill and Weir (2006) derived a method to compute \mathbf{T} for three or more loci, using a haploid model with random mating and selfing, although its dimensions rise rapidly: for example there are 16 terms with three loci and 139 with four. Thus, this method quickly becomes unwieldy.

As $X_{ABC} = X_{AB}X_{C|AB}$, then utilizing the order of loci on the chromosome, we can predict $X_{ABC} = X_{AB}X_{C|B} = X_{AB}X_{BC}/X_B$. Similarly, non-IBD at A, B, C and D can be calculated sequentially as $X_{ABCD} = X_{ABC}X_{D|C} = X_{AB}X_{BC}X_{CD}/X_BX_C$. Hence,

$$X_{1\dots k} = \prod_{i=1}^{k-1} X_{i,i+1} / X_1^{k-2}, \quad [1] \quad \text{where}$$

$X_{i,i+1}$ are pair-wise non-IBD probabilities at consecutive loci, and $X_1 = X_A$ is the single locus non-IBD assumed to be the same for all loci. Weir and Cockerham (1974) developed an exact theory for obtaining two-locus (non-)IBD coefficients in a wide range of breeding structures that makes this method more generally applicable.

Relationship between non-IBD and IBD. Notwithstanding the relationships $X_{ABC} = X_{AB}X_{C|AB}$ and [1], $F_{ABC} \neq F_{AB}F_{BC}/F_B$, basically because A and B can be IBD on one haplotype and B and C on another, but these may not be on the same haplotype. Hence we adopt the algorithm for multi-locus non-IBD to obtain multi-locus IBD coefficients under the same assumptions of population structure using $F_{AB} = 1 - X_A - X_B + X_{AB}$ and in general at k loci simultaneously

$$F_{1\dots k} = 1 - \sum_{i=1}^k X_i + \sum_{i<j}^k X_{ij} - \dots + (-1)^k X_{1\dots k}, \quad [2]$$

where numbers substitute letters to denote loci.

Segment IBD. The probability of inbreeding at a given DNA segment of fixed length is $\lim_{k \rightarrow \infty} F_{1\dots k}$. However, it is not computationally efficient to calculate equation [2] directly, as it involves $2^k - 1$ terms. More efficient methods may involve assuming equal distances between consecutive loci, and grouping equivalent non-IBD terms together so that they are computed just once. This approach involves calculating k terms for $F_{1\dots k}$.

RESULTS AND DISCUSSION

Table 1 shows F_{ABC} for a diploid population obtained via simulations (x1000 reps), with the sequential method of equations [1] and [2], and directly with a regression method proposed by Hernández-Sánchez *et al.* (2004). We assume random mating and selfing, and no migration, mutation or selection. In these examples, both methods perform well.

Table 1. Simulations and two predictions of inbreeding at three loci

c_{AB}, c_{BC}	0.0025, 0.01			0.01, 0.01		
	Sim	Seq	Reg	Sim	Seq	Reg
25	0.0895 (.00090)	0.0903	0.0896	0.0770 (.00084)	0.0780	0.0760
50	0.1444 (.00111)	0.1461	0.1438	0.1171 (.00102)	0.1183	0.1123
100	0.2369 (.00135)	0.2341	0.2300	0.1846 (.00123)	0.1849	0.1764
150	0.3143 (.00147)	0.3212	0.3177	0.2548 (.00138)	0.2606	0.2542
200	0.4216 (.00156)	0.4098	0.4076	0.3495 (.00151)	0.3463	0.3424
300	0.5791 (.00156)	0.5758	0.5754	0.5188 (.00158)	0.5230	0.5212
400	0.7077 (.00144)	0.7102	0.7103	0.6603 (.00150)	0.6744	0.6742

c_{AB}, c_{BC} = recombination rate; t = generation; $N = 100$; Sim = Simulation (s.e. in brackets); Seq = sequential method; Reg = regression method.

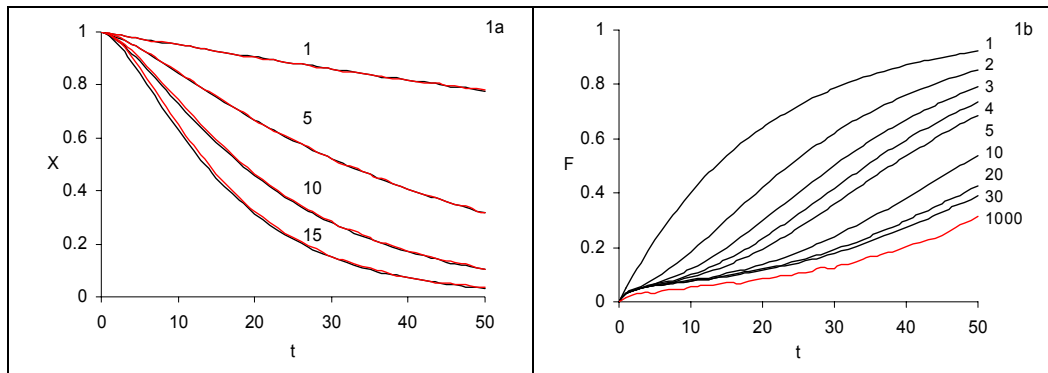


Figure 1. a) Non-IBD ($X_{1...k}$) at 1, 5, 10 and 15 loci, with 10cM between adjacent loci, and $N=100$. b) Predictions of IBD, $F_{1...k}$ for a 50cM DNA segment with $N=10$ for $k=1, 2, 3, 4, 5, 10, 20, 30$ and 1000 loci (the latter obtained with simulations)

Figure 1a shows predictions (black) and simulations (red) for $X_1, X_{1...5}, X_{1...10}$ and $X_{1...15}$ for a diploid population with random mating and selfing with $N = 100$, and consecutive markers were spaced 10 cM apart, e.g. $X_{1...10}$ was predicted for a region spanning 90cM. X_1 and X_{12} are exact predictions and the rest are from [1] using X_1 and X_{12} . We can see that the accuracy of approximations is good, although it decreases for large k and low t . This accuracy increases with shorter distances or larger populations sizes. Figure 1b shows the asymptotic behavior of $F_{1...k}$ as k increases for $N=10$, total map length of 50cM, and breeding structure as in Fig. 1a. This probability was compared against a segment IBD probability obtained via simulations ($\times 1000$ reps) for an equivalent distance using 1000 evenly spread markers. The approach to the asymptote is seen to be slow. Repeated use of Aitken's approximation enabled quite good predictions of the asymptote to be obtained for shorter distances, e.g. 25cM (not shown).

CONCLUSION

We have shown the relationship between non-IBD and IBD coefficients. The most useful parameters are the latter, and an easy way to compute them is via the intermediary step of computing non-IBD. In general, both the present sequential and regression methods (Hernández-Sánchez *et al.* 2004) rendered very similar results. However, we have observed that for low F and low ΔF (rate of inbreeding) the sequential method was closer to simulation results than the regression method (unpublished results). An obvious advantage of the sequential method over the regression method is the ease in which it is extended to predict non-IBD at any number of loci, from the key relationship $X_{ABC} = X_{AB}X_{BC}/X_B$, and hence obtain predictions of $F_{1...k}$. Thus, the probability of IBD of a DNA segment is $\lim_{k \rightarrow \infty} F_k$, which can be

approximated using a sufficiently large finite k . This approximation requires efficient methods to compute $F_{1...k}$. The theory developed here has been incorporated into GridQTL, a grid-based software package for QTL mapping that develops and extends the QTL Express software (Seaton *et al.* 2002). In particular, this theory has helped to improve IBD predictions given marker data and history information. These IBD predictions can then be utilised in variance component analyses for fine gene mapping.

ACKNOWLEDGEMENTS

We acknowledge BBSRC for financial support and B.S. Weir for inspiration.

REFERENCES

- Hayes, B.J., Visscher, P.M., McPartlan, H.C. and Goddard, M.E. (2003) *Genome Res.* **13** : 635-643
- Hernández-Sánchez, J., Haley, C.S. and Woolliams, J.A. (2004) *Genet. Res.* **83** : 113-120
- Hernández-Sánchez, J., Haley, C.S. and Woolliams, J.A. (2006) *Genet. Sel. Evol.* (in press)
- Hill, W.G. and Weir, B.S. (2006) *Theor. Pop. Biol.* (submitted)
- Lynch, M. and Walsh, B. (1998) «Genetics and analysis of quantitative traits». Sinauer, Sunderland, US.
- Meuwissen, T.H.E. and Goddard, M.E. (2004) *Genet. Sel. Evol.* **36** : 261-279
- Seaton, G., Haley, C.S., Knott, S.A., Kearsley, M. and Visscher, P.M. (2002) *Bioinformatics* **18** : 339-340
- Weir, B.S. and Cockerham, C.C. (1974) *Theor. Pop. Biol.* **6** : 323-354