Linkage Analysis for Misclassified Markers

Chang-Xing Ma[1*], Rameela Chandrasekhar[1], Rongling Wu[2]

[1] *Department of Biostatistics, University at Buffalo, Buffalo, NY 14214 USA*

[2] *Department of Statistics, University of Florida, Gainesville, Florida 32611 USA*

[*] *Corresponding author*

Abstract

Quantitative trait loci (QTL) mapping relates portions of the gene to a phenotypic trait and can be used to explain genetic variance. The accuracy of QTL mapping is greatly affected by linkage maps which in turn are affected by factors like viability and misclassification in markers. Markers are usually misclassified due to human errors in failing to identify the phenotypes correctly or due to the failure of a gene to produce its expected consequences. Misclassification of markers results in discrepancy or a biased estimate of the gene loci associate with a trait. In this paper we have described a statistical method incorporating misclassified markers for linkage analysis. Statistical properties on applying the method was examined using a real example and from simulation studies for a One-Gene and a Two-Gene model. The results showed the recombination fraction to be affected by the degree of misclassification. Understanding misclassification and incorporating it in linkage analysis may help in a more accurate QTL mapping.

Keywords: misclassification, QTL mapping , linkage analysis.

Introduction

Misclassification is one of the main sources of disturbance in linkage experiments. A typical occurence of misclassification is when an organism, known from its antecedents to be carrying a dominant gene **A** and to be of genotype **Aa**, fails to exhibit the dominant character and is misclassified as the recessive **aa**. A similar situation can also arise with recessive characters. Thus an organism homozygous for the recessive gene **a** may fail to show the recessive character, and so be wrongly classified as a dominant. This happens only in a certain proportion of cases. Bailey (1961) discussed several models for analyzing misclassified markers in a great detail. Effect such as misclassification can be incorporated into linkage analysis using modified methods.

In this article we describe a method for estimating the recombination fraction between markers subject to misclassification by introducing into the statistical analysis additional parameters to be estimated. The method incorporates maximum likelihood and EM algorithm

to estimate the parameters. Linkage analysis is widely studied using the F_2 or backcross pedigree since both recombinant and nonrecombinant gamete types can be counted. When two homozygous individuals are crossed, it results in a heterozygous F_1 offspring. This F_1 progeny, when backcrossed to each of their parents generate two backcrosses while crossing with each other produces the F_2 generation. In order to understand the influence of misclassification on linkage analysis, simulation studies are conducted for both the backcross and F_2 population.

Analysis of linkage is performed using two or more loci. The statistical approach for estimating and testing the recombinant fraction between two different markers is known as two-point analysis whereas the analysis of linkage involving more than two markers is commonly referred to as multiple point linkage analysis. Though the two-point analysis is most commonly used, multiple point analysis is seen to have a greater advantage concerning efficiency and power. It also increases the precision of the estimates of the recombination fractions when markers are not fully informative (Thompson 1984; Wu et al. 2002) and also provides a way of determining the optimal order of different markers. But ease of implementation and less computation time dictates the use of two-point analysis in our study.

Our simulations deal with a model in which only one of the two markers **A** and **B** subject to linkage analysis is affected by misclassification. The new method is also compared using a real example for the more general case in which both markers subject to linkage analysis are affected by misclassification.

One-gene Model

In practice, it is possible to misclassify one genotype as the other due to human errors. For example, in a double backcross, there may be a proportion of λ of allele A that is misclassified as a (irrespective of whether they are associated with B or b). The appropriate expected frequencies and observed numbers of genotypes or gametes in a backcross population are shown below:

Gamete	Expected	Observed
AB	$\frac{1}{2}(1-r)(1-\lambda)$	n_1
Ab	$\frac{1}{2}r(1-\lambda)$	n_2
aB	$\frac{1}{2}[r+\lambda(1-r)]$	n_3
ab	$\frac{1}{2}[(1-r)+\lambda r]$	n_4
Total	1	n

The expectations are derived because a proportion λ of each of the first two gamete classes involving A has been transferred to the corresponding one involving a, but with the same B or b classification.

The maximum likelihood estimation of the two parameters r and λ is straightforward. We have

$$\hat{r} = \frac{n_2(n_1+n_3)}{n_2(n_1+n_3)+n_1(n_2+n_4)},$$

$$\hat{\lambda} = \frac{n_3n_4-n_1n_2}{(n_1+n_3)(n_2+n_4)}.$$

The sampling variances of these estimators are:

$$\mathrm{var}(\hat{r}) = \frac{2r(1-r)}{n}\left[\frac{1}{1-\lambda}-2r(1-r)\right],$$

$$\mathrm{var}(\hat{\lambda}) = \frac{2(1-\lambda)^2}{n}\left[\frac{\lambda}{1-\lambda}+2r(1-r)\right].$$

Although A is misclassified as a, it is possible to have the reverse misclassification, i.e., $a \rightarrow A$. The formulae for this alternative pattern can be derived by suitably changing the observational symbols n_1, n_2, n_3 and n_4.

When misclassification occurs in the formation of genotypes in the F_2, similar formulas can be derived to explore its influences on the estimates of the recombination fraction. In so doing, we assume that such misclassification arises from the ambiguity of individual alleles. Thus, the above formulation for the backcross can be extended to model misclassification in the F_2. The expected frequencies of the F_2 genotypes after gene **A** is misclassified are expressed as

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \left[\begin{array}{ccc} \frac{1}{4}(1-\lambda)^2(1-r)^2 & \frac{1}{2}(1-\lambda)^2 r(1-r) & \frac{1}{4}(1-\lambda)^2 r^2 \\ \frac{1}{2}[(1-\lambda)(1-r)(r+\lambda(1-r))] & \frac{1}{2}(1-\lambda)[r^2+(1-r)^2+2\lambda r(1-r)] & \frac{1}{2}[(1-\lambda)r(1-r+\lambda r)] \\ \frac{1}{4}[r+\lambda(1-r)]^2 & \frac{1}{2}[\lambda r^2+\lambda(1-r)^2+(1+\lambda^2)r(1-r)] & \frac{1}{4}[(1-r)+\lambda r]^2 \end{array}\right] \end{array}. \qquad (1)$$

We derive the EM algorithm to estimate the recombination fraction r. It is not difficult to find the expected numbers of recombinants within each cell in the above matrix (1), expressed as

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 2 \\ \phi_1 & 2\phi_2 & \phi_3 \\ 2\phi_4 & \phi_5 & \phi_6 \end{array} \right] \end{array},$$

where

$$\phi_1 = \frac{r(1-r)}{r(1-r)+\lambda(1-r)^2}, \quad (2)$$

$$\phi_2 = \frac{r^2+\lambda r(1-r)}{r^2+(1-r)^2+2\lambda r(1-r)}, \quad (3)$$

$$\phi_3 = 1+\frac{\lambda r^2}{r(1-r)+\lambda r^2}, \quad (4)$$

$$\phi_4 = \frac{r}{r+\lambda(1-r)}, \quad (5)$$

$$\phi_5 = \frac{2\lambda r^2+(1+\lambda^2)r(1-r)}{\lambda r^2+\lambda(1-r)^2+(1+\lambda^2)r(1-r)}, \quad (6)$$

$$\phi_6 = \frac{2\lambda r}{(1-r)+\lambda r}. \quad (7)$$

The recombination fraction can be estimated using equation

$$\hat{r} = \frac{1}{2n}(n_{21}+2n_{20}+\phi_1 n_{12}+2\phi_2 n_{11}+\phi_3 n_{10}+2\phi_4 n_{02}+\phi_5 n_{01}+\phi_6 n_{00}). \quad (8)$$

In the E step, the expected numbers of recombinants for each genotype are calculated using equations (2) – (7). These expected numbers are used to update the estimate of r with equation (8) in the M step. These two steps are iterated until r converges to a stable value.

The MLE of λ in the F_2 is given by solving the third-order polynomial equation

$$\frac{n_1}{1-\lambda} = \frac{n_2}{\lambda+\frac{r}{1-r}}+\frac{n_3}{\lambda+\frac{1-r}{r}}+\frac{n_{11}}{\lambda+\frac{r^2+(1-r)^2}{2r(1-r)}}. \quad (9)$$

Two-gene Model

When two genes are both misclassified, we should introduce an additional proportion for allele B misclassified as b. Let the misclassified proportions be λ_1 and λ_2 for markers **A** and **B**, respectively. Assuming that these two proportions are independent, we have the expected numbers of each of the four backcross genotypes, along with their observations, as follows

Gamete	Expected	Observed
AB	$\frac{1}{2}(1-\lambda_1)(1-\lambda_2)(1-r)$	n_1
Ab	$\frac{1}{2}(1-\lambda_1)[r+\lambda_2(1-r)]$	n_2
aB	$\frac{1}{2}(1-\lambda_2)[r+\lambda_1(1-r)]$	n_3
ab	$\frac{1}{2}[r(\lambda_1+\lambda_2)+(1-r)(1+\lambda_1\lambda_2)]$	n_4
Total	1	n

The MLEs of the recombination fraction and the proportions of misclassification can be derived as

$$\hat{r} = 1 - \frac{n_1 n}{2(n_1+n_2)(n_1+n_3)},$$

$$\hat{\lambda}_1 = \frac{(n_3+n_4)-(n_1+n_2)}{n},$$

$$\hat{\lambda}_2 = \frac{(n_2+n_4)-(n_1+n_3)}{n}.$$

The sampling variances of the MLEs of r, λ_1 and λ_2 is

$$\mathrm{var}(\hat{r}) = \frac{nn_1[(n(n_1^2+n_2n_3)-n_1(n_1+n_2)(n_1+n_3)]}{4(n_1+n_2)^3(n_1+n_3)^3},$$

$$\mathrm{var}(\hat{\lambda}_1) = \frac{1-{\lambda_1}^2}{n},$$

$$\mathrm{var}(\hat{\lambda}_2) = \frac{1-{\lambda_2}^2}{n}.$$

It is difficult to estimate the recombination fraction when both markers are misclassified in the F_2. With the assumption that misclassification arises from the ambiguity of individual alleles, we derive the expected frequencies of the F_2 genotypes after both genes **A** and **B** are misclassified, expressed as

Genes		Expected Frequencies	Obs.
AA	BB	$(1-\lambda_1)^2(1-\lambda_2)^2(1-r)^2/4$	n_{22}
AA	Bb	$(1-\lambda_1)^2(1-\lambda_2)(1-r)(r+\lambda_2(1-r))/2$	n_{21}
AA	bb	$(1-\lambda_1)^2(r+\lambda_2(1-r))^2/4$	n_{20}
Aa	BB	$(1-\lambda_1)(1-\lambda_2)^2(1-r)(r+\lambda_1(1-r))/2$	n_{12}
Aa	Bb	$(1-\lambda_1)(1-\lambda_2)[2r(1-r)(\lambda_1+\lambda_2)+r^2+(1-r)^2(2\lambda_1\lambda_2+1)]$	n_{11}
Aa	bb	$(1-\lambda_1)(r+\lambda_2(1-r))((1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r)/2$	n_{10}
aa	BB	$(1-\lambda_2)^2(r+\lambda_1(1-r))^2/4$	n_{02}
aa	Bb	$((1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r)(r+\lambda_1(1-r))(1-\lambda_2)/2$	n_{01}
aa	bb	$((1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r)^2/4$	n_{00}

(10)

The expected numbers of recombinants within each cell in the above table (10) are expressed as

$$\begin{array}{c} \\ AA \\ Aa \\ aa \end{array} \begin{array}{c} \begin{array}{ccc} BB & Bb & bb \end{array} \\ \left[\begin{array}{ccc} 0 & \phi_1 & 2\phi_2 \\ \phi_3 & 2\phi_4 & \phi_5 \\ 2\phi_6 & \phi_7 & \phi_8 \end{array} \right] \end{array},$$

where

$$\phi_1 = \frac{r(1-r)}{r(1-r)+\lambda_2(1-r)^2}, \tag{11}$$

$$\phi_2 = \frac{r}{r+\lambda_2(1-r)}, \tag{12}$$

$$\phi_3 = \frac{r(1-r)}{r(1-r)+\lambda_1(1-r)^2}, \tag{13}$$

$$\phi_4 = \frac{r(1-r)(\lambda_1+\lambda_2)+r^2}{2r(1-r)(\lambda_1+\lambda_2)+r^2+(1-r)^2(2\lambda_1\lambda_2+1)}, \tag{14}$$

$$\phi_5 = \frac{2r^2(\lambda_1+\lambda_2)+(\lambda_2^2+2\lambda_1\lambda_2+1)r(1-r)}{[(1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r](r+\lambda_2(1-r))}, \tag{15}$$

$$\phi_6 = \frac{r}{r+\lambda_1(1-r)}, \tag{16}$$

$$\phi_7 = \frac{2r^2(\lambda_1+\lambda_2)+(\lambda_1^2+2\lambda_1\lambda_2+1)r(1-r)}{[(1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r](r+\lambda_1(1-r))}, \tag{17}$$

$$\phi_8 = \frac{2(\lambda_1+\lambda_2)r}{(1-r)(1+\lambda_1\lambda_2)+(\lambda_1+\lambda_2)r}. \tag{18}$$

The recombination fraction can be estimated using equation

$$\hat{r} = \frac{1}{2n}(\phi_1 n_{21} + 2\phi_2 n_{20} + \phi_3 n_{12} + 2\phi_4 n_{11} + \phi_5 n_{10} + 2\phi_6 n_{02} + \phi_7 n_{01} + \phi_8 n_{00}). \qquad (19)$$

The EM algorithm is formulated to estimate r. In the E step, the expected numbers of recombinants for each genotype are calculated using equations (11) – (18). These expected numbers are used to update the estimate of r with equation (19) in the M step. These two steps are iterated until r converges to a stable value.

The MLEs of λ_1 and λ_2 in the F_2 are given by solving the third-order polynomial equations

$$\frac{n_1}{1-\lambda_1} = \frac{n_2}{\lambda_1 + \frac{r}{1-r}} + \frac{n_3}{\lambda_1 + \frac{1-r+\lambda_2 r}{(1-r)\lambda_2 + r}} + \frac{n_{11}}{\lambda_1 + \frac{r^2+(1-r)^2+2\lambda_2 r(1-r)}{2r(1-r)+2\lambda_2(1-r)^2}},$$

$$\frac{m_1}{1-\lambda_2} = \frac{m_2}{\lambda_2 + \frac{r}{1-r}} + \frac{n_3}{\lambda_2 + \frac{1-r+\lambda_1 r}{(1-r)\lambda_1 + r}} + \frac{n_{11}}{\lambda_2 + \frac{r^2+(1-r)^2+2\lambda_1 r(1-r)}{2r(1-r)+2\lambda_1(1-r)^2}},$$

where

$$\begin{aligned}
n_1 &= 2(n_{22} + n_{21} + n_{20}) + n_{12} + n_{11} + n_{10}, \\
n_2 &= n_{12} + 2n_{02} + n_{01}, \\
n_3 &= n_{10} + 2n_{00} + n_{01}, \\
m_1 &= 2(n_{22} + n_{12} + n_{02}) + n_{21} + n_{11} + n_{01}, \\
m_2 &= n_{21} + 2n_{20} + n_{10}.
\end{aligned}$$

The sampling variances of the MLEs of r, λ_1, and λ_2 are

$$\begin{gathered}
\mathrm{var}(\hat{r}) = \\
\{(1-r)[24(1 + \lambda_1^2\lambda_2^2 + \lambda_2^2 - 2\lambda_2^2\lambda_1 - 2\lambda_1 - 2\lambda_2 + \lambda_1^2 + 4\lambda_1\lambda_2 - 2\lambda_1^2\lambda_2)r^4 \\
+(100\lambda_2 - 56\lambda_2^2 + 100\lambda_1 - 56\lambda_1^2 - 240\lambda_1\lambda_2 - 44 - 84\lambda_1^2\lambda_2^2 + 140\lambda_1^2\lambda_2 + 140\lambda_2^2\lambda_1)r^3 \\
+(-144\lambda_2^2\lambda_1 + 42\lambda_1^2 + 42\lambda_2^2 + 110\lambda_1^2\lambda_2^2 - 144\lambda_1^2\lambda_2 + 34 - 76\lambda_2 + 212\lambda_1\lambda_2 - 76\lambda_1)r^2 \\
+(59\lambda_2^2\lambda_1 - 12 + 59\lambda_1^2\lambda_2 - 64\lambda_1^2\lambda_2^2 + 27\lambda_1 - 11\lambda_1^2 - 82\lambda_1\lambda_2 - 11\lambda_2^2 + 27\lambda_2)r \\
+(2 + \lambda_2^2 - 7\lambda_2^2\lambda_1 - 7\lambda_1^2\lambda_2 - 3\lambda_1 + 14\lambda_1^2\lambda_2^2 - 3\lambda_2 + 14\lambda_1\lambda_2 + \lambda_1^2)]\}/ \\
\{2n(1-\lambda_2)(1-\lambda_1)[6(\lambda_1\lambda_2 + 1 - \lambda_1 - \lambda_2)r^2 + 2(3\lambda_1 - 5\lambda_1\lambda_2 - 3 + 3\lambda_2)r \\
+4\lambda_1\lambda_2 - \lambda_1 - \lambda_2 + 2]\}, \\
\mathrm{var}(\hat{\lambda}_1) = \frac{1-\lambda_1^2}{2n}, \\
\mathrm{var}(\hat{\lambda}_2) = \frac{1-\lambda_2^2}{2n}.
\end{gathered}$$

Simulations

The influence of marker misclassification on the estimate of the recombination fraction is investigated through simulation studies. Two markers are simulated for both the backcross and F_2 with different degrees of marker misclassification (λ_1, λ_2). The sample sizes considered are 80 and 200, and there are degrees of linkage ($r = 0.05$ and 0.35). The simulated data are analyzed by both the model (1) that incorporates marker misclassification and the model (2) that does not.

When $\lambda_1 = \lambda_2 = 0$ (i.e., there is no marker misclassification), models 1 and 2 provide similarly good results (Tables 1 and 2). When one of the markers is misclassified to some extent, model 1 quickly ill-behaves, and its estimate is very biased, especially for a tight linkage. For example, when $\lambda_2 = 0.1$, model 1 estimates the true $r = 0.05$ as over 0.09. This is not improved when the size of the sample is increased. Model 1 can generally provide good estimates of r even when both markers are misclassified. The estimation accuracy of r by model 1 can be dramatically increased with increasing sample size. Because model 2 has fewer parameters to be estimated than model 1, the former displays superior estimation precision over the latter. But this advantage of model 1 is not useful given its large biased estimate. The estimation precision of model 1 is increased with increasing sample size.

The backcross and F_2 display similar trends for parameter estimation. But it is observed that the F_2 (Table 2) is better in terms of estimation accuracy and precision than the backcross (Table 1). This may be because the F_2 contains a larger amount of information than the backcross.

A working example

To demonstrate the benefits of incorporating misclassified markers for QTL analysis, we applied the method developed to a real data of backcross population in rice (n=123) and F_2 population in mice. For the rice data, we consider a two-point analysis of 10 markers on rice chromosome 1. The rice marker data consists of a total of 175 polymorphic markers including 146 RFLPs, 8 isozymes, 14 RAPDs and 12 cloned genes. (Huang et al. 1997). We consid-

ered a Two-Gene model where both genes are misclassified using the primary assumption that the misclassified proportions are independent. To demonstrate the effect of misclassified markers on linkage analysis, we obtained an estimate of the recombination fraction and the misclassified proportions along with their standard errors under two models, one that incorporated misclassified markers into the analysis and one that did not. The quantitative trait loci maps were generated for both cases and the differences noted. Reconstruction of the linkage maps with misclassified markers gives us a visual representation of how the QTL maps change on considering misclassification among markers.

For the mice data, we consider a two-point analysis of 6 markers on mice chromosome 1. Cheverud et al.(1996) constructed a linkage map using 75 microsatellite markers in a population of 535 F_2 progeny derived from two strains, the Large(LG/J) and Small(SM/J). The F_2 progeny were measured for body mass at ten weekly intervals starting at age 7 days. The raw weights were corrected for the effects of each covariate due to dam, litter size at birth, and parity but not for the effect due to sex.

In the F_2 population when both markers are misclassified we would have to use the iterative EM algorithm to obtain an estimate of the recombination fraction and to obtain the estimate of the degrees of misclassification, a third order polynomial equation would have to be solved. For the backcross population in rice, it is not very difficult to estimate the parameters and their standard errors. The estimate of the recombination fraction and the degrees of misclassification can be obtained from the maximum likelihood method. From the simulations we can note that the model that does not incorporate misclassification tends to overestimate or underestimate the recombination fraction and hence the linkage map obtained by accounting for misclassification tends to be much shorter or longer.

But on applying the method to the rice data, except for a few, nearly all estimates of recombination fraction incorporating misclassification seem to equal or be much larger than the estimates from the traditional model. This may be due to extreme imbalance in the numbers of genotype observed especially in the last three makers. This may also be due to a low sample size, which from the simulation results can be seen to affect the estimation precision. Due to this, the linkage map using the new method for the rice data tends to

have distances greater than the linkage map that does consider misclassified markers. The method fails to work well when the observed number of any specific genotype occur in a much greater proportion than the other.

Table 3 and Table 4 contains the estimate of the recombination fraction for the rice and mice data respectively without considering misclassification ($\hat{r}_0$), the estimate of the recombination fraction after controlling for misclassification ($\hat{r}_1$) and the estimate of the coefficient of marker misclassification ($\hat{\lambda}_1$ and $\hat{\lambda}_2$).

Discussion

QTL mapping is essential in identifying the location of genes and in genetic study. Many statistical methods have been described for linkage analysis, but the analysis depends a lot on whether the data is affected by misclassification of certain phenotypes. Statistical methods described till now for linkage analysis can be used only for markers whose segregation follows the Mendelian ratio (1:1 for backcross or 1:2:1 for F_2). Care should be taken when using traditional methods for estimating the recombination fraction between markers since in practical molecular experiments, the influence of misclassified markers is not taken into consideration. The existence of misclassified markers has been ignored in the estimation of the recombination fraction between markers and construction of QTL maps and this leads to a biased estimate of the recombination fraction and results in discrepancies in QTL detection and mapping.

Genotyping errors occur when the genotype determined after molecular analysis does not correspond to the real genotype of the individual under consideration (Bonin et al. 2004). Misclassification of markers occur mainly due to variation in DNA sequence, low quantity or quality of DNA, biochemical artefacts and human error (Pompanon et al.). Various studies have been conducted evaluating the effect of genotyping errors of markers in the construction of genetic linkage maps (See Shields et al. 1991, Hackett and Broadfoot 2003, Buetow 1991 to name a few for more details). Hackett and Broadfoot note how even a low frequency of typing errors have a substantial impact on the order and length of a linkage map. To deal

with misclassification errors, Cartwright et al. extends the traditional likelihood model used for genetic mapping to include the possibility of genotyping errors where each individual marker is assigned an error rate. Lincoln and Lander overcame the disadvantages caused due to misclassified markers by incorporating the possibility of error into the usual likelihood model for linkage analysis.

Since marker segregation is disturbed by misclassified markers, these effects have to be incorporated into linkage analysis using modified methods. In this paper, we have developed a method to estimate the recombination fraction by including extra parameters to incorporate misclassification using the maximum likelihood approach and EM algorithm. Extra parameters are included in the method depending on whether misclassification occurs in just one gene or in both genes. For a One-Gene model, λ can be defined as the proportion of allele **A** that is accidentally classified as **a** while for a Two-Gene model we have two parameters λ_1 and λ_2 representing the misclassified proportions for markers **A** and **B**.

To compare the new method with the traditional one, a simulation study is conducted for both backcross and F_2 populations. To understand the effect of misclassification on the recombination fraction, the simulation study has also considered different degrees of marker misclassification. The simulated results show that the new method succeeds in accurately estimating the recombination fraction even when both markers are misclassified. The study also establishes the importance of a large sample size in increasing estimation precision.

The new method is also applied to the real backcross population rice dataset and the F_2 population mice data. In the rice dataset, one hundred and twenty-three DH plants derived from two inbred lines, semi-dwarf IR64 and tall Azucena were genotyped for a total of 175 polymorphic markers (Huang et al. 1997). In Huang's paper a two-point analysis was performed to estimate pair wise recombination fractions for these 175 markers. Based on a cluster analysis of the 175 by 175 matrix for recombination fractions, Huang et al. sorted these markers into 12 different groups each representing a rice chromosome. Here for illustration, we have only considered the first ten markers on rice chromosome 1. We have considered the Two-Gene model where both markers are misclassified with the assumption that misclassification arises from the ambiguity of individual alleles. In the mice data,

Large(LG/J) and Small(SM/J) inbred mouse strains were chosen of which ten SM/J males were mated with 10 LG/J females producing 41 F_1 progeny (Chevrud et al. 1996). These F_1 offsprings were then crossed with each other to produce 535 F_2 offsprings. Seventy five microsatellite loci were identified using the interval mapping methods described by Lander and Botstein (1989) and a genetic map was constructed. For purposes of this work, we have only considered a two-point analysis of 6 markers on chromosome 1.

To obtain the most probable order, we used the Sum of Adjacent Recombination Frequencies (SAR). SAR scores are calculated by summing recombination frequency estimates of neighboring intervals in a contiguous sequence of intervals (a locus order) where the most probable order will have the shortest map (lowest SAR score). For the rice data, the best order of these 10 markers was found to be RG472, RG246, K5, U10, RG532, W1, RG173, Amy1B, RZ276 and RG146. The genetic distances between all adjacent markers were estimated and the corresponding genetic distances were calculated using the Haldane map function (Haldane 1919). Considered to be the simplest second to the Morgan map function, it assumes that crossovers occur at random and independently of each other. The Haldane map function given by $r = \frac{1}{2}[1 - Prob(X = 0)] = \frac{1}{2}(1 - e^{-2d})$ had the disadvantage that it may not be accurate at small distances. But empirical observations show that the probability of having two crossovers occur in close proximity toe ach other is often less than predicted by the Haldane map funciton. A 10 by 10 matrix of estimates of r_0, r_1, λ_1 and λ_2 along with their standard errors are given in Table 3. The same technique is applied to analyze the F_2 population and the results are as in Table 4.

Figure 1 and Figure 2 gives the linkage map for the backcross and F_2 data respectively with the genetic distances and marker names at the left and right sides of the chromosome. We can see that the linkage map has changed considerably after incorporating misclassification. For the rice data, it can be observed that the order of 10 markers change along with the distance between them. The recombination fraction between markers considering misclassification is seen to be really close to 0.5 for the last three markers. This could be an indication to show that these markers may not belong to chromosome 1 and that it could be a part of another chromosome. Regarding the mice data, the best order of the 6 markers on chromosome 1 remained unchanged when compared with the model that did not consider

misclassification though the distances between the markers were seen to differ.

In the simulation, the traditional method seems to be overestimating the recombination fraction, and hence the linkage maps tend to be longer than the new method. But in the working example the method does not seem to work very well. This may be due to the imbalance in the observed genotypes or due to the small sample size. The method depends on the assumption that the misclassified proportions in a Two-Gene model are independent. The method fails to work for the rice data when the observed number of any specific genotype occur in a much greater proportion than the other.

This technique can be be applied to datasets after we first test if the markers follow mendelian segregation. Depending on the test, we can decide if a two-gene or a one-gene misclassification model should be applied.

References

Bailey NTJ. Introduction to the Mathematical Theory of Genetic Linkage. (1961); Clarendon Press, Oxford.

Cheverud JM, Routman EJ, Duarte FAM, van Swinderen B, Cothran K, Perel C. Quantitative trait loci for murine growth. Genetics. (1996);142:1305-1319.

Huang N, Parco A, Mew T, Magpantay G, McCouch S, Guiderdoni E, Xu J, Subudhi P, Angeles ER, Khush GS. RFLP mapping of isozymes, RAPD and QTLs for grain shape, brown planthopper resistance in a doubled haploid rice population. Molecular Breeding. (1997); 3:105-113.

Wu RL, Ma CX, Casella G. Statistical genetics of quantitative traits. (2007). Springer, Berlin Heidelberg New York.

Hackett CA, Broadfoot LB. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity.(2003); 90:3338.

Shields DC, Collins AK, Buetow H, Morton NE. Error filtration, interference, and the human linkage map. Proceedings of the National Academy of Sciences. (1991); 88:65016505.

Cartwright DA, Troggio M, Velasco R, Gutin A. Genetic Mapping in the Presence of Genotyping Errors. Genetics. (2007); 176:25212527.

Lincoln SE, Lander ES. Systematic detection of errors in genetic linkage data. Genomics. (1992); 14(3):604-610.

Buetow KH. Influence of aberrant observations on high resolution linkage analysis outcomes. The American Journal of Human Genetics. (1991); 49: 985994.

Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P. How to track and assess genotyping errors in population genetics studies. Molecular Ecology. (2004); 13:32613273.

Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: causes consequences, and solutions. Nature Reviews Genetics. (2005);6:847859.

Haldane JBS. The combination of linkage values and the calculation of distance between loci

of linked factors. Journal of Genetics. (1919); 8:299309.

Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. (1989); 121:185-199.

Wu RL, Ma CM, Wu SS, Zeng ZB. Linkage mapping of sex-specific differences. Genetical Research. (2002); 79:85-96.

Thompson EA. Information gain in joint linkage analysis. IMA Journal of Mathematics Applied in Medicine and Biology. (1984); 1:31-49.

Table 1: Linkage Analysis for Misclassified Markers (Backcross, 1000 simulations for each case)

$(\lambda_1, \lambda_2, r)$	$\hat{r} \pm SE$	$\hat{\lambda_1} \pm SE$	$\hat{\lambda_2} \pm SE$	$\hat{r_0} \pm SE$
n=80				
(0.0, 0.0, 0.05)	0.040 ± 0.104	0.000 ± 0.110	-0.000 ± 0.110	0.051 ± 0.026
(0.0, 0.0, 0.35)	0.345 ± 0.061	0.006 ± 0.112	0.005 ± 0.112	0.349 ± 0.053
(0.0, 0.1, 0.05)	0.034 ± 0.106	0.005 ± 0.111	0.106 ± 0.111	0.095 ± 0.033
(0.0, 0.1, 0.35)	0.351 ± 0.064	-0.002 ± 0.111	0.098 ± 0.105	0.367 ± 0.053
(0.1, 0.2, 0.05)	0.043 ± 0.101	0.097 ± 0.107	0.197 ± 0.107	0.166 ± 0.040
(0.1, 0.2, 0.35)	0.346 ± 0.085	0.103 ± 0.114	0.204 ± 0.112	0.381 ± 0.055
(0.2, 0.2, 0.05)	0.038 ± 0.116	0.204 ± 0.108	0.200 ± 0.109	0.192 ± 0.043
(0.2, 0.2, 0.35)	0.345 ± 0.091	0.200 ± 0.103	0.203 ± 0.110	0.383 ± 0.054
(0.2, 0.3, 0.05)	0.041 ± 0.127	0.202 ± 0.113	0.299 ± 0.106	0.219 ± 0.049
(0.2, 0.3, 0.35)	0.344 ± 0.097	0.200 ± 0.109	0.301 ± 0.102	0.384 ± 0.053
n=200				
(0.0, 0.0, 0.05)	0.046 ± 0.065	0.001 ± 0.069	-0.001 ± 0.071	0.050 ± 0.015
(0.0, 0.0, 0.35)	0.348 ± 0.038	0.002 ± 0.072	0.001 ± 0.068	0.350 ± 0.034
(0.0, 0.1, 0.05)	0.046 ± 0.064	0.000 ± 0.068	0.100 ± 0.069	0.095 ± 0.020
(0.0, 0.1, 0.35)	0.350 ± 0.043	-0.001 ± 0.072	0.101 ± 0.073	0.366 ± 0.035
(0.1, 0.2, 0.05)	0.047 ± 0.069	0.101 ± 0.070	0.198 ± 0.070	0.166 ± 0.027
(0.1, 0.2, 0.35)	0.349 ± 0.049	0.101 ± 0.068	0.199 ± 0.073	0.381 ± 0.034
(0.2, 0.2, 0.05)	0.048 ± 0.070	0.198 ± 0.069	0.196 ± 0.069	0.192 ± 0.027
(0.2, 0.2, 0.35)	0.349 ± 0.053	0.200 ± 0.068	0.201 ± 0.071	0.384 ± 0.033
(0.2, 0.3, 0.05)	0.045 ± 0.076	0.198 ± 0.071	0.299 ± 0.068	0.217 ± 0.029
(0.2, 0.3, 0.35)	0.348 ± 0.063	0.200 ± 0.071	0.302 ± 0.067	0.386 ± 0.035

Table 2: Linkage Analysis for Misclassified Markers (F2, 1000 simulations for each case)

$(\lambda_1, \lambda_2, r)$	$\hat{r} \pm SE$	$\hat{\lambda_1} \pm SE$	$\hat{\lambda_2} \pm SE$	$\hat{r_0} \pm SE$
n=80				
(0.0, 0.0, 0.05)	0.047 ± 0.075	-0.001 ± 0.080	-0.006 ± 0.132	0.052 ± 0.016
(0.0, 0.0, 0.35)	0.351 ± 0.050	-0.001 ± 0.078	0.001 ± 0.078	0.353 ± 0.047
(0.0, 0.1, 0.05)	0.042 ± 0.075	-0.044 ± 1.351	0.099 ± 0.089	0.096 ± 0.025
(0.0, 0.1, 0.35)	0.349 ± 0.058	0.001 ± 0.081	0.101 ± 0.078	0.367 ± 0.049
(0.1, 0.2, 0.05)	0.045 ± 0.082	0.094 ± 0.200	0.199 ± 0.079	0.168 ± 0.033
(0.1, 0.2, 0.35)	0.345 ± 0.069	0.101 ± 0.074	0.198 ± 0.077	0.376 ± 0.047
(0.2, 0.2, 0.05)	0.047 ± 0.087	0.199 ± 0.080	0.199 ± 0.078	0.193 ± 0.035
(0.2, 0.2, 0.35)	0.344 ± 0.076	0.199 ± 0.077	0.204 ± 0.074	0.372 ± 0.045
(0.2, 0.3, 0.05)	0.049 ± 0.091	0.197 ± 0.076	0.295 ± 0.076	0.217 ± 0.038
(0.2, 0.3, 0.35)	0.334 ± 0.083	0.202 ± 0.079	0.299 ± 0.071	0.366 ± 0.043
n=200				
(0.0, 0.0, 0.05)	0.047 ± 0.049	-0.001 ± 0.054	0.001 ± 0.054	0.050 ± 0.012
(0.0, 0.0, 0.35)	0.349 ± 0.035	-0.001 ± 0.050	-0.003 ± 0.052	0.349 ± 0.032
(0.0, 0.1, 0.05)	0.050 ± 0.045	-0.002 ± 0.049	0.096 ± 0.049	0.096 ± 0.016
(0.0, 0.1, 0.35)	0.351 ± 0.037	-0.002 ± 0.050	0.100 ± 0.051	0.367 ± 0.032
(0.1, 0.2, 0.05)	0.047 ± 0.052	0.101 ± 0.050	0.200 ± 0.049	0.167 ± 0.021
(0.1, 0.2, 0.35)	0.348 ± 0.047	0.101 ± 0.050	0.203 ± 0.050	0.377 ± 0.032
(0.2, 0.2, 0.05)	0.047 ± 0.054	0.200 ± 0.050	0.200 ± 0.049	0.190 ± 0.022
(0.2, 0.2, 0.35)	0.348 ± 0.051	0.201 ± 0.049	0.199 ± 0.045	0.374 ± 0.030
(0.2, 0.3, 0.05)	0.049 ± 0.056	0.200 ± 0.048	0.301 ± 0.049	0.218 ± 0.023
(0.2, 0.3, 0.35)	0.348 ± 0.057	0.197 ± 0.049	0.298 ± 0.049	0.373 ± 0.030

Marker	*RG472*	*RG246*	*K5*	*U10*	*RG532*	*W1*	*RG173*	*Amy1B*	*RZ276*	*RG146*
RG472		0.17±0.04 0.16±0.00 0.06±0.1 −0.03±0.1	0.28±0.04 0.3±0.00 0.06±0.1 −0.22±0.1	0.32±0.05 0.32±0.00 0.08±0.11 −0.18±0.11	0.29±0.04 0.31±0.00 0.07±0.1 −0.24±0.1	0.43±0.05 0.43±0.00 0.09±0.1 −0.43±0.09	0.43±0.05 0.44±0.00 0.06±0.1 −0.65±0.07	0.48±0.05 0.48±0.00 0.05±0.1 −0.8±0.06	0.5±0.05 0.49±0.00 0.06±0.1 −0.72±0.07	0.44±0.05 0.46±0.00 0.04±0.1 −0.69±0.07
RG246	0.17±0.04 0.16±0.00 −0.03±0.1 0.06±0.1		0.14±0.03 0.25±0.00 −0.09±0.1 −0.26±0.09	0.19±0.04 0.25±0.00 −0.02±0.11 −0.18±0.1	0.17±0.04 0.26±0.00 −0.07±0.1 −0.25±0.09	0.31±0.04 0.37±0.00 −0.03±0.09 −0.46±0.08	0.37±0.05 0.44±0.00 −0.07±0.1 −0.67±0.07	0.46±0.05 0.5±0.00 −0.07±0.09 −0.79±0.06	0.46±0.05 0.49±0.00 −0.05±0.09 −0.7±0.07	0.44±0.05 0.48±0.00 −0.06±0.1 −0.69±0.07
K5	0.28±0.04 0.3±0.00 −0.22±0.1 0.06±0.1	0.14±0.03 0.25±0.00 −0.26±0.09 −0.09±0.1		0.05±0.02 0.18±0.02 −0.15±0.11 −0.2±0.11	0.02±0.01 0.2±Inf −0.2±0.1 −0.24±0.1	0.18±0.04 0.35±0.00 −0.22±0.1 −0.47±0.09	0.26±0.04 0.43±0.00 −0.26±0.1 −0.71±0.07	0.38±0.05 0.49±0.00 −0.23±0.09 −0.79±0.06	0.38±0.05 0.48±0.00 −0.22±0.1 −0.75±0.06	0.35±0.05 0.47±0.00 −0.22±0.1 −0.72±0.07
U10	0.32±0.05 0.32±0.00 −0.18±0.11 0.08±0.11	0.19±0.04 0.25±0.00 −0.18±0.1 −0.02±0.11	0.05±0.02 0.18±0.02 −0.2±0.11 −0.15±0.11		0.06±0.03 0.16±0.01 −0.14±0.11 −0.12±0.11	0.13±0.04 0.29±0.00 −0.19±0.11 −0.35±0.1	0.24±0.05 0.4±0.00 −0.2±0.11 −0.62±0.08	0.39±0.05 0.47±0.00 −0.16±0.1 −0.76±0.07	0.4±0.05 0.48±0.00 −0.16±0.1 −0.69±0.08	0.36±0.05 0.46±0.00 −0.16±0.11 −0.66±0.08
RG532	0.29±0.04 0.31±0.00 −0.24±0.1 0.07±0.1	0.17±0.04 0.26±0.00 −0.25±0.09 −0.07±0.1	0.02±0.01 0.2±Inf −0.24±0.1 −0.2±0.1	0.06±0.03 0.16±0.01 −0.12±0.11 −0.14±0.11		0.15±0.03 0.33±0.00 −0.22±0.09 −0.42±0.09	0.26±0.04 0.42±0.00 −0.26±0.09 −0.66±0.07	0.35±0.04 0.47±0.00 −0.22±0.09 −0.75±0.06	0.36±0.05 0.47±0.00 −0.21±0.09 −0.71±0.07	0.34±0.05 0.46±0.00 −0.23±0.09 −0.68±0.07
W1	0.43±0.05 0.43±0.00 −0.43±0.09 0.09±0.1	0.31±0.04 0.37±0.00 −0.46±0.08 −0.03±0.09	0.18±0.04 0.35±0.00 −0.47±0.09 −0.22±0.1	0.13±0.04 0.29±0.00 −0.35±0.1 −0.19±0.11	0.15±0.03 0.33±0.00 −0.42±0.09 −0.22±0.09		0.11±0.03 0.41±0.00 −0.49±0.08 −0.68±0.07	0.24±0.04 0.46±0.00 −0.43±0.08 −0.77±0.06	0.26±0.04 0.46±0.00 −0.41±0.09 −0.69±0.07	0.24±0.04 0.45±0.00 −0.43±0.09 −0.69±0.07
RG173	0.43±0.05 0.44±0.00 −0.65±0.07 0.06±0.1	0.37±0.05 0.44±0.00 −0.67±0.07 −0.07±0.1	0.26±0.04 0.43±0.00 −0.71±0.07 −0.26±0.1	0.24±0.05 0.4±0.00 −0.62±0.08 −0.2±0.11	0.26±0.04 0.42±0.00 −0.66±0.07 −0.26±0.09	0.11±0.03 0.41±0.00 −0.68±0.07 −0.49±0.08		0.13±0.03 0.46±0.00 −0.67±0.07 −0.78±0.06	0.16±0.03 0.46±0.00 −0.67±0.07 −0.72±0.07	0.14±0.03 0.46±0.00 −0.65±0.07 −0.71±0.07
Amy1B	0.48±0.05 0.48±0.00 −0.8±0.06 0.05±0.1	0.46±0.05 0.5±0.00 −0.79±0.06 −0.07±0.09	0.38±0.05 0.49±0.00 −0.79±0.06 −0.23±0.09	0.39±0.05 0.47±0.00 −0.76±0.07 −0.16±0.1	0.35±0.04 0.47±0.00 −0.75±0.06 −0.22±0.09	0.24±0.04 0.46±0.00 −0.77±0.06 −0.43±0.08	0.13±0.03 0.46±0.00 −0.78±0.06 −0.67±0.07		0.04±0.02 0.43±Inf −0.77±0.06 −0.68±0.07	0.04±0.02 0.44±Inf −0.77±0.06 −0.7±0.07
RZ276	0.5±0.05 0.49±0.00 −0.72±0.07 0.06±0.1	0.46±0.05 0.49±0.00 −0.7±0.07 −0.05±0.09	0.38±0.05 0.48±0.00 −0.75±0.06 −0.22±0.1	0.4±0.05 0.48±0.00 −0.69±0.08 −0.16±0.1	0.36±0.05 0.47±0.00 −0.71±0.07 −0.21±0.09	0.26±0.04 0.46±0.00 −0.69±0.07 −0.41±0.09	0.16±0.03 0.46±0.00 −0.72±0.07 −0.67±0.07	0.04±0.02 0.43±Inf −0.68±0.07 −0.77±0.06		0.06±0.02 0.43±0.00 −0.7±0.07 −0.68±0.07
RG146	0.44±0.05 0.46±0.00 −0.69±0.07 0.04±0.1	0.44±0.05 0.48±0.00 −0.69±0.07 −0.06±0.1	0.35±0.05 0.47±0.00 −0.72±0.07 −0.22±0.1	0.36±0.05 0.46±0.00 −0.66±0.08 −0.16±0.11	0.34±0.05 0.46±0.00 −0.68±0.07 −0.23±0.09	0.24±0.04 0.45±0.00 −0.69±0.07 −0.43±0.09	0.14±0.03 0.46±0.00 −0.71±0.07 −0.65±0.07	0.04±0.02 0.44±Inf −0.7±0.07 −0.77±0.06	0.06±0.02 0.43±0.00 −0.68±0.07 −0.7±0.07	

Table 3: The MLEs of the recombination fraction between two misclassified markers in the backcross progeny of the rice data. Each cell contains the estimate of the recombination fraction without considering misclassification ($\hat{r}_0$), the estimate of the recombination fraction after controlling for misclassification ($\hat{r}_1$) and the estimate of the degree of marker misclassification ($\hat{\lambda}_1$ and $\hat{\lambda}_2$).

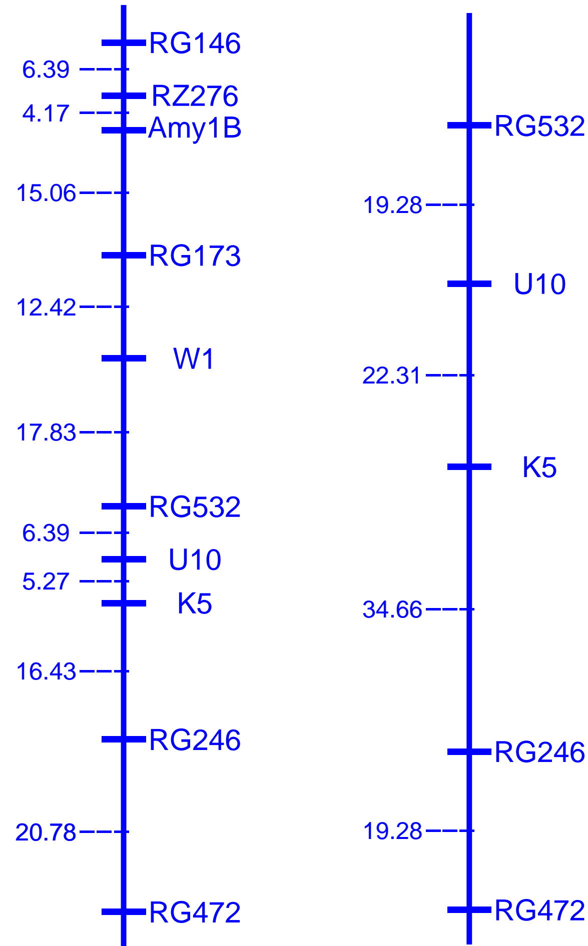
Marker	*D1Mit3*	*D1Mit20*	*D1Mit7*	*D1Mit11*	*D1Mit14*	*D1Mit17*
D1Mit3		0.0752±0.0088 0.0169±0.0311 0.0454±0.0321 0.0825±0.032	0.2982±0.0178 0.273±0.0226 0.0444±0.031 0.0792±0.031	0.3751±0.0211 0.3635±0.024 0.0571±0.0325 0.0444±0.0325	0.4587±0.0219 0.4567±0.0232 0.043±0.0312 0.0117±0.0312	0.5±0.0223 0.499±0.0235 0.0596±0.0315 −0.0159±0.0315
D1Mit20	0.0752±0.0088 0.0169±0.0312 0.0825±0.032 0.0454±0.0321		0.27±0.0172 0.2347±0.0238 0.0789±0.0317 0.0729±0.0317	0.3483±0.0206 0.3286±0.0248 0.0943±0.033 0.0351±0.0331	0.4438±0.0223 0.4388±0.0243 0.0761±0.032 0.0021±0.0321	0.5065±0.0226 0.5064±0.0246 0.0839±0.0319 −0.0117±0.032
D1Mit7	0.2982±0.0178 0.273±0.0228 0.0792±0.031 0.0444±0.031	0.27±0.0172 0.2347±0.0238 0.0729±0.0317 0.0789±0.0317		0.1018±0.0103 0.0478±0.0296 0.083±0.0321 0.0436±0.0322	0.2983±0.0178 0.2782±0.0225 0.077±0.0309 0.0154±0.031	0.4723±0.022 0.4687±0.0238 0.0801±0.0311 −0.0176±0.0312
D1Mit11	0.3751±0.0211 0.3635±0.0239 0.0444±0.0325 0.0571±0.0325	0.3483±0.0206 0.3286±0.0244 0.0351±0.0331 0.0943±0.033	0.1018±0.0103 0.0478±0.0295 0.0436±0.0322 0.083±0.0321		0.1982±0.0148 0.1744±0.0245 0.0527±0.0324 0.0232±0.0325	0.4358±0.0225 0.4332±0.0235 0.0403±0.0325 −0.0085±0.0325
D1Mit14	0.4587±0.0219 0.4567±0.0231 0.0117±0.0312 0.043±0.0312	0.4438±0.0223 0.4388±0.0241 0.0021±0.0321 0.0761±0.032	0.2983±0.0178 0.2782±0.0222 0.0154±0.031 0.077±0.0309	0.1982±0.0148 0.1744±0.0245 0.0232±0.0325 0.0527±0.0324		0.2857±0.0176 0.2858±0.0209 0.0139±0.0315 −0.0158±0.0315
D1Mit17	0.5±0.0223 0.499±0.0234 −0.0159±0.0315 0.0596±0.0315	0.5065±0.0226 0.5064±0.0243 −0.0117±0.032 0.0839±0.0319	0.4723±0.022 0.4687±0.0235 −0.0176±0.0312 0.0801±0.0311	0.4358±0.0225 0.4332±0.0234 −0.0085±0.0325 0.0403±0.0325	0.2857±0.0176 0.2858±0.0209 −0.0158±0.0315 0.0139±0.0315	

Table 4: The MLEs of the recombination fraction between two misclassified markers in the F_2 progeny of the mice data. Each cell contains the estimate of the recombination fraction without considering misclassification ($\hat{r}_0$), the estimate of the recombination fraction after controlling for misclassification ($\hat{r}_1$) and the estimate of the degree of marker misclassification ($\hat{\lambda}_1$ and $\hat{\lambda}_2$).

LEGENDS

FIGURE 1 - QTL map for Backcross Rice data before and after considering misclassification among markers.

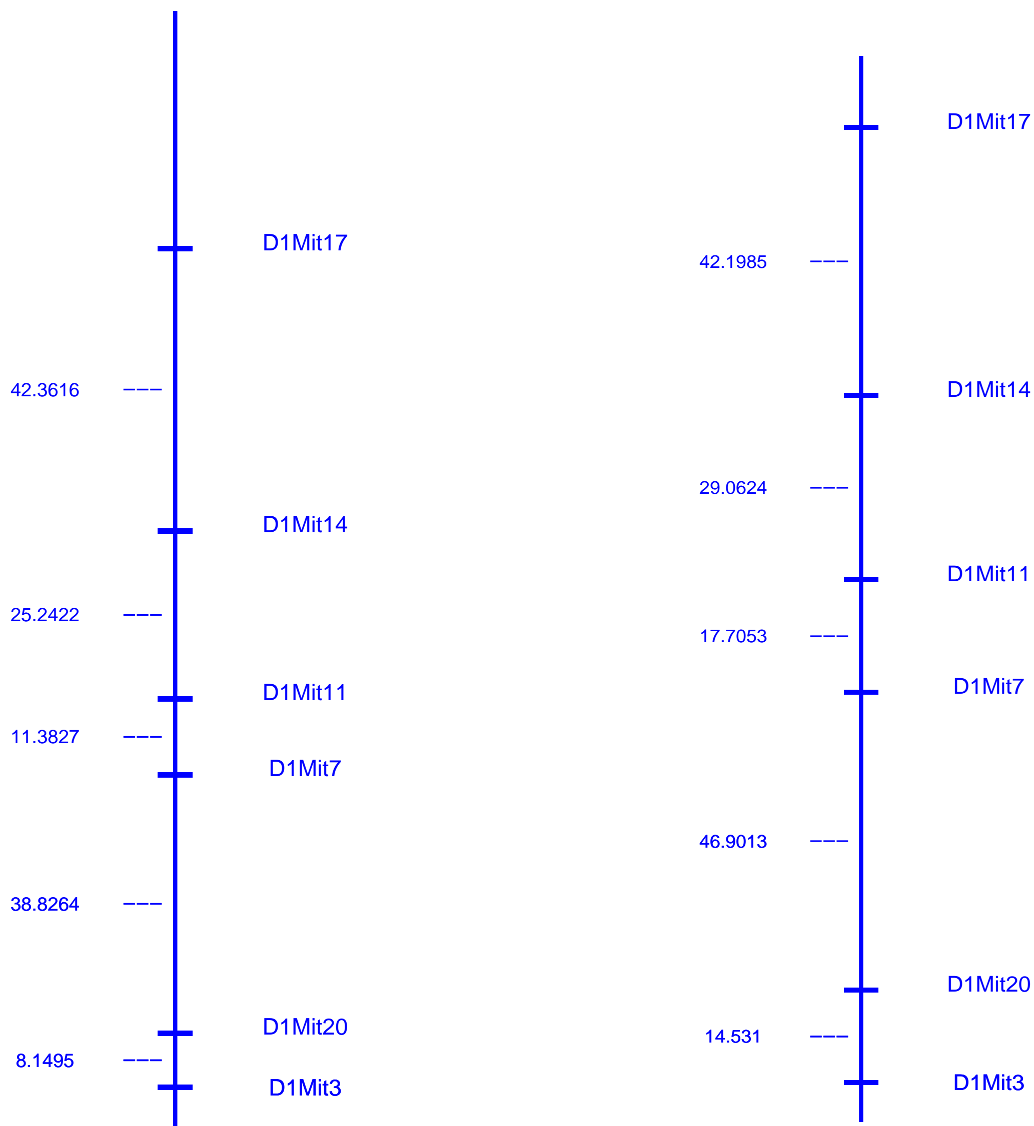
FIGURE 2 - QTL map for F2 mice data before and after considering misclassification among markers



(a) QTL map without considering misclassification

(b) QTL map after considering misclassification

Figure 1: QTL map for Chromosome 1 for Backcross Rice data



(a) QTL map without considering misclassification

(b) QTL map after considering misclassification

Figure 2: QTL map for Chromosome 1 for F2 mice data