Theor Appl Genet (1996) 92: 524–531

R. Cheng · Y. Takano · A. Saito · Y. Ukai

Estimation of unbiased recombination values in the presence of misclassification of a trait using RFLP maps

Received: 7 September 1995 / Accepted: 15 December 1995

Abstract The effect of misclassification of phenotypes of a trait on the estimation of recombination value was investigated. The effect was larger for closer linkage. If a locus is dominant and linked with the misclassfied trait locus in the repulsion phase, then the effect on the recombination value between the two loci is largest. A method for estimating the unbiased recombination value and the misclassification rate using maximum likelihood associated with an EM algorithm is also presented. This method was applied to a numerical example from rice genome data. It was concluded that the present method combined with the metric multi-dimensional scaling method is useful for the detection of misclassified markers and for the estimation of unbiased recombination values.

Key words RFLP linkage map · Recombination value · Misclassification · Maximum likelihood · EM algorithm

Introduction

In the construction of a linkage map of the rice genome (Saito et al. 1991) using restriction fragment length polymorphisms (RFLP), morphological physiological traits and isozymes, it was found that some loci relating to traits and an isozyme marker showed a contradictory relationship concerning the estimated gene order. In the case of three loci *A*, *B* and *C*, for instance, the order *A-B-C* was estimated from multiloci LOD scores, but the map distance *A-C* was significantly lower than the sum of distances between *A-B* and *B-C*. The order of the markers was also confirmed by the application of the metric multi-dimensional scaling (MDS) method of Torgerson (1952) adopted in a microcomputer program MAPL developed by Ukai et al. (1990, 1995). But a marked deviation from the expected linearity of the relative positions of the markers in terms of map distance on a scatter diagram of the 1st and 2nd principal axes of MDS was exhibited. It was suggested that such deviations may be related to overestimation of the recombination values between a trait locus (*B*) and its flanking markers (*A* and *C*), which are mostly RFLP markers, and that this overestimation may be due to misclassification with respect to the trait *B*. Such a trait can be quantitative or qualitative, but is expressed in binary form. Even phenotypes of a qualitative trait can be misclassified due to incomplete penetration of phenotypes, phenocopy, and other causes in the survey of a segregating generation. Polygenically controlled quantitative traits are not usually targets for mapping but rather for QTL analysis. But quantitative traits controlled by a single locus, e.g. dwarfness, earliness and leaf color, can be used for mapping just like an ordinary marker. Such quantitative traits may be much more subject to misclassification as compared with qualitative traits, the degree of misclassification depending on the magnitude of the gene effect relative to the environmental effect.

In this paper we show firstly how the estimated recombination value is biased if misclassification of a trait is involved in segregation data, and secondly we present a method for the estimation of the unbiased recombination value as well as for the misclassification rate.

Communicated by J. W. Snape

R. Cheng (✉) · Y. Takano · Y. Ukai
Laboratory of Biometrics,
Division of Agriculture and Agricultural Life Sciences,
The University of Tokyo, Yayoi 1-1-1, Bunkyo,
Tokyo 113, Japan

A. Saito
Kyushu Agricultural Experimental Station MAFF,
Nishigoshi, Kikuchi, Kumamoto 861-11, Japan

Theory

Effect of misclassification on the estimate of recombination value

There are a number of quantitative traits which are under single-gene control. When the gene effect is sufficiently

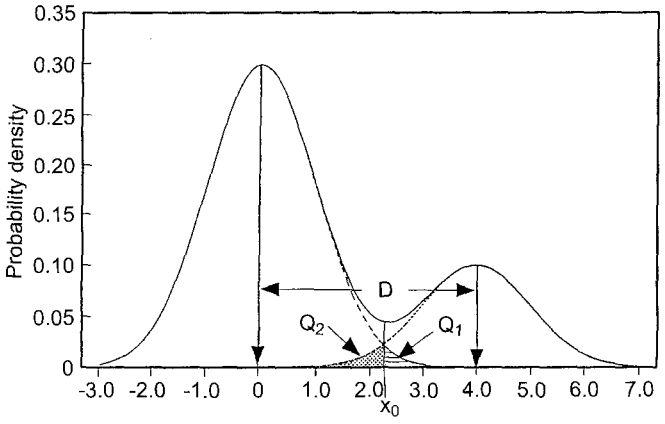

Fig. 1 A schematic representation of misclassification. Normal distributions for dominant (left) and recessive (right) phenotypic groups of individuals with misclassified portions Q_1 for the dominant and Q_2 for the recessive group, and with the difference between means D. Areas under the normal distribution curve for dominant and recessive groups are in a ration of 3:1; x_0 is the intersection of the two curves

large as compared with the environmental variation, the frequency distribution for the phenotypes in a segregating generation can show two or three discontinuous curves related to dominant and co-dominant segregation, respectively. In some cases, however, the segregation of a single locus controlling the trait may not be discrete but rather to involve the overlap of groups of individuals belonging to different genotypes. In such cases, a classification by phenotypes of individuals in the overlapping portion is impossible, though researchers often obtain segregation ratios by setting a boundary. Unless the degree of overlapping is considerable, the bias in the estimation of the segregation ratio invoked by the adoption of such a procedure may not be large, since the frequencies of misclassified individuals are to some degree cancelled between the two groups. In the estimation of recombination values, however, such misclassification may lead to a marked bias.

Firstly, we examine the effect of misclassification at a locus, say *B*, on the estimation of recombination value between *B* and a locus *A* which is linked to *B*. We restrict our consideration to an F_2 generation derived from a cross between two pure lines. We suppose that the phenotype at the locus *B* is quantitative and influenced by environmental fluctuation. For simplicity, let *B* be completely dominant over b, and hence a 3:1 segregation ratio for the two groups of individuals with dominant (*BB* and *Bb*) and recessive (*bb*) phenotypes is expected in the F_2. With respect to segregation at the locus *A*, three cases were investigated; (1) *A* is dominant over *a* and the F_1 is in coupling phase, (2) *A* is dominant over *a* and the F_1 is in repulsion phase and (3) *A* is co-dominant with *a* and 1:2:1 ratio is expected for segregation at the locus. We assume that the phenotypic values of a dominant (x_1) and recessive (x_2) individual follow normal distributions $N_1(0,1)$ and $N_2(D,1)$, respectively (Fig. 1), where D is positive and represents the difference between the means of dominant and recessive groups of individuals. The distribution of phenotypes at the locus *B* is a compound distribution of dominant and recessive groups of individuals:

$$f(x) = \frac{3}{4\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) + \frac{1}{4\sqrt{2\pi}} \exp\left[\frac{-(x-D)^2}{2}\right].$$

Let x_0 be the phenotypic value at the point where the expected distribution curves for the dominant and recessive groups cross. Then the following relationship holds,

$$\frac{3}{4\sqrt{2\pi}} \exp\left(\frac{-x_0^2}{2}\right) = \frac{1}{4\sqrt{2\pi}} \exp\left[\frac{-(x_0-D)^2}{2}\right]. \quad (1)$$

The solution of this equation is x_0=D/2+(ln3)/D.

Now suppose that one divides the F_2 individuals, taking this cross point as the boundary between the dominant and recessive groups of phenotypes. Then misclassification of individuals with respect to genotypes occurs in the tail part of the distribution beyond this point. Let the rate of misclassification of dominant individuals into the recessive, and of recessive individuals into the dominant, group be Q_1 and Q_2, respectively (Fig. 1), then

$$Q_1 = \frac{1}{\sqrt{2\pi}} \int_{x_0}^{\infty} \exp\left(\frac{-u^2}{2}\right) du \quad (2)$$

$$Q_2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x_0-D)} \exp\left(\frac{-u^2}{2}\right) du. \quad (3)$$

Q_1 and Q_2 are functions of D and easily obtained from a cumulative normal distribution. Let the recombination value between the markers *A* and *B* be *r*, then the expected frequencies of the four genotypes *AB*, *Ab*, *aB* and *ab* of gametes which are derived from the F_1 in coupling phase (*AB/ab*) are (1-*r*)/2, *r*/2, *r*/2 and (1-*r*)/2, respectively, and those from the F_1 in repulsion (*Ab/aB*) are *r*/2, (1-*r*)/2, (1-*r*)/2 and *r*/2. The expected frequencies in the F_2 of six genotypes for co-dominance and four genotypes for dominance, with or without misclassification are shown in Table 1. If misclassification, is present, the proportion Q_1 of the individuals which are dominant with respect to the locus *B* would be erroneously classified as recessive, and conversely the proportion Q_2 of the recessive individuals as would be classified dominant. Hence, the expected frequencies of the seemingly dominant individuals would be $f_i(1-Q_1) + f_{i+1}Q_2$ (i=odd) and those of recessive individuals $f_i(1-Q_2) + f_{i-1}Q_1$ (i=even). These frequencies will be denoted as g_i hereafter.

Recombination values were estimated by the maximum-likelihood method based on the frequencies g_i for varying values of D and the true recombination value *r*. The procedures of estimating recombination value with the maximum-likelihood method were introduced by several researchers (Fisher 1925; Mather 1938; Bailey 1961). The recombination value can be estimated by maximizing the logarithm of the likelihood (L). Here the observed numbers of individuals a_i were replaced by their expectation

Table 1 Expected segregation frequencies of F_2 phenotypes with respect to RFLP (*A-a*) and a trait (*B-b*) with or without misclassification of the trait for dominance and co-dominance of the *A-a* locus

Co-dominant			Dominant				With misclassification
Phenotype (1:2:1)	Expected freq. without misclassification		Phenotype (3:1)	Without misclassification: Coupling[a]	Repulsion		
*AAB_*	$1-r^2$ [b]	(f_1)	*A_B_*	$3-2r+r^2$	$2+r^2$	(f_1)	$g_1=(1-Q_1)f_1+f_2Q_2$
AAbb	r^2	(f_2)	*A_bb*	$r(2-r)$	$1-r^2$	(f_2)	$g_2=(1-Q_2)f_2+f_1Q_1$
*AaB_*	$2(1-r+r^2)$	(f_3)	*aaB_*	$r(2-r)$	$1-r^2$	(f_3)	$g_3=(1-Q_1)f_3+f_4Q_2$
Aabb	$2r(1-r)$	(f_4)	*aabb*	$(1-r)^2$	r^2	(f_4)	$g_4=(1-Q_2)f_4+f_3Q_1$
*aaB_*	$r(2-r)$	(f_5)					$g_5=(1-Q_1)f_5+f_6Q_2$
aabb	$(1-r)^2$	(f_6)					$g_6=(1-Q_2)f_6+f_5Q_1$

[a] Linkage phase between the loci *A-a* and *B-b*
[b] All frequencies are multiplied by 1/4

Table 2 Estimated recombination values between (*A-a*) and (*B-b*) in the presence of misclassification

True recombination value	Co-dominant (*A-a*)				Dominant (*A-a*)							
					Coupling				Repulsion			
	D=3	D=4	D=5	D=6	D=3	D=4	D=5	D=6	D=3	D=4	D=5	D=6
0.01	0.073	0.032	0.016	0.011	0.066	0.029	0.015	0.011	0.181	0.108	0.058	0.029
0.10	0.152	0.118	0.105	0.101	0.147	0.116	0.104	0.101	0.204	0.145	0.115	0.103
0.20	0.240	0.214	0.204	0.201	0.236	0.212	0.203	0.201	0.260	0.223	0.207	0.202
0.30	0.327	0.309	0.303	0.301	0.325	0.309	0.302	0.301	0.334	0.312	0.304	0.301
0.40	0.414	0.405	0.401	0.400	0.413	0.405	0.401	0.400	0.415	0.405	0.402	0.400

ng_i; the results are shown in Table 2. In the presence of misclassification, recombination values were always overestimated. The smaller the true recombination value, the larger was the bias of an estimated recombination value. Among the three cases investigated with respect to the dominance relationship and linkage phase of F_1, the deviation was most conspicuous when the locus *A* is dominant and the F_1 is in a repulsion phase. Naturally, as the distance (D) between the two group means increases, the bias of the estimated recombination value decreased.

Estimation of unbiased recombination value and misclassification rate

We propose here a method for the estimation of unbiased recombination values in the presence of misclassification. Suppose *B-b* is the marker being misclassified and *A-a* and *C-c* are its flanking markers without misclassification, with the order of the loci being *A-B-C*. No chiasma interference is assumed. Let the recombination values between *A* and *B*, between *B* and *C*, and between *A* and *C* be r_1, r_2 and r_{1+2}, respectively. Also let Q_1 and Q_2 be the rates of misclassification of the dominant individuals into the recessive group and of the recessive individuals into the dominant group, respectively, as described above. Unlike the above case, the distribution of phenotypes for the locus *B-b* was not restricted to a normal distribution. The expected frequencies of the eight genotypes, *ABC*, *ABc*, *AbC*, *Abc*, *aBC*, *aBc*, *abC* and *abc*, of gametes derived from an F_1 (*ABC/abc*) with respect to the three loci concerned are $(1-r_1)(1-r_2)/2$, $(1-r_1)r_2/2$, $r_1r_2/2$, $r_1(1-r_2)/2$, $r_1(1-r_2)/2$, $r_1r_2/2$, $(1-r_1)r_2/2$ and $(1-r_1)(1-r_2)/2$, respectively. The expected frequencies f_i (i=1, 2, ...,18) of F_2 phenotypes for the locus *B-b* in the absence of misclassification are functions of r_1 and r_2 as shown in Table 3. If misclassification is involved, then the expected frequencies of the phenotypes g_i (i = 1, 2, ..., 18) can be expressed as a function of r_1, r_2, Q_1, and Q_2 as follows:

$$g_{2k-1}(r_1, r_2, Q_1, Q_2) = f_{2k-1}(r_1, r_2) \times (1 - Q_1) + f_{2k}(r_1, r_2) \times Q_2 \quad \text{(for odd i)} \tag{4a}$$

$$g_{2k}(r_1, r_2, Q_1, Q_2) = f_{2k}(r_1, r_2) \times (1 - Q_2) + f_{2k-1}(r_1, r_2) \times Q_1 \quad \text{(for even i)} \tag{4b}$$

(k = 1, 2, ..., 9).

The right hand sides of the above equations (4a and 4b) contain two parts, $f_i(1-Q_1)$ and $f_{i+1}Q_2$ for odd i, or $f_i(1-Q_2)$ and $f_{i-1}Q_1$ for even i. The former and latter parts correspond respectively to non-misclassified individuals belonging to genotypic class i and individuals which belong to the other genotype as regards the locus *B-b* and misclassified into the class i. If we express the observed number of F_2 individuals with phenotypic class i in the presence of

Table 3 Expected segregation frequencies of F_2 phenotypes with respect to two RFLPs (*A-a*), (*C-c*) and a trait (*B-b*) without misclassification of the trait

Phenotype	*B_(BB, Bb)*		*bb*	
AACC	$(1-r_1)^2(1-r_2)^2+2(1-r_1)(1-r_2)r_1 r_2$	(f_1)	$r_1^2 r_2^2$ [a]	(f_2)
AACc	$2(1-r_1)^2(1-r_2)r_2+2(1-r_1)(1-r_2)^2 r_1 +2(1-r_1)r_1 r_2^2$	(f_3)	$2r_1^2(1-r_2)r_2$	(f_4)
AAcc	$(1-r_1)^2 r_2^2+2(1-r_1)(1-r_2)r_1 r_2$	(f_5)	$r_1^2(1-r_2)^2$	(f_6)
AaCC	$2(1-r_1)(1-r_2)^2 r_1+2r_1^2(1-r_2)r_2 +2(1-r_1)^2(1-r_2)r_2$	(f_7)	$2(1-r_1)r_1 r_2^2$	(f_8)
AaCc	$4(1-r_1)(1-r_2)r_1 r_2+2(1-r_1)^2 r_2^2 +2(1-r_1)^2(1-r_2)^2+2r_1^2(1-r_2)^2 +2r_1^2 r_2^2$	(f_9)	$4(1-r_1)(1-r_2)r_1 r_2$	(f_{10})
Aacc	$2(1-r_1)r_1 r_2^2+2r_1^2(1-r_2)r_2 +2(1-r_1)^2(1-r_2)r_2$	(f_{11})	$2(1-r_1)(1-r_2)^2 r_1$	(f_{12})
aaCC	$r_1^2(1-r_2)^2+2(1-r_1)(1-r_2)r_1 r_2$	(f_{13})	$(1-r_1)^2 r_2^2$	(f_{14})
aaCc	$2r_1^2(1-r_2)r_2+2(1-r_1)(1-r_2)^2 r_1 +2(1-r_1)r_1 r_2^2$	(f_{15})	$2(1-r_1)^2(1-r_2)r_2$	(f_{16})
aacc	$r_1^2 r_2^2+2(1-r_1)(1-r_2)r_1 r_2$	(f_{17})	$(1-r_1)^2(1-r_2)^2$	(f_{18})

[a] All frequencies are multiplied by 1/4

misclassification by a_i ($i = 1, 2, \ldots, 18$), it can be evaluated as a sum of two parts, i.e. $a_i = a_{i,1} + a_{i,2}$. Here:

$$a_{2k-1,1} = a_{2k-1} \times \frac{f_{2k-1}(1-Q_1)}{f_{2k-1}(1-Q_1)+f_{2k}Q_2}; \quad (5a)$$

$$a_{2k-1,2} = a_{2k-1} \times \frac{f_{2k}\,Q_2}{f_{2k-1}(1-Q_1)+f_{2k}Q_2}$$

$$a_{2k,1} = a_{2k} \times \frac{f_2(1-Q_2)}{f_{2k}(1-Q_2)+f_{2k-1}Q_1}; \quad (5b)$$

$$a_{2k,2} = a_{2k} \times \frac{f_{2k-1}Q_1}{f_{2k}(1-Q_2)+f_{2k-1}Q_1}$$

($k = 1, 2, \ldots, 9$), with 2k–1 for odd i, and 2k for even i.

The likelihood can be obtained as

$$e^L \propto [f_1(1-Q_1)]^{a_{1,1}}[f_2Q_2]^{a_{1,2}} \ldots [f_{18}(1-Q_2)^{a_{18,1}}[f_{17}Q_1]^{a_{18,2}}$$
$$= \prod_{k=1}^{9}\{[f_{2k-1}(1-Q_1)]^{a_{2k-1,1}}[f_{2k}Q_2]^{a_{2k-1,2}} [f_{2k}(1-Q_2)]^{a_{2k,1}}[f_{2k-1}Q_1]^{a_{2k,2}}\}.$$

There are four parameters (r_1, r_2, Q_1, and Q_2) to be estimated and we have three equations of score as follows

$$S_{Q_1} = \frac{\partial L}{\partial Q_1} = \frac{1}{Q_1}\sum_{k=1}^{9} a_{2k,2} - \frac{1}{1-Q_1}\sum_{k=1}^{9} a_{2k-1,1} = 0 \quad (6)$$

$$S_{Q_2} = \frac{\partial L}{\partial Q_2} = \frac{1}{Q_2}\sum_{k=1}^{9} a_{2k-1,2} - \frac{1}{1-Q_2}\sum_{k=1}^{9} a_{2k,1} = 0 \quad (7)$$

$$S_{r_j} = \frac{\partial L}{\partial r_j} = \sum_{k=1}^{9}\left(\frac{a_{2k-1,1}+a_{2k,2}}{f_{2k-1}(r_1,r_2)} \times \frac{\partial f_{2k-1}(r_1,r_2)}{\partial r_j} + \frac{a_{2k-1,2}+a_{2k,1}}{f_{2k}(r_1,r_2)} \times \frac{\partial f_{2k}(r_1,r_2)}{\partial r_j}\right) = 0 (j = 1,2). \quad (8)$$

The maximum-likelihood estimates can be obtained by solving these equations. The EM algorithm (Dempster et al. 1977) was used for obtaining the final maximum-likelihood estimates of the parameters. Initially an arbitrarily chosen value in the interval (0, 0.5) is put into each of the parameters r_1, r_2, Q_1, and Q_2 to be estimated. Let $\theta_1 = r_1$, $\theta_2 = r_2$, $\theta_3 = Q_1$ and $\theta_4 = Q_2$, and let n be the total number of F_2 individuals, then an information matrix **I** is given by,

$$\mathbf{I} = \begin{bmatrix} I_{11} & I_{12} & 0 & 0 \\ I_{21} & I_{22} & 0 & 0 \\ 0 & 0 & I_{33} & 0 \\ 0 & 0 & 0 & I_{44} \end{bmatrix}.$$

The elements of the information matrix will be obtained by

$$I_{ij} = -E\left(\frac{\partial^2 L}{\partial\theta_i\,\theta_j}\right) \quad (9)$$

(i, j = 1, 2, 3, 4).

Here $I_{ij} = I_{ji}$, that is the matrix is symmetrical. The diagonal elements are given by,

$$I_{11} = I_{r_1} = -E\left(\frac{\partial^2 L}{\partial r_1^2}\right) = n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_1}\right)^2 \quad (10)$$

$$I_{22} = I_{r_2} = -E\left(\frac{\partial^2 L}{\partial r_2^2}\right) = n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_2}\right)^2 \quad (11)$$

$$I_{33} = I_{Q_1} = -E\left(\frac{\partial^2 L}{\partial Q_1^2}\right) = \frac{3n}{Q_1(1-Q_1)} \quad (12)$$

$$I_{44} = I_{Q_2} = -E\left(\frac{\partial^2 L}{\partial Q_2^2}\right) = \frac{n}{Q_2(1-Q_2)}. \quad (13)$$

Only I_{12} and I_{21} are non-zeros in the off-diagonal elements, and they are given by,

$$I_{12} = I_{21} = -E\left(\frac{\partial^2 L}{\partial r_1\,\partial r_2}\right) = n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_1}\right)\left(\frac{\partial f_i}{\partial r_2}\right). \quad (14)$$

The covariance matrix for the four estimates is given by the inverse of the information matrix as follows

$$\mathbf{I}^{-1} = \begin{bmatrix} I^{11} & I^{12} & 0 & 0 \\ I^{21} & I^{22} & 0 & 0 \\ 0 & 0 & I^{33} & 0 \\ 0 & 0 & 0 & I^{44} \end{bmatrix},$$

where

$$I^{11} = \frac{I_{22}}{I_{11}I_{22} - I_{12}^2} \quad (15)$$

$$I^{22} = \frac{I_{11}}{I_{11}I_{22} - I_{12}^2} \quad (16)$$

$$I^{33} = \frac{1}{I_{33}} = \frac{Q_1(1-Q_1)}{3n} \quad (17)$$

$$I^{44} = \frac{1}{I_{44}} = \frac{Q_2(1-Q_2)}{n} \quad (18)$$

$$I^{12} = I^{21} = \frac{-I_{12}}{I_{11}I_{22} - I_{12}^2} \text{ or } \frac{-I_{21}}{I_{11}I_{22} - I_{21}^2} \quad (19)$$

are the non-zero elements of the covariance matrix; since the maximum-likelihood estimates $\hat{\Theta}$ follow a multi-dimensional normal distribution in large samples with Θ and covariance matrix $\mathbf{I}^{-1}$, the variances of the four estimated parameters ($\hat{r}_1$, $\hat{r}_2$, $\hat{Q}_1$ and $\hat{Q}_2$) are I^{11}, I^{22}, I^{33} and I^{44}, respectively, and the corresponding standard errors are given by $\sqrt{I^{ii}}$ (i = 1, 2, 3, 4).

Table 4 Observed frequencies of F_2 phenotypes with regard to *A* (RFLP No. 180), *B* (sp) and *C* (RFLP No. 44) loci

Phenotype	*BB, Bb*	*bb*
AACC	22	3
AACc	15	0
AAcc	1	0
AaCC	9	2
AaCc	45	2
Aacc	9	1
aaCC	4	0
aaCc	5	6
aacc	1	14
Total	111	28

Numerical example

In the course of mapping RFLP markers in an F_2 from a cross Kasalath×FL144 (Saito et al. 1991), it was suggested that misclassification was involved in the segregation data of a trait (marker No.3) in linkage group XI. The marker was concerned with the length of panicle (*sp*). The segregation data at the *sp* locus and its flanking RFLP markers No.180 and No.44 are shown in Table 4. The locus *sp*, RFLP markers No.180 and No. 44 will be called hereafter *B*, *A*, and *C*, respectively. The method described in the previous section was used to estimate the recombination values r_1 (r_{AB}), r_2 (r_{BC}) and the misclassification rates Q_1 and Q_2. The results are shown in Table 5. The estimated recombination values r_1 and r_2 were 0.0867 and 0.1694, respectively. The recombination value r_{1+2} between the two RFLP markers No. 180 and No. 44 was 0.2267 when calculated using the formula $r_{1+2} = r_1 + r_2 - 2r_1r_2$ (Trow

Table 5 Estimated recombination values and misclassification rates

Linkage group	A pair of loci	Recombination value		Misclassification rate
		Saito et al. (1991)	Estimated by our method	
XI	180-*sp*	0.183 ± 0.038[a]	0.0867 ± 0.0117	Q_1 = 0.0544 ± 0.0111
	sp-44	0.269 ± 0.047	0.1694 ± 0.0147	Q_2 = 0.1152 ± 0.0271
	180-44	0.244 ± 0.031	0.2267[b](0.2419[c])	

[a] Standard errors of the estimates
[b] 0.2267 was calculated by a formula of Trow (1913), $r_{1+2} = r_1 + r_2 - 2r_1r_2$
[c] 0.2419 was calculated by a formula of Kosambi (1944), $r_{1+2} = (r_1 + r_2)/(1 + 4r_1r_2)$

1913) which corresponds to 0.2419 by the formula $r_{1+2} = (r_1 + r_2)/(1 + 4r_1r_2)$ (Kosambi 1944). It can be seen that the estimate of r_{1+2} is close to 0.244 which was directly estimated by Saito et al. (1991). The standard errors of the estimates for r_1 and r_2 by the method presented in the present paper were smaller than those of Saito et al. (1991). The estimated misclassification rates of Q_1, Q_2 were rather high at 0.0544 and 0.1152, respectively.

Discussion

A method for estimating the recombination values between a lethal-factor locus and neighboring molecular markers, and the relative viability of gametes or zygotes affected by the lethal factor in an F_2 population, has been developed by us (unpublished). The effect of misclassification on the estimation of segregation rate and recombination value in mapping was shown to be quite different from that of a lethal factor. In the presence of a lethal factor, the segregation ratios of loci in the vicinity of the factor becomes markedly distorted, while the estimate of recombination value between flanking markers is not affected at all, irrespective of the dominance vs co-dominance segregation of the markers. On the contrary, it was shown here that if misclassification is involved in the segregation data of a trait, the recombination value between the locus of the trait and its flanking markers is always overestimated, while the segregation ratio of the misclassified locus is little affected. The closer the linkage, the greater is the degree of overestimation. The bias in the recombination value differs with dominance vs co-dominance, the phenotypic segregation of the trait, and the linkage phase in the F_1 between the trait locus misclassified and the neighboring markers under consideration, being largest for dominance of the marker with the F_1 in repulsion phase.

Overestimation of recombination values due to misclassification was also shown by Ott (1977) in double-backcross data, assuming the same misclassification error for each genotype of a trait. From numerical comparison it was found that for the same proportion of misclassified individuals the degree of overestimation is smaller in the backcross than in the F_2 with co-dominant or dominant flanking markers.

If the proportion of individuals misclassified is low, the order of the trait locus and the two flanking markers would be correctly estimated by a simple three-point test of linkage. But if large, it may alter the estimated order itself. The map distances estimated by an ordinary method for the interval *A-B*, *B-C* and *A-C* in the numerical example shown above were 19.2, 30.1 and 26.7 cM, respectively (Saito et al. 1991), the distance of *A-C* being much smaller than the sum of the distances of *A-B* and *B-C* and even smaller than that of *B-C*. If a three-point test were adopted, one would conclude that the location of *A* was intermediate between *B* and *C*.

Analysis by the metric multi-dimensional scaling method is useful for detecting a locus with misclassification, and obtaining the correct order of the loci, including the misclassified one (Ukai et al. 1990, 1995). Using map distances between the pairs of markers in all possible combinations as a measure of dissimilarity, or "relative distance" in terms of the method, the relative positions of the markers can be described by the coordinates of a space of the dimension which is equal to the number of markers to be examined. Variations in the positions of the markers in the original space can usually be summarized in a space of much reduced dimension. Since the markers belonging to the same linkage group are located linearly on a chromosome, the major part of the variations can be accounted for by the first principal components, and the markers are expected to show positions along the first principal axis in the analysis. If a locus suffers from misclassification, the map distances between the locus and neighboring markers are more or less overestimated and hence the locus is expected to exhibit a location deviating from the first principal axis. Estimation of correct order by multi-dimensional scaling is valid only for cases where most markers are scored without error, and only a few are misclassified at a certain rate. If all markers to be mapped are mis-scored at approximately the same rate, we cannot detect the misclassification by this method.

If we investigate the segregation in F_3 lines derived from F_2 individuals for which misclassification was suspected, we can confirm whether the F_2 individuals were correctly classified or not. But usually it is uncertain which individuals in the F_2 are suspected of being misclassified, and so a survey of segregation in the F_3 would be necessary for a large proportion, if not all, of the F_2 individuals. In practical situations, particularly in field surveys where environmental conditions may vary with years, the investigation of segregation for many F_3 lines is not always feasible. For instance, a disease symptom which was clearly observed in the field in one year cannot be always expected to appear in the next year.

It was shown in the present study that if linkage between two loci is not so close, the effect of misclassification on the estimation of recombination value is not conspicuous. This may be the reason why the problem of misclassification in linkage studies was not so important in classical mapping in which only loci of a limited number of traits were used as markers and linkage between any two loci was usually loose. In the mapping of DNA markers segregation data for multi-loci is available and multi-point cosegregation data make it possible to detect any misclassification involved. Further, it is for such a detailed linkage map that misclassification exerts its influence markedly on the estimation of recombination value. In the integration of a DNA polymorphism linkage map and a trait map the effect of misclassification cannot be dismissed. The present method combined with the metric multi-dimensional scaling method can be effectively used to determine the order of the loci and to estimate the unbiased recombination value in such a case. The proposed method is based on an F_2 generation, but can easily be extended to other segregating generations such as backcross and dihaploid.

Appendix

Elements of the information matrix

For r_1 and r_2,

$$-E\left(\frac{\partial^2 L}{\partial r_l\,\partial r_m}\right)=-E\left[\frac{\partial}{\partial r_m}\left(\frac{\partial L}{\partial r_l}\right)\right]$$

$$=-\frac{\partial}{\partial r_m}\sum_{k=1}^{9}\left[\frac{a_{2k-1,1}+a_{2k,2}}{f_{2k-1}(r_l,r_m)}\times\frac{\partial f_{2k-1}(r_l,r_m)}{\partial r_l}\right.$$

$$\left.+\frac{a_{2k-1,2}+a_{2k,1}}{f_{2k}(r_l,r_m)}\times\frac{\partial f_{2k}(r_l,r_m)}{\partial r_l}\right]$$

$$=\sum_{k=1}^{9}\left[\frac{a_{2k-1,1}+a_{2k,2}}{f_{2k-1}^2}\times\left(\frac{\partial f_{2k-1}}{\partial r_l}\right)\left(\frac{\partial f_{2k-1}}{\partial r_m}\right)\right.$$

$$-\frac{a_{2k-1,1}+a_{2k,2}}{f_{2k-1}}\times\frac{\partial^2 f_{2k-1}}{\partial r_l\,\partial r_m}$$

$$+\frac{a_{2k-1,2}+a_{2k,1}}{f_{2k}^2}\times\left(\frac{\partial f_{2k}}{\partial r_l}\right)\left(\frac{\partial f_{2k}}{\partial r_m}\right)$$

$$\left.-\frac{a_{2k-1,2}+a_{2k,1}}{f_{2k}}\times\frac{\partial^2 f_{2k}}{\partial r_l\,\partial r_m}\right].$$

We substitute the expected values for the observed ones, i.e. ng_i for a_i, then

$$a_{2k-1,1}=a_{2k-1}\times\frac{f_{2k-1}\times(1-Q_1)}{g_{2k-1}}=nf_{2k-1}(1-Q_1).$$

$$a_{2k-1,2}=a_{2k-1}\times\frac{f_{2k}Q_2}{g_{2k-1}}=nf_{2k}Q_2$$

$$a_{2k,1}=a_{2k}\times\frac{f_{2k}(1-Q_2)}{g_{2k}}=nf_{2k}(1-Q_2),$$

$$a_{2k,2}=a_{2k}\times\frac{f_{2k-1}Q_1}{g_{2k}}=nf_{2k-1}Q_1$$

so that

$$a_{2k-1,1}+a_{2k,2}=nf_{2k-1}$$

$$a_{2k-1,2}+a_{2k,1}=nf_{2k},$$

this gives

$$-E\left(\frac{\partial^2 L}{\partial r_l\,\partial r_m}\right)=n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_l}\right)\left(\frac{\partial f_i}{\partial r_m}\right)-n\sum_{i=1}^{18}\frac{\partial^2 f_i}{\partial r_l\,\partial r_m}.$$

Since

$$n\sum_{i=1}^{18}\frac{\partial^2 f_i}{\partial r_l\,\partial r_m}=\frac{\partial^2}{\partial r_l\,\partial r_m}\sum_{i=1}^{18}f_i=0,$$

we finally obtain

$$-E\left(\frac{\partial^2 L}{\partial r_l\,\partial r_m}\right)=n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_l}\right)\left(\frac{\partial f_i}{\partial r_m}\right).$$

Particularly, when l = m = 1,

$$I_{11}=I_{r_1}=n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_l}\right)^2,$$

l = m = 2,

$$I_{22}=I_{r_2}=n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_2}\right)^2,$$

and l = 1, m = 2 or l = 2, m = 1,

$$I_{12}=I_{21}=n\sum_{i=1}^{18}\frac{1}{f_i}\left(\frac{\partial f_i}{\partial r_l}\right)\left(\frac{\partial f_i}{\partial r_2}\right).$$

For Q_1 and Q_2

$$-E\left(\frac{\partial^2 L}{\partial Q_1^2}\right)=\frac{1}{Q_1^2}\sum_{k=1}^{9}a_{2k,2}+\frac{1}{(1-Q_1)^2}\sum_{k=1}^{9}a_{2k-1.1}.$$

We substitute the expected values for the observed ones as above,

$$-E\left(\frac{\partial^2 L}{\partial Q_1^2}\right)=\frac{nQ_1}{Q_1^2}\sum_{k=1}^{9}f_{2k-1}+\frac{n(1-Q_1)}{(1-Q_1)^2}\sum_{k=1}^{9}f_{2k-1}$$

$$=\left(\frac{n}{Q_1}+\frac{n}{1-Q_1}\right)\sum_{k=1}^{9}f_{2k-1}.$$

Here

$$\sum_{k=1}^{9}f_{2k-1}=3,$$

we obtain

$$-E\left(\frac{\partial^2 L}{\partial Q_1^2}\right)=\frac{3n}{Q_1(1-Q_1)}.$$

Similarly

$$-E\left(\frac{\partial^2 L}{\partial Q_2^2}\right)=\frac{nQ_2}{Q_2^2}\sum_{k=1}^{9}f_{2k}+\frac{n(1-Q_2)}{(1-Q_2)^2}\sum_{k=1}^{9}f_{2k}$$

$$=\left(\frac{n}{Q_2}+\frac{n}{1-Q_2}\right)\sum_{k=1}^{9}f_{2k},$$

where

$$\sum_{k=1}^{9}f_{2k}=1,$$

then

$$-E\left(\frac{\partial^2 L}{\partial Q_2^2}\right)=\frac{n}{Q_2(1-Q_2)}$$

is obtained.

Acknowledgments This study was partly supported by the Japan Society for the Promotion of Science (JSPS) for postdoctoral fellowship of foreign researchers. We gratefully acknowledge all of the members in the group involved in constructing the rice RFLP linkage map (Saito et al. 1991) for providing us with valuable data.

References

Bailey NTJ (1961) Introduction to the mathematical theory of genetic linkage. Oxford University Press, London

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38

Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, London

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172–175

Mather K (1938) Measurement of linkage in heredity. 1st edn reprinted, Methuen and Co., London

Saito A, Yano M, Kishimoto N, Nakagahra M, Yoshimura A, Saito K, Kuhara S, Ukai Y, Kawase M, Nagamine T, Yoshimura S, Ideta O, Ohsawa R, Hayano Y, Iwata N, Sugiura M (1991) Linkage map of restriction fragment length polymorphism loci in rice. Japan J Breed 41: 665–670

Ott J (1977) Linkage analysis with misclassification at one locus. Clin Genet 12:119–124

Torgerson WS (1952) Multi-dimentional scaling. I. Theory and method. Psychometrika 17:401–419

Trow AH (1913) Forms of reproduction: primary and secondary. J Genet 2:313–324

Ukai Y, Ohsawa R, Saito A (1990) Automatic determination of the order of RFLPs in a linkage group by a metric multi-dimensional scaling method. Japan J Breed 40 (Suppl. 2): 302–303

Ukai Y, Ohsawa R, Saito A, Hayashi T (1995) MAPL: a package of computer programs for construction of DNA polymorphism linkage maps and analysis of QTLs. Breed Sci 45:139–142