July 2001

Predictors of marker-informativeness for an outbred F_2 design

J. L. Rocha
University of Nebraska - Lincoln

Daniel Pomp
University of Nebraska - Lincoln, dpomp1@unl.edu

L. Dale Van Vleck
University of Nebraska - Lincoln, dvan-vleck1@unl.edu

Merlyn K. Nielsen
University of Nebraska - Lincoln, mnielsen1@unl.edu

Predictors of marker-informativeness for an outbred F_2 design

J. L. Rocha*, D. Pomp*, L. D. Van Vleck† and M. K. Nielsen*

*Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE, USA. †USDA, ARS, USMARC, Lincoln, NE, USA

Summary

Generalization of the polymorphism information content (PIC) index to represent marker informativeness (MI) for a three-generation F_2 design requires that two additional sources of non-informativeness be added to the PIC formula: the probability of matings between like-heterozygous F_1 individuals, of which one is non-informative; and that of matings between like-heterozygous F_1 individuals, which are both fully informative but where line of origin of the same alleles is reciprocal. Given the dense marker-maps currently available for some species, this F_2 informativeness parameter constitutes the natural criterion for marker selection in F_2 designs, and two computer programs to predict MI from grandparental marker-genotypes were developed for an F_2 population originating from two divergent selection lines of outbred mice ($F \sim 0.2$). A total of 403 markers had been genotyped for the F_0 grandparents ($n = 31$), and 14 markers had also been genotyped in the complete pedigree including 559 F_2 individuals. One program was based on assumptions of random-mating (RM), while the other (PED) accounted for the pedigreed mating structure. For the 403 markers, the correlation between MI from RM and from PED was 0.95, and the average deviation between the two predictions was 0.005 MI units (MI ranged from 0 to 1). Correlations between predicted and realized MI for the 14 fully genotyped markers were 0.97 for PED and 0.94 for RM, while the corresponding average of deviations between predicted and actual values were 0.01 and 0.04, respectively. Absolute deviations from realized MI never exceeded 0.09 and 0.16 for PED and RM, respectively. Simulated optimization of the mating system to maximize average MI of 28 markers on one chromosome led to improvements in the range of 15–20% average MI (0.07–0.09 MI units). The degree of relative advantage conferred by the F_2 generalization of the PIC index over the traditional index was found to be of minor significance.

Keywords genetic marker, linkage disequilibrium, outbred cross, polymorphism information content.

Introduction

Availability of dense marker maps affords the possibility of selecting markers to maximize marker informativeness (MI). Linkage analysis requires at least one parent to be heterozygous for the loci under study (Guo & Elston 1999), so the best (most informative) markers are those with the highest frequencies of heterozygous parents. Thus, MI and marker heterozygosity in the parental generation are positively correlated for the purposes of linkage analyses (Botstein *et al.* 1980; Da *et al.* 1999; Guo & Elston 1999).

Haley *et al.* (1994) have clearly demonstrated how marker information content (or polymorphism) is also directly and positively related to the mean maximum test statistic in a quantitative trait loci (QTL) analysis, which, in turn, affects sample sizes required for detection of QTL at a given level of statistical power (Da *et al.* 1999). Hence, within the framework of achieving relatively equal marker spacing for comprehensive genomic coverage, MI should be optimized

Address for correspondence

D. Pomp, Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE 68583-0908, USA.
E-mail: dpomp@unl.edu

Accepted for publication 30 July 2001

in a mapping project and should be the foundation for selection of markers.

The polymorphism information content (PIC) index (Botstein *et al.* 1980) estimates MI for a specific two-generation model. Other studies (e.g. Da *et al.* 1999) have also dealt with polymorphism measures in the context of two-generation models. However, in the context of a three-generation outbred F_2 design, where MI is the fraction of F_2 alleles for which grandparental line (or breed) of origin can be unambiguously ascertained, additional sources of non-informativeness need to be considered because of marker-allele sharing between lines (or breeds).

This study reports a new formula that generalizes the PIC index for an F_2 three-generation model, and the development and comparison of MI predictors for marker selection in an outbred F_2 design. We applied these predictors to estimate MI in a project identifying QTL for energy balance using an F_2 cross between two divergent selection lines of outbred mice ($F \sim 0.2$; Nielsen *et al.* 1997), where genotypes for all F_0 grandparents ($n = 31$) had been collected for a large number of markers (403). We also consider the relationships of MI predictors with accrual of inbreeding and with measurements of linkage disequilibrium. Finally, optimization of mating systems to maximize average MI for a given set of markers was evaluated and a computer program developed for that purpose.

Materials and methods

A new formula that generalizes the PIC index for an F_2 three-generation model was calculated by incorporating sources of non-informativeness specific to the F_2 context (Figs 1 & 2). Two computer programs to predict MI from grandparental marker-data were developed using SAS (SAS Institute Inc. 1985) and are available upon request: one implements the general formula under assumptions of random-mating (RM), and the other tracks the actual pedigreed mating structure used to generate the F_2 progeny (PED). These programs were applied to grandparental genotypic data for 403 markers [using F_0 from the low-high (LH) intercross population described by Moody *et al.* (1999)], and the correlation between the two predictions of MI was computed. The average deviation between the two predictions was computed for each marker, and the UNIVARIATE procedure of SAS (SAS Institute Inc. 1985) was used to test whether these deviations followed a normal distribution with mean 0. Variation observed in MI accounted for by the number of marker alleles was studied with correlations and linear regression models fitted under the REG procedure of SAS (SAS Institute Inc. 1985).

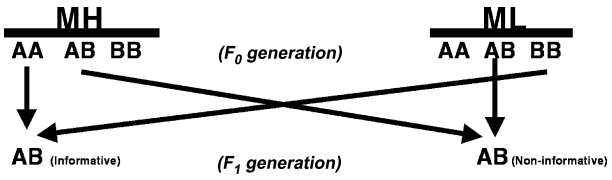

Figure 1 The F_2 generalization of the PIC index – uninformative situation I. MH and ML represent two grandparental non-inbred lines. One of the F_1 individuals is informative (left) while the other is uninformative (right). If these two individuals are mated to produce F_2 offspring, then half of the informativeness of the individual on the left is lost (because its meiosis will only be informative when producing F_2 homozygotes). This situation leads to the third subtractive term in Equation (1).

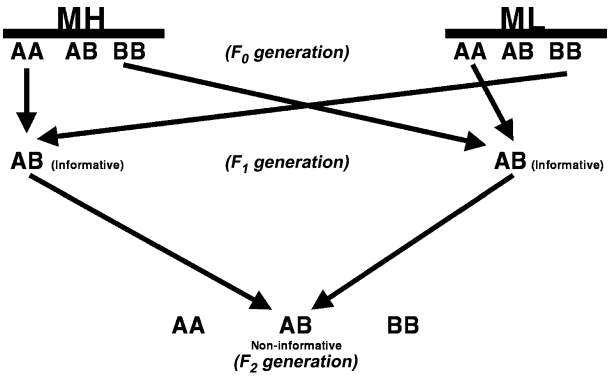

Figure 2 The F_2 generalization of the PIC index – uninformative situation II. Both these heterozygous F_1 individuals are informative. However, the same allele indicates reciprocal lines of origin in the two different F_1s. A mating between these two individuals will produce heterozygous F_2 individuals that are non-informative (fourth subtractive term in Equation 1, see text). Note: matings between homozygous grandparents are represented, but exactly the same situation is derived from equivalent matings involving homozygous and heterozygous grandparents – the last three terms in Equation (1).

A sample of 14 marker-loci was available for which all F_2 progeny ($n = 559$) had been genotyped (Moody *et al.* 1999). This sample allowed evaluation of the accuracy of the MI predictions. Correlations between predicted and actual MI values for these 14 markers, as well as corresponding average and extreme value deviations, were computed. This sample of 14 markers was also used to assess the degree of relative advantage conferred by the F_2 generalization of the PIC index (formula 1) over more simplistic approaches to approximate MI, and correlations were also calculated between actual MI values and those estimated by pedigree-based predictions of F_1 heterozygosity and of conventional PIC index (Botstein *et al.* 1980) values.

Inbreeding and MI

The same outbred selection lines used to generate the F_2 progeny had subsequently been subject to full-sib matings to

develop partially inbred lines (F ~ 0.8). Because predictions of MI described above quantify degree of allele sharing between the two selection lines, the hypothesis was that the MI predictions would also be good predictors of fixation of the same or different marker-alleles in the two resulting inbred lines (i.e. the higher the prediction of MI for a given marker, the greater the likelihood that the two lines would be fixed for different alleles at that marker). This hypothesis was tested by assigning marker-loci to three categories in the inbred lines: those which had reached fixation for different alleles (DIF); those which were still segregating between the lines (SEG), and those which had reached fixation for the same allele (ID). The average predicted MIs (PED) in the outbred lines for these three marker-categories were then computed and contrasted with an analysis of variance (ANOVA procedure, SAS Institute Inc. 1985).

Linkage disequilibrium and MI

The mouse selection lines (Nielsen *et al.* 1997) originated from a four-way composite of outbred lines. The F_2 progeny evaluated here were produced after 16 generations of divergent selection and three generations of relaxed selection. The formation of the initial four-way composite should have resulted in considerable linkage disequilibrium (LD), the erosion of which may have been slowed by the 16 generations of divergent selection implemented (Nielsen *et al.* 1997). Correlations between the MI indices of pairs of markers separated by varying genetic distances were computed, under the assumption that they should reflect existing LD. In an attempt to avoid artificially inflating or deflating these correlations, monomorphic markers (MI = 0) and pairs of markers with very different numbers of alleles were excluded.

Optimization of the mating system to maximize MI

To investigate the extent to which optimization of the mating system could improve average MI for a set of markers, a computer program was developed (available upon request) to optimize the mating system with respect to only the two first subtractive terms of Equation (1) below, relying only on grandparental data. For each marker the program computes an $m \times n$ matrix containing probabilities of F_1 heterozygosity (subject to the conditions in formula 1) for each of the possible matings among grandparents (m and n being the number of grandparents in lines MH and ML, respectively; see Figs 1 & 2). For each successive marker that is processed, the newly created matrix is added to the previous one. The final output is the sum of matrices for all markers. The cells with the highest numeric value identify those grandparental matings that would result in the maximum (expected) average MI for the set of markers processed. Two optimization schemes were tested with the full set of 28 markers on chromosome 1, encompassing the (1) 10 best and (2) six best grandparental matings. Both optimization schemes were subject to the conditions that one female could not be mated to two males, and that one male could only be mated to a maximum of three females. The F_1 matings were at random, subject only to the condition of no full-sib matings.

Results and discussion

Generalization of the PIC index for an F_2 design

Informativeness in F_2 progeny ranges from 0 to 1 and is the fraction of F_2 alleles for which grandparental line of origin can be unambiguously ascertained. Knowledge of line of origin is key for QTL analyses and requires F_1 heterozygosity. However, not every heterozygous F_1 will be informative or will lead to full informativeness in the F_2 (Figs 1 & 2). Simultaneous consideration of all uninformative possibilities leads to the following expression for MI for a three-generation F_2 model (assuming a segregation ratio of 1:2:1 in the F_2):

$$\begin{aligned}\mathrm{MI} = 1 &- \left(\sum_{i=1}^{n} f_{i\mathrm{H}} f_{i\mathrm{L}}\right) - \left(\sum_{i=1,j=2,i\neq j}^{n} 2 f_{i\mathrm{H}} f_{j\mathrm{H}} f_{i\mathrm{L}} f_{j\mathrm{L}}\right) \\ &- 1/2 \sum_{i=1,j=2,i\neq j}^{n} \left[\left(f_{i\mathrm{H}} f_{j\mathrm{L}} + f_{i\mathrm{L}} f_{j\mathrm{H}} - 2 f_{i\mathrm{H}} f_{j\mathrm{H}} f_{i\mathrm{L}} f_{j\mathrm{L}}\right)\left(2 f_{i\mathrm{H}} f_{j\mathrm{H}} f_{i\mathrm{L}} f_{j\mathrm{L}}\right)\right] \\ &- \left(\sum_{i=1,j=2,i\neq j}^{n} f_{i\mathrm{H}}^2 f_{j\mathrm{L}}^2 f_{j\mathrm{H}}^2 f_{i\mathrm{L}}^2\right) - \left[\sum_{i=1,j=2,i\neq j}^{n} \left(f_{i\mathrm{H}}^2 f_{i\mathrm{L}} f_{j\mathrm{L}} + f_{i\mathrm{H}} f_{j\mathrm{H}} f_{j\mathrm{L}}^2\right) \right. \\ &\qquad \left. \times \left(f_{i\mathrm{H}} f_{j\mathrm{H}} f_{i\mathrm{L}}^2 + f_{j\mathrm{H}}^2 f_{i\mathrm{L}} f_{j\mathrm{L}}\right)\right] \\ &- \sum_{i=1,j=2,i\neq j}^{n} \left[\left(f_{i\mathrm{H}}^2 f_{j\mathrm{L}}^2\right)\left(f_{i\mathrm{H}} f_{j\mathrm{H}} f_{i\mathrm{L}}^2 + f_{j\mathrm{H}}^2 f_{i\mathrm{L}} f_{j\mathrm{L}}\right)\right] \\ &- \sum_{i=1,j=2,i\neq j}^{n} \left[\left(f_{j\mathrm{H}}^2 f_{i\mathrm{L}}^2\right)\left(f_{i\mathrm{H}}^2 f_{i\mathrm{L}} f_{j\mathrm{L}} + f_{i\mathrm{H}} f_{j\mathrm{H}} f_{j\mathrm{L}}^2\right)\right] \qquad (1)\end{aligned}$$

where $f_{i\mathrm{H}}$ and $f_{i\mathrm{L}}$ represent the frequencies of allele i in the grandparental lines MH and ML (Figs 1 & 2), respectively, and n is the total number of alleles for a given marker-locus. It should be noticed that the first three terms of this equation amount to the PIC index developed by Botstein *et al.* (1980).

For X-linked markers (with no Y chromosome homologue), MI was computed as:

$$[(\text{Probability of } F_1 \text{ heterozygosity}) \times 2/3] + 1/3 \qquad (2)$$

because paternal meioses are always informative with a maximum of three F_2 alleles available for QTL analyses, for any given mating of F_1 parents.

Although a specific F_2 context was considered in this study, it should be noted that generalization of the PIC index to represent MI for a three-generation outbred backcross design would be a very similar exercise. The outcome for the outbred backcross design would encompass only the first four terms of formula (1).

Computer programs and accuracy of MI predictions

The computer program RM estimates marker-allele frequencies in the grandparental samples ($n = 31$) and implements formula (1). However, a few qualifications are necessary. First, actual grandparental genotypic frequencies were used in RM, rather than the corresponding products of allele frequencies as depicted in (1). Secondly, because of their cumbersome nature and small magnitude, the third and the last two subtractive terms in (1) were not included in RM; for the sample of 14 markers previously mentioned, the sum of these three terms averaged only 0.012 and ranged from 0 to 0.031 MI units.

The computer program PED implements the same underlying principles while tracking the actual pedigree of the F_2 progeny. For this experiment there were 12 grandparental matings, each involving four (not necessarily different) alleles. For each of these matings PED makes four comparisons of alleles corresponding to the first two subtractive terms in formula (1) (the probability of F_1 heterozygosity, adjusted for uninformative heterozygotes resulting from the mating of like-heterozygote grandparents).

Subsequently PED processes F_1 matings to account for the uninformative situations illustrated in Figs 1 and 2 [corresponding to the last five subtractive terms in (1)]. For this experiment there were 12 basic F_1 mating-types, each involving eight grandparental alleles (F_1 genotypes not available). For each of these F_1 matings PED does the array of allele comparisons required to compute probabilities of uninformative situations. The final prediction of MI is then computed reflecting these probabilities.

Both programs were applied to the grandparental genotypic data from 403 markers. Comparison of their predictions yielded a correlation of 0.95, while the deviations between predictions ranged from –0.30 to +0.30, and averaged 0.005 MI units (MI ranges from 0 to 1). These deviations followed a normal distribution with mean not significantly different from 0 ($P = 0.21$).

Number of alleles for the 403 microsatellite markers averaged 2.9 and ranged from one to eight. Correlations of MI with number of marker-alleles were 0.65 for PED, 0.70 for RM (samples of 403 markers), and 0.66 for actual MI (sample of 14 markers). Regression of MI on number of alleles yielded intercepts not significantly different from 0 ($P = 0.82$ and 0.90, for PED and RM, respectively), and regression coefficients of +0.15 MI units per additional marker-allele ($P = 0.0001$ for both models). Similar results were obtained with the regression model involving actual MI. There were 54 markers with a PED-predicted MI of 0. Of these 54, seven markers had more than one allele. When these 54 markers were excluded from computations, correlations between MI and allele number became 0.48 for PED, and 0.54 for RM, while the correlation between PED and RM remained very high (0.92).

For the 14 markers for which all F_2 progeny had been genotyped, the comparison between predictions and actual MI is summarized in Table 1. As expected, the pedigree-based program (PED) was the best predictor for the marker selection process. The average deviation from realized MI was not significantly different from 0 ($P = 0.42$). However, the relative advantage of PED over pedigree-based simple-PIC predictions [the first three terms of formula (1)] was found to be of minor significance (Table 1). The rank-correlations among the pedigree-based predictions and actual MI values were 0.96 for PED, 0.94 for PIC and 0.88 for F_1 heterozygosity. That same rank-correlation was 0.91 for RM. Very high correlations between PED and pedigree-based simple-PIC predictions were also observed for other samples of markers (a correlation of 0.995 and a rank-correlation of 0.99 for 26 markers on chromosome 1).

Although PED, based on exclusively grandparental data, was found to be a good predictor of MI, whenever possible simultaneous genotyping of F_1 parents would be recommended. Then, the only source of error would be the possibility of segregation distortion, which would have only a minimal impact as the last five cross-product terms in (1) are in general very small.

Table 1 Comparison of actual and predicted marker informativeness (MI) for a sample of 14 markers.

Method[1]	Correlation[2]	Average deviation[3]	Range of deviations
PED	0.97	–0.01	–0.091 + 0.078
RM	0.94	–0.04	–0.156 + 0.085
F_1HET	0.86	–0.07	–0.322 + 0.078
PIC	0.96	–0.03	–0.126 + 0.078

[1]MI predicted by alternative approaches: PED is a computer program that implements the principles in formula (1) (see text) while tracking the pedigreed mating structure of the F_2 population; RM is a computer program that implements (1) under assumptions of random-mating; F_1HET is the pedigree-based prediction of F_1 heterozygosity from grandparental marker-data [the first two terms in (1)]; PIC is the PIC-predicted index from pedigreed grandparental marker-data [the first three terms in (1)].
[2]Correlation between actual and predicted MI.
[3]Deviations computed as 'actual MI – predicted MI'.

Inbreeding and MI

The hypothesis that MI of outbred populations has some predictive power concerning the outcome of subsequent inbreeding programs for specific marker-loci was verified. The average outbred MIs for the different inbred marker-categories were 0.608, 0.470 and 0.420, for DIF, SEG and ID, respectively. Differences among these means were highly significant ($P = 0.0001$), but the specific pair-wise difference between SEG and ID was not significant ($P = 0.10$).

Linkage disequilibrium and MI

If genotypes of pairs of marker-loci are independently distributed, then their MI indices should not be correlated. However, if there is LD leading to genotypic associations and dependencies between pairs of markers, then the degree and extent of these associations should be reflected in correlations between the MI indices of these markers. Figure 3 displays the pattern of correlations observed between the MI indices of pairs of markers separated by varying genetic distances, and suggests retention of LD through generation 19 of these outbred lines of mice for genetic distances up to 13 cM.

Figure 3 also displays the theoretical expectation for the decay in linkage disequilibrium after 19 generations (Falconer & Mackay 1996). This expectation is relative to a degree of initial linkage disequilibrium (D_0), which is unknown. For the purpose of the graphical representation in Fig. 3, D_0 was assumed to have corresponded to an MI correlation of 0.5 in the first generation of the four-way composite. After 19 generations, the theoretical expectation is that 83% of the initial D_0 will still be retained within 1-cM regions, while only 7% of D_0 will be retained between loci that are separated by 13-cM intervals.

The small sample sizes involved (average $n = 57$ pairs of markers) do not lead to a smooth, linearly declining pattern, but for distances of up to 13 cM all correlations were positive, some of them being fairly high and statistically significant. Beyond this distance the correlations declined sharply, were always near 0, and were sometimes negative. The overall pattern suggests the existence of linkage disequilibrium within 13-cM regions in these populations. These observations are in reasonable agreement with results of simulations conducted by Stephens *et al.* (1994).

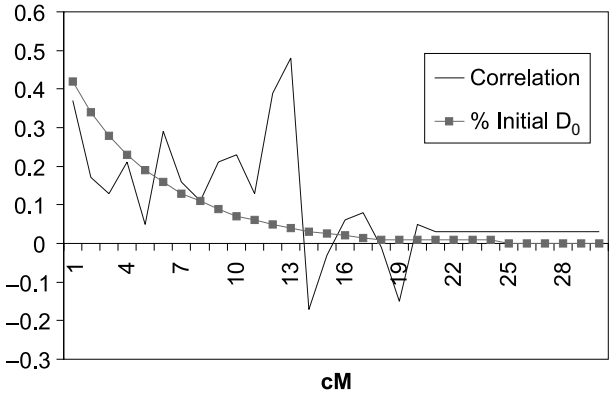

Figure 3 Correlations between MI indices of linked loci in generation 19 of a four-way composite. Percentage of initial D_0 is a theoretical expectation of the degree of linkage disequilibrium retained, based on Falconer & Mackay (1996) and assuming D_0 amounted to an MI correlation of 0.5 in the first generation of the four-way composite (see text).

Table 2 Marker informativeness (MI) in optimized and non-optimized mating systems.

Marker	MI-non[1]	MI-opt10[2]	MI-opt6[2]
1	0.40	0.63	0.56
2	0.44	0.28	0.17
3	0.71	0.70	0.75
4	0.42	0.52	0.49
5	0.66	0.61	0.70
6	0.29	0.40	0.48
7	0.12	0.50	0.67
8	0.00	0.10	0.00
9	0.56	0.46	0.29
10	0.21	0.23	0.33
11	0.46	0.78	0.83
12	0.35	0.66	0.73
13	0.63	0.45	0.50
14	0.00	0.25	0.17
15	0.57	0.65	0.63
16	0.76	0.98	0.96
17	0.85	1.0	1.0
18	0.80	0.45	0.42
19	0.35	0.50	0.50
20	0.58	0.55	0.65
21	0.77	0.80	0.88
22	0.79	0.90	0.83
23	0.39	0.63	0.71
24	0.57	0.41	0.67
25	0.65	0.65	0.58
26	0.72	0.87	0.85
27	0.32	0.45	0.58
28	0.31	0.29	0.31
Average	0.489	0.561	0.580
No. MI > 0.48	14	17	21

[1]MI-non: PED-predicted MI from non-optimized mating system.
[2]MI-opt10 and MI-opt6: PED-predicted MI from optimized mating systems including the 10 or the six best grandparental matings, respectively.

Optimization of the mating system to maximize MI

The full results for the set of 28 markers are presented in Table 2. Over this set of 28 markers, the optimization procedure resulted in an improvement in average (predicted) MI of 0.07 MI units (14.7%) if a mating system with the 10 best grandparental matings would be adopted (close to the original mating system involving 12 grandparental matings). If the mating system would include only the six best grandparental matings, then the improvement in average (predicted) MI was 0.09 MI units (18.6%). Assuming an arbitrary MI threshold of ~0.50 for marker-selection, the optimization procedures would increase the number of markers selected from 14 (50% – random-mating) to 17 (61% – 10 best matings) or to 21 (75% – six best matings; Table 2).

Considering the large number of markers involved, the improvement obtained in average MI (Table 2) can be considered reasonable. The larger the number of markers considered, the more this improvement would be diluted, and obviously, positional considerations would also need to be brought into this process. However, the results in Table 2 indicate that for projects involving a small to moderate number of markers, such as confirmation and fine-mapping studies, genotyping of grandparents and consideration of procedures for mating system optimization can lead to significant improvement in the levels of average MI.

Acknowledgements

The authors are grateful to Katherine Gilson for grandparental marker-genotyping, and to Dan Yardley for assistance with the computer programming of the optimization algorithm. We also utilized data previously collected by Diane Moody and acknowledge her efforts. JLR acknowledges the support of the Portuguese Foundation for Science and Technology. Partial support for this work was through a grant to DP (GM60029) from the National Institutes of Health, National Institute of General Medical Sciences. This work is published as paper 13268 of the Journal Series, Agricultural Research Division, University of Nebraska.

References

Botstein D., White R.L., Skolnick M. & Davis R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32,** 314–31.

Da Y., Van Raden P.M., Ron M., Beever J.E., Paszek A.A., Song J., Wiggans G.R., Ma R., Weller J.I. & Lewin H.A. (1999) Standardization and conversion of marker polymorphism measures. *Animal Biotechnology* **10,** 25–35.

Falconer D.S. & Mackay T.F. (1996) *Introduction to Quantitative Genetics*. Longman Group Ltd., Harlow, UK.

Guo X. & Elston R.C. (1999) Linkage Information Content of polymorphic genetic markers. *Human Heredity* **49,** 112–8.

Haley C.S., Knott S.A. & Elsen J.-M. (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136,** 1195–207.

Moody D.E., Pomp D., Nielsen M.K. & Van Vleck L.D. (1999) Identification of QTL influencing traits related to energy balance in mice. *Genetics* **152,** 699–711.

Nielsen M.K., Jones L.D., Freking B.A. & DeShazer J.A. (1997) Divergent selection for heat loss in mice. I. Selection applied and direct responses through fifteen generations. *Journal of Animal Science* **75,** 1461–8.

SAS Institute Inc. (1985) *SAS User's Guide: Basics*. SAS, Cary, NC.

Stephens J.C., Briscoe D. & O'Brien S.J. (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics* **55,** 809–24.