Genet. Sel. Evol. 34 (2002) 353–369

© INRA, EDP Sciences, 2002
DOI: 10.1051/gse:2002012

Original article

A Bayesian approach for constructing genetic maps when markers are miscoded

Guilherme J.M. Rosa[a]*, Brian S. Yandell[b], Daniel Gianola[c]

[a] Department of Biostatistics, UNESP, Botucatu, SP, Brazil
[b] Departments of Statistics and of Horticulture, University of Winconsin, Madison, WI, USA
[c] Departments of Animal Science and of Biostatistics & Medical Informatics, University of Wisconsin, Madison, WI, USA

(Received 10 September 2001; accepted 8 February 2002)

Abstract – The advent of molecular markers has created opportunities for a better understanding of quantitative inheritance and for developing novel strategies for genetic improvement of agricultural species, using information on quantitative trait loci (QTL). A QTL analysis relies on accurate genetic marker maps. At present, most statistical methods used for map construction ignore the fact that molecular data may be read with error. Often, however, there is ambiguity about some marker genotypes. A Bayesian MCMC approach for inferences about a genetic marker map when random miscoding of genotypes occurs is presented, and simulated and real data sets are analyzed. The results suggest that unless there is strong reason to believe that genotypes are ascertained without error, the proposed approach provides more reliable inference on the genetic map.

genetic map construction / miscoded genotypes / Bayesian inference

1. INTRODUCTION

The advent of molecular markers has created opportunities for a better understanding of quantitative inheritance and for developing novel strategies for genetic improvement in agriculture. For example, the location and the effects of quantitative trait loci (QTL) can be inferred by combining information from marker genotypes and phenotypic scores of individuals in a population in linkage disequilibrium, such as in experiments with line crosses, *e.g.*, using backcross or F2 progenies. A QTL analysis relies on the availability of accurate estimates of the genetic marker map, which includes information

* Correspondence and reprints
E-mail: rosag@msu.edu
Current address: Departments of Animal Science and of Fisheries & Wildlife, Michigan State University, East Lansing, MI 48824, USA

on the order and on genetic distances between marker loci order. Genetic maps are inferred from recombination events between markers, which are genotyped for each individual. Several statistical methods have been suggested for map construction. Lathrop *et al.* [14], Ott [17] and Smith and Stephens [21] discussed maximum likelihood procedures for marker map inferences, and George *et al.* [9] presented a Bayesian approach for ordering gene markers. Jones [10] reviewed a variety of statistical methods for gene mapping. At present, most statistical methods used for map construction ignore the possibility that molecular (marker) data may be read with error. Often, however, there is ambiguity about genotypes and, if ignored, this can adversely affect inferences [3,15]. The problem of miscoded genotypes has received the attention of some investigators. Most of their research, however, has focused on error detection and data cleaning [4,11,15]. The objective of our work is to discuss possible biases in marker map estimates when miscoding of genotypes is ignored and to suggest a robust approach for more realistic inferences about marker positions and their distances. The approach simultaneously estimates the genotyping error rate and corrects for possible miscoded genotypes, while making inferences on the order and distances between genetic markers.

The plan of the paper is as follows. In Section 2, the problem of miscoding genotypes is discussed, as well as the systematic bias that this imposes on genetic map estimation. In Section 3, a Bayesian approach for inferences about a genetic map, when miscoding is ignored, is reviewed. In Section 4, the methodology is extended to handle situations with miscoded genotypes, when these occur at random. Simulated and real data are analyzed in Sections 5 and 6, respectively, and the results are discussed. Concluding remarks are presented in Section 7.

2. THE PROBLEM CAUSED BY MISCODED GENOTYPES

First consider the estimation of the genetic distance between two marker loci having a recombination rate θ. In simple situations, *e.g.*, with double haploid or backcross designs, each individual has one of two possible genotypes (say 0 or 1) at each marker locus. Inferences about genetic distance between loci are based on recombination events, which are observed by genotyping individuals. If marker genotypes could be read without error, the probability of observing a recombination event in a randomly drawn individual would be θ. However, it will be supposed that there is ambiguity in the assignment of genotypes to individuals. For example, a genotype 0 may be coded as 1 (or *vice-versa*), with probability π. Here, given the genotype for a specific marker and the probability of miscoding (π), the distribution of the observed genotypes can be

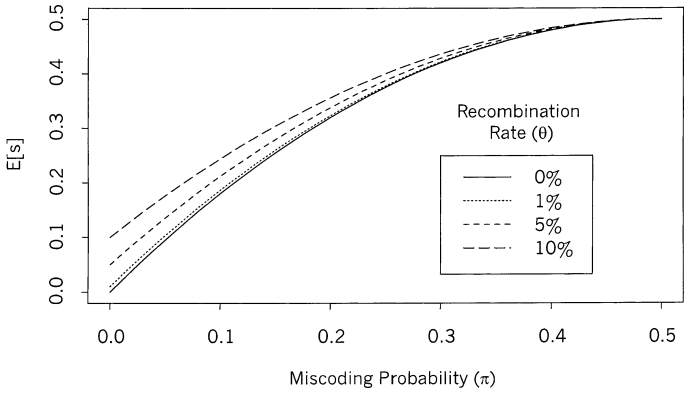

Figure 1. Expected recombination events observed on different values of miscoding probabilities (π), for some selected values of recombination rates (θ).

written as:

$$p[m_{ij}|g_{ij}, \pi] = \pi^{|m_{ij}-g_{ij}|}(1-\pi)^{1-|m_{ij}-g_{ij}|},$$

where m_{ij} and g_{ij} are the observed and true genotypes (m_{ij}, $g_{ij} = 0, 1$), respectively, for locus j ($j = 1, 2$) of individual i ($i = 1, 2, \ldots, n$).

If a "recombination event" between the loci is observed, this may be due to either a true genetic recombination between them, or to an artifact caused by miscoding. Hereinafter, a "recombination" observed by genotyping the markers will be denoted as the "apparent recombination", to distinguish between observed and "true" recombination events.

The probability of observing an apparent recombination between markers 1 and 2 for individual i can be written as:

$$\begin{aligned}\Pr(s_i = 1) &= \Pr[r_i = 1]\,(\Pr[\text{no miscod.}] + \Pr[\text{double miscod.}]) \\ &\quad + \Pr[r_i = 0]\,\Pr[\text{one miscod.}] \\ &= \theta\left[\pi^2 + (1-\pi)^2\right] + 2(1-\theta)\pi(1-\pi) \\ &= \theta + 2\pi(1-\pi)(1-2\theta), \end{aligned} \tag{1}$$

where $s_i = |m_{i1} - m_{i2}|$ and $r_i = |g_{i1} - g_{i2}|$ stand for apparent and real recombination events, respectively; and $\Pr[r_i = k] = \theta^k(1-\theta)^{1-k}$, with $k = 0, 1$.

It is easy to realize, therefore, that recombination rates estimated from recombinations observed by genotyping the marker loci, ignoring the possibility of miscoding, would be biased upwards whenever the markers are linked ($\theta < 0.5$) and $\pi > 0$. Figure 1 shows the expected apparent recombination rates as function of π, for some selected recombination rate values. It seems that the smaller the genetic recombination rate, the worse the relative bias produced by miscoded genotypes.

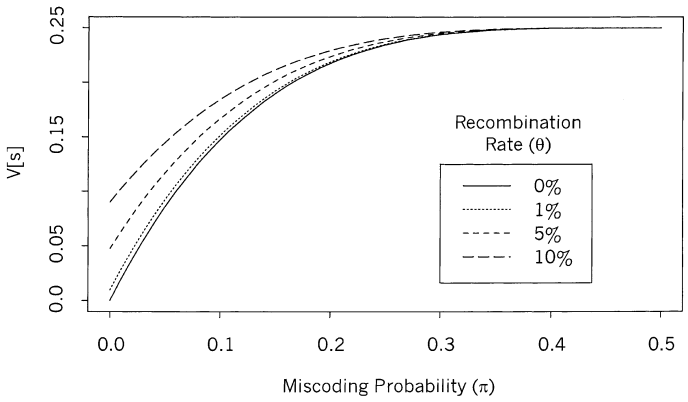

Figure 2. Variance of recombination events observed on different values of miscoding probabilities (π), for some selected values of recombination rates (θ).

The variance of the apparent recombination event is equal to:

$$\begin{aligned} Var[s_i] &= \Pr[s_i = 1]\,(1 - \Pr[s_i = 1]) \\ &= [\theta + 2\pi(1-\pi)(1-2\theta)][1 - \theta - 2\pi(1-\pi)(1-2\theta)] \\ &= \theta(1-\theta) + 2\pi(1 - 3\pi + 4\pi^2 - 2\pi^3)(1-2\theta)^2. \end{aligned} \tag{2}$$

Thus, the variance of apparent recombination events is larger than the variance of the real recombination events whenever the markers are linked ($\theta < 0.5$) and $\pi > 0$. Figure 2 shows the variance of the apparent recombination events as a function of π, for some different values of recombination rates.

In view of the possibility of miscoding for each marker genotype (*i.e.* ambiguity about their genotypes), standard methods commonly used for genetic map inferences overestimate the recombination rate between loci (or, in other words, underestimate genetic linkage), and underestimate its precision [15]. For example, the maximum likelihood estimator of the recombination rate between the loci (if the possibility of miscoding is ignored) is:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} |m_{i1} - m_{i2}|,$$

with expectation and variance given by (1) and (2), respectively.

In more general situations, we have more than just two marker loci, and the goal is to construct the genetic map, *i.e.*, to order these marker loci and to estimate the genetic distances between them. Again, all inferences are based on recombination events observed (apparent recombinations) between the marker loci. The problem of ignoring miscoding may lead to even worse difficulties, *e.g.*, to the mistaken ordering of the loci, specially with dense maps.

3. BAYESIAN APPROACH FOR GENETIC MAP CONSTRUCTION

First, we will review a Bayesian approach for map construction when miscoding is not taken into account [9]. Consider the genotype of m markers for the individual i as $\boldsymbol{g}_i = (g_{i1}, g_{i2}, \ldots, g_{im})$. In a backcross design, for example, $g_{ij} = 0$ if the individual i is homozygous for the locus j, and 1 otherwise. The sampling model of $\boldsymbol{g}_i$, assuming the Haldane map function, is given by:

$$p(\boldsymbol{g}_i|\lambda, \boldsymbol{\theta}) \propto \prod_{j=1}^{m-1} \theta_j^{|g_{ij}-g_{i,j+1}|}(1-\theta_j)^{1-|g_{ij}-g_{i,j+1}|}, \tag{3}$$

where λ is the order of the genetic marker loci and θ_j is the recombination rate between the loci j and $j+1$. Considering a sample of n independent individuals, the likelihood of λ and θ is given by:

$$\begin{aligned} L(\lambda, \boldsymbol{\theta}|\boldsymbol{G}) = p(\boldsymbol{G}|\lambda, \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(\boldsymbol{g}_i|\lambda, \boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{m-1} \theta_j^{|g_{ij}-g_{i,j+1}|}(1-\theta_j)^{1-|g_{ij}-g_{i,j+1}|}, \end{aligned} \tag{4}$$

where $\boldsymbol{G}$ is the $(n \times m)$ matrix of marker genotypes, with each row representing one individual, and each column related to one marker locus.

In a Bayesian context, rather than maximizing the likelihood, it is modified by a prior and integrated to produce inference summaries for the unknown components in the model. The prior can be chosen based on earlier studies or information from the literature. Here, we use a prior expressed as:

$$p(\lambda, \boldsymbol{\theta}|\boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta})p(\lambda|\boldsymbol{\tau}), \tag{5}$$

where $p(\lambda|\boldsymbol{\tau})$ is a probability distribution over the $m!/2$ different orders for the m markers, $\boldsymbol{\tau}$ is a set of prior probabilities of each order, and $p(\boldsymbol{\theta}|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{m-1} p(\theta_j|\lambda, \alpha_j, \beta_j)$, where $\theta_j|\lambda, \alpha_j, \beta_j \sim Beta(\alpha_j, \beta_j)$ is the recombination rate between genetic markers j and $j+1$. A special case of these prior distributions would be uniform across different gene orders, and *Uniform* (0, 0.5) distributions for each θ_j.

The Bayes theorem combines the information from the data and the prior knowledge to produce a posterior distribution over all unknown quantities. In this case, the posterior density of λ and $\boldsymbol{\theta}$ is given by:

$$p(\lambda, \boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto p(\boldsymbol{G}|\lambda, \boldsymbol{\theta})p(\boldsymbol{\theta}|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta})p(\lambda|\boldsymbol{\tau}). \tag{6}$$

Distribution (6) is intractable analytically but MCMC methods such as the Gibbs sampler and the Metropolis-Hastings algorithm [7,8] can be used to draw samples, from which features of marginal distributions of interest can be inferred.

3.1. Fully conditional posterior distributions

The Gibbs sampler draws samples iteratively from conditional posterior distributions deriving from (6). The fully conditional posterior distribution of each recombination rate θ_j is:

$$p(\theta_j|\lambda, \boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1}\prod_{i=1}^{n}\theta_j^{|g_{ij}-g_{i,j+1}|}(1-\theta_j)^{1-|g_{ij}-g_{i,j+1}|}$$
$$\propto \theta_j^{q_j+\alpha_j-1}(1-\theta_j)^{n-q_j+\beta_j-1}, \tag{7}$$

where $q_j = \sum_{i=1}^{n}|g_{i,j} - g_{i,j+1}|$ is the number of recombination events between the loci j and $j+1$. This is the kernel of a Beta distribution with parameters $(q_j + \alpha_j)$ and $(n - q_j + \beta_j)$.

The updating for the gene order λ involves moves between a set of models, because for distinct ordering, the recombination rates have different meanings. George *et al.* [9] discuss a reversible jump algorithm, for which recombination rates are converted into map distances, and reverted to new recombination rates after shifting a randomly selected marker around a pivot marker.

Here, another Metropolis-Hastings [12] scheme is presented for the MCMC updating of λ and $\boldsymbol{\theta}$, simultaneously. A new gene ordering is proposed according to a candidate generator density $q(.)$, and new recombination rates are simulated for this new order, using (7). The Markov chain moves from the current state $T = (\lambda, \boldsymbol{\theta})$ to $T^* = (\lambda^*, \boldsymbol{\theta}^*)$ with probability:

$$\pi(T^*, T) = \min\left[1, \frac{p(\lambda^*, \boldsymbol{\theta}^*|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\lambda, \boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\frac{q(\lambda, \lambda^*)}{q(\lambda^*, \lambda)}\right], \tag{8}$$

where $p(\lambda, \boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the joint conditional posterior distribution of the gene ordering λ and recombination rates $\boldsymbol{\theta}$, given by:

$$p(\lambda, \boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto p(\lambda|\boldsymbol{\tau})\prod_{j=1}^{m-1}\theta_j^{q_j+\alpha_j-1}(1-\theta_j)^{n-q_j+\beta_j-1}.$$

Under these circumstances, the choice of $q(.)$ is extremely important for an efficient implementation of the MCMC, especially in situations with a large number of marker loci. A bad choice of $q(.)$ would generate a large number of unlikely orders, or even generate inconsistent orders, in relation to the data set. In order to have a better implementation and mixing of the MCMC, some alternatives for the generation of candidate orders for the Metropolis-Hastings step are described in the Appendix.

3.2. Missing data

In practice, some marker genotypes are missing. The missing data can be handled by the MCMC approach, with an additional step for updating each missing genotype based on this fully conditional density. For instance, suppose g_{ij} is missing, the genotype for the j-th marker of the individual i. Its fully conditional distribution is Bernoulli with probability $p_{ij} = \Pr(g_{ij} = 1|\boldsymbol{G}_{-ij})$ given by:

$$p_{ij} = \frac{p(g_{ij} = 1|\boldsymbol{\theta}, \boldsymbol{G}_{-ij}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\sum_k p(g_{ij} = k|\boldsymbol{\theta}, \boldsymbol{G}_{-ij}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})},$$

where $\boldsymbol{G}_{-ij}$ refers to all elements in $\boldsymbol{G}$ but g_{ij}, and $k = 0, 1$. Under the Haldane independence assumption, $p(g_{ij} = k|\boldsymbol{\theta}, \boldsymbol{G}_{-ij}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ depends just on the recombination rates between the locus j and its flanking neighbors, as well as on the genotypes of these neighbor loci, so it can be written as $p(g_{ij} = k|\theta_{j-1}, \theta_j, g_{i,j-1}, g_{i,j+1})$.

4. THE PROBABILITY OF MISCODING GENOTYPES

At present, the methods commonly used for map construction ignore the possibility that molecular (marker) data may be read with error, or the error rate has a fixed and known value, as in Lincoln and Lander [15]. Often, however, there is ambiguity about the genotypes. To address these situations, we introduced a new parameter into the model, the probability π of miscoding a genotype. Now we consider that the matrix $\boldsymbol{G}$ of genotypes is unknown, and that we observe a matrix $\boldsymbol{M}$ of genotypes, possibly with some miscoding. The probability of observing a genotype m_{ij}, *i.e.* the genotype of locus j for individual i, given that the actual genotype is g_{ij} , may be expressed as:

$$\Pr(m_{ij} = k_1|g_{ij} = k_2) = \pi^{|k_1-k_2|}(1-\pi)^{1-|k_1-k_2|},$$

where k_1 and k_2 assume values equal to 0 or 1.

Assuming independence between miscodings in different loci and individuals, and considering that the miscoding rate is constant over the genome, the probability of observing a matrix $\boldsymbol{M}$ of genotypes, given the matrix $\boldsymbol{G}$ of actual genotypes, can be expressed as:

$$p(\boldsymbol{M}|\boldsymbol{G}) = \pi^t(1-\pi)^{nm-t}, \tag{9}$$

where n is the number of individuals, m is the number of marker loci, and $t = \sum_{i=1}^{n}\sum_{j=1}^{m}|m_{ij} - g_{ij}|$ is the number of miscoding genotypes in the data set. Note that under these circumstances, $\boldsymbol{M}$ is the observed data, and $\boldsymbol{G}$ is now

an auxiliary and non-observed matrix. The joint posterior distribution of all unknowns in the model is written now as the product of (9) by (4), (5) and the prior distribution of π, which gives:

$$p(\boldsymbol{G}, \lambda, \boldsymbol{\theta}, \pi|\boldsymbol{M}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b)$$
$$\propto \pi^t(1-\pi)^{nm-t} \prod_{i=1}^{n}\prod_{j=1}^{m-1} \theta_j^{|g_{ij}-g_{i,j+1}|}(1-\theta_j)^{1-|g_{ij}-g_{i,j+1}|}$$
$$\times\, p(\boldsymbol{\theta}|\lambda, \boldsymbol{\alpha}, \boldsymbol{\beta})p(\lambda|\boldsymbol{\tau})p(\pi|a, b). \tag{10}$$

Assuming a uniform prior probability distribution for λ; $Beta(\alpha_j, \beta_j)$ as prior for each θ_j; and $Beta(a, b)$ as the prior distribution for π, the expression (10) becomes:

$$p(\boldsymbol{G}, \lambda, \boldsymbol{\theta}, \pi|\boldsymbol{M}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b)$$
$$\propto \pi^{a+t-1}(1-\pi)^{b+nm-t-1} \prod_{j=1}^{m-1} \theta_j^{\alpha_j+q_j-1}(1-\theta_j)^{\beta_j+n-q_j-1}$$

where $q_j = \sum_{i=1}^{n} |g_{ij} - g_{i,j+1}|$, as already defined, is the number of recombination events between the loci j and $j+1$. Note that the dependence of this distribution on λ is rendered implicit by the definition of θ_j as the recombination rate between the ordered loci j and $j+1$.

4.1. Fully conditional posterior distributions

The fully conditional posterior distributions of λ and of each θ_j have the same forms as discussed before. In the case of $\boldsymbol{G}$, its conditional distribution is:

$$p(\boldsymbol{G}|\boldsymbol{M}, \lambda, \boldsymbol{\theta}, \pi, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b)$$
$$\propto \pi^{a+t-1}(1-\pi)^{b+nm-t-1} \prod_{j=1}^{m-1} \theta_j^{\alpha_j+q_j-1}(1-\theta_j)^{\beta_j+n-q_j-1}.$$

Given the independence between the recombination events in different intervals (by the Haldane map function), each element in $\boldsymbol{G}$ can be updated independently. If $j = 1$, *i.e.* g_{ij} refers to genotypes at one end of the linkage group, its fully conditional posterior distribution can be written as:

$$p(g_{i1}|\boldsymbol{G}_{-i1}, \boldsymbol{M}, \lambda, \boldsymbol{\theta}, \pi, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b)$$
$$\propto \pi^{|g_{i1}-m_{i1}|}(1-\pi)^{1-|g_{i1}-m_{i1}|}\theta_1^{|g_{i1}-g_{i2}|}(1-\theta_1)^{1-|g_{i1}-g_{i2}|},$$

where $\boldsymbol{G}_{-i1}$ represents all the elements in $\boldsymbol{G}$ but g_{i1}, and similarly for g_{im}.

For genotypes at interior markers in the linkage group, the fully conditional posterior distribution becomes:

$$p(g_{ij}|\boldsymbol{G}_{-ij}, \boldsymbol{M}, \lambda, \boldsymbol{\theta}, \pi, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \propto \pi^{|g_{ij}-m_{ij}|}(1-\pi)^{1-|g_{ij}-m_{ij}|} \times \theta_{j-1}^{|g_{ij}-g_{i,j-1}|}(1-\theta_{j-1})^{1-|g_{ij}-g_{i,j-1}|}\theta_j^{|g_{ij}-g_{i,j+1}|}(1-\theta_j)^{1-|g_{ij}-g_{i,j+1}|},$$

for $j = 2, 3, \ldots, m-1$. The conditional distribution of the probability of miscoding π is given by:

$$p(\pi|\boldsymbol{M}, \boldsymbol{G}, \lambda, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta}, a, b) \propto \pi^{a+t-1}(1-\pi)^{b+nm-t-1},$$

which is the kernel of a Beta distribution with parameters $(a+t)$ and $(b+nm-t)$.

5. SIMULATION STUDY

5.1. Example 1

Three data sets were simulated to examine the ability of the model discussed in Section 4 to correctly estimate genetic distances and the probability of miscoding. Each simulation considered 300 individuals with genotypes for 5 loci, denoted as *ABCDE*. The recombination rates between consecutive loci were assumed to be $\theta_{AB} = 0.09$, $\theta_{BC} = 0.11$, $\theta_{CD} = 0.05$ and $\theta_{DE} = 0.14$. The data sets were generated considering $\pi = 0$, 0.02 and 0.04, and 3% of missing data for each.

These data sets were analyzed using models with and without the miscoding parameter (π). An equal probability distribution was adopted as prior for the different loci orders. For each recombination rate, a *Uniform* (0, 0.5) process was considered as prior distribution. Computations were performed using the IML procedure of SAS [19]. Graphical inspection and the Raftery and Lewis diagnostic [18] for the Gibbs output using CODA [1] were used for assessing convergence to the equilibrium distribution, the joint posterior. A burn-in period of 1 000 iterations was adopted, followed by 60 000 iterations with thinning intervals of 20, based on a lag-correlation study. Hence, 3 000 samples were retained for the post-Gibbs analysis.

For all data sets, the gene order was estimated perfectly by both models, with 100% of the MCMC iterations sampling the order *ABCDE*. It seems that, up to certain levels, inferences about gene ordering is robust to miscoding genotypes, if these occur at random. As discussed earlier (Sect. 2), the effect of miscoding is larger for smaller genetic distances between loci, such as in fine mapping studies. In these cases, the miscoding may lead to ordering estimated with some positions switched for tightly linked markers, as discussed in the next example.

Table I. True parameter values and posterior means and standard deviations (in parenthesis) of the recombination rates considering the data set without miscoding genotypes and the two models, with and without the miscoding parameter.

	Recombination rates					
Model	θ_{AB}	θ_{BC}	θ_{CD}	θ_{DE}	π	$p(\boldsymbol{y}\|\text{Model})$
W/o miscoding	0.0929	0.1060	0.0528	0.1214	–	2.61×10^{-162}
	(0.0172)	(0.0181)	(0.0132)	(0.0186)		
With miscoding	0.0892	0.1028	.0511	0.1193	0.0027	5.36×10^{-161}
	(0.0178)	(0.0179)	(0.0131)	(0.0194)	(0.0024)	
Parameter values	0.09	0.11	0.05	0.14	0	–

Table I shows the posterior mean and standard deviation for each recombination rate, for the data set without miscoding. The estimates obtained by each model do not present any relevant difference, so it seems that the introduction of the extra parameter (π) into the model, in situations where there is no miscoding, does not affect the estimated genetic map. In this example, the estimate for π was very close to zero, denoting the ability of the model to recognize situations without miscoding. However, because $\pi = 0$ relies on the boundary of the parameter space of π, to test for the absence of miscoding for a particular data set, another approach should be employed, such as comparing both models (with and without miscoding) using some criteria, *e.g.*, the Bayes factor or the likelihood ratio test.

The Bayes factors may be computed by taking ratios between estimates of the marginal densities of the data (after integrating out all parameters). If models are taken as equally probable, *a priori*, then the Bayes factor gives the ratio between the posterior probabilities of the corresponding models. Here, the marginal densities were estimated by calculating harmonic means of likelihoods evaluated at the posterior draws of the Gibbs output [16], and these are presented in Table I. The Bayes factor (in favor of the model without the miscoding parameter) of 20.5 does not denote important differences between both models for modeling this data set.

The results obtained by both models for the data set with 2% miscoding ($\pi = 0.02$) are presented in Table II. As expected, the model that ignores the miscoding problem had estimates biased upwards. When the probability of miscoding was introduced into the model, there was improvement on the estimates. In addition, the probability of miscoding was adequately estimated. For the robust model, all the parameter values were inside a credible set of 0.95 of probability. The Bayes factor of 2.01×10^6, in favor of the model with the miscoding parameter, denotes its greater plausibility, when compared to the model ignoring miscoding genotypes.

Table II. True parameter values and posterior means and standard deviations (in parenthesis) of the recombination rates considering the data set with 2% of miscoding genotypes and the two models, with and without the miscoding parameter.

	Recombination rates					
Model	θ_{AB}	θ_{BC}	θ_{CD}	θ_{DE}	π	$p(\boldsymbol{y}\|\text{Model})$
W/o miscoding	0.0982	0.1361	0.0934	0.1950	–	4.70×10^{-200}
	(0.0178)	(0.0195)	(0.0172)	(0.0226)		
With miscoding	0.0739	0.1096	0.0561	0.1624	0.0223	9.45×10^{-194}
	(0.0192)	(0.0200)	(0.0175)	(0.0251)	(0.0067)	
Parameter values	0.09	0.11	0.05	0.14	0.02	–

Table III. True parameter values and posterior means and standard deviations (in parenthesis) of the recombination rates considering the data set with 4% of miscoding genotypes and the two models, with and without the miscoding parameter.

	Recombination rates					
Model	θ_{AB}	θ_{BC}	θ_{CD}	θ_{DE}	π	$p(\boldsymbol{y}\|\text{Model})$
W/o miscoding	0.1681	0.1640	0.1413	0.1758	–	1.93×10^{-235}
	(0.0223)	(0.0215)	(0.0204)	(0.0217)		
With miscoding	0.1327	0.1208	0.0828	0.1307	0.0374	3.37×10^{-213}
	(0.0252)	(0.0228)	(0.0204)	(0.0251)	(0.0087)	
Parameter values	0.09	0.11	0.05	0.14	0.04	–

Similar results were found for the data set with 4% miscoding (Tab. III). The Bayes factor, in this case, was of 1.75×10^{22} in favor of the model with the miscoding parameter.

5.2. Example 2

Thirty data sets were simulated, where half had no miscoding and half had 5% miscoding. Here, our main interest was to examine the performance of the models (with and without miscoding) to correctly estimate the gene order with relatively small data sets, both under situations without miscoding and with high levels of miscoding genotypes. Each data set had 100 individuals with genotypes for five markers; no missing data were considered in this study. The recombination rates between consecutive loci were: $\theta_{AB} = 0.05$, $\theta_{BC} = 0.18$, $\theta_{CD} = 0.02$ and $\theta_{DE} = 0.07$. Prior distributions and computations were similar to those described for the previous simulation study.

For the 15 data sets without miscoding, both models (with and without miscoding) yielded the highest posterior probability for the correct order *ABCDE*.

The Bayes factor was favorable to the model without miscoding for 8 of the data sets (and favorable to the model with the miscoding parameter for the remaining data sets), but with no expressive or relevant values (ranging from 1.2 to 57.9). For the datasets with 5% miscoding, the Bayes factor was always favorable to the model considering miscoding genotypes, with values that ranged from 2.42×10^2 to 8.31×10^{10}.

The model ignoring the miscoding gave the highest posterior probability for a gene ordering other than *ABCDE* in 8 of the 15 data sets. In the case of the model with the miscoding parameter, just four data sets had the highest posterior probability for a wrong order. If credibility sets with minimum probability of 0.90 are considered, three of the data sets had the correct gene ordering outside of the set. For the model with the miscoding parameter, just one data set presented a probability set that did not contain the correct order.

The results suggest that the model ignoring the miscoding overstates the precision in relation to the gene ordering, sometimes concentrating posterior probability on the wrong (set of) order(s).

6. ANALYSIS OF EXPERIMENTAL DATA

The data set refers to the RFLP study with *Brassica napus* using F1-derived double haploid lines. Materials and methods related to the DNA extraction and a preliminary linkage map construction (using maximum likelihood approach) are presented by Ferreira *et al.* [6]. These data, combined with phenotypic information (flowering time under one of the three vernalization treatments considered in that study), were also analyzed by Ferreira *et al.* [5] and by Satagopan *et al.* [20] for the study on quantitative trait loci.

Here, we focus on the estimation of the probability of miscoding, and also on robust construction of the linkage map for a set of marker loci. To illustrate the methods, we consider the data for 105 progeny and 10 maker loci related to the linkage group 9, for which 9% of the genotypes were missing. For simplicity, the marker loci are denoted here by letters (from A through J), according to the order that was estimated by Ferreira *et al.* [5].

Prior distributions and computations were similar to those describe in Section 5, for the simulation studies. In this case, a longer burn-in period of 2 000 iterations was adopted, followed by 100 000 iterations with thinning intervals of 20. Hence, 5 000 samples were used for the post-Gibbs analysis.

The miscoding rate for this data set was estimated at a level of approximately 1.5%, with a 95% probability set [0.0055; 0.0266]. The model with the miscoding parameter was much more plausible than the one without it, with the Bayes Factor of 4.91×10^6. Posterior probabilities for different gene ordering, estimated by both models (with and without the miscoding parameter) are presented in Table IV. The most probable order for both models was the

Table IV. Posterior probabilities of different ordering of the markers in the *Brassica* data by using the two models, with and without miscoding.

	Models	
Order	W/o miscoding	With miscoding
ABCDEFGHIJ	0.7516	0.6824
BACDEFGHIJ	0.1818	0.1658
ABDCEFGHIJ	0.0412	0.0632
ABCEDFGHIJ	0.0128	0.0404
BACEDFGHIJ	0.0036	0.0094
Others [(1)]	0.0090	0.0388

[(1)] Loci orders with posterior probabilities smaller than 0.0050.

Table V. Posterior means and standard deviations (in parenthesis) of the recombination rates and probability of miscoding in the *Brassica* data by using the two models, with and without the miscoding parameter.

	Models	
Parameter	W/o miscoding	With miscoding
θ_{AB}	0.0923 (0.0291)	0.0693 (0.0308)
θ_{BC}	0.1135 (0.0318)	0.0961 (0.0304)
θ_{CD}	0.0719 (0.0255)	0.0390 (0.0224)
θ_{DE}	0.0730 (0.0259)	0.0392 (0.0223)
θ_{EF}	0.0899 (0.0289)	0.0853 (0.0298)
θ_{FG}	0.1096 (0.0319)	0.0773 (0.0308)
θ_{GH}	0.1084 (0.0320)	0.0749 (0.0309)
θ_{HI}	0.0684 (0.0271)	0.0681 (0.0287)
θ_{IJ}	0.1353 (0.0358)	0.1320 (0.0379)
π	–	0.0151 (0.0054)
$p(\boldsymbol{y}\|\text{Model})$	4.81×10^{-137}	2.36×10^{-130}

sequence *ABCDEFGHIJ*, with approximate posterior probabilities of 0.68 and 0.75, respectively for models with and without the miscoding parameter. Some uncertainty on the order arose with the position of the two first markers (*A* and *B*), but very high posterior probabilities were found for the sequence *CDEFGHIJ* of the other eight loci (respectively 0.85 and 0.93 for the models with and without the miscoding parameter).

The recombination rates, as expected, presented higher estimates by the model ignoring the possibility of miscoding (Tab. V). Figure 3 shows the estimated map from these two models, using the inverse of the Haldane map function to the recombination rates drawn at each Gibbs iteration.

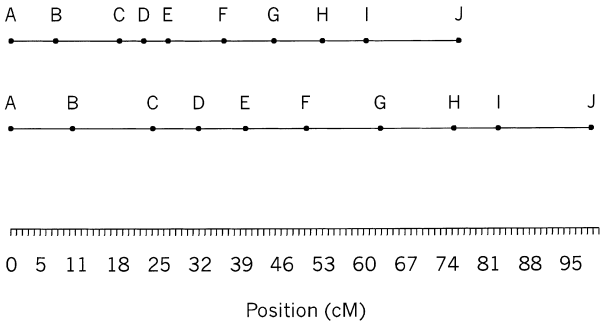

Figure 3. Estimated genetic map of the markers of the *Brassica* data, by using the robust model (first map) and the model ignoring miscoding (second map).

7. CONCLUDING REMARKS

The model discussed in this paper provides an appealing robust alternative for genetic map construction in the presence of non-systematic miscoding genotypes. The MCMC implementation of the Bayesian analysis is straightforward, with just some caution to be addressed in relation to the Metropolis-Hastings step for updating the gene ordering. This approach provides more reliable estimates for subsequent studies that use information on genetic maps, such as quantitative trait loci (QTL) search and marker assisted selection.

High values of miscoding probability estimates, however, should raise concern about the molecular data, and a revaluation of the marker genotypes may be a good approach. In situations with relatively large rates of miscoding, the high frequency of apparent recombinations may not be recognized as the reflect of miscoding genotypes, but due to bigger values of real genetic recombinations. For these cases, a multilocus feasible map function, which assumes interdependence between different marker intervals [17], could be a better alternative to the Haldane map function.

This paper can be extended in various ways to analyze genetic data originated from different designs (*e.g.* F2 progenies, granddaughter design, etc.). Furthermore, the idea of considering the possibility of miscoding genotypes may be used for QTL analysis as well. The methodology for robust estimation under miscoding genotypes can be adapted to handle multiallelic loci situations. For example, consider that g_{ij} can assume one of t genotypes, denoted as $1, 2, \ldots, t$. In these cases, the probability of observing a genotype m_{ij} equal to r ($r = 1, 2, \ldots, t$) would be:

$$\begin{aligned}\Pr(m_{ij} = r) &= \Pr(g_{ij} = r)\Pr(m_{ij} = r|g_{ij} = r) + \Pr(g_{ij} \neq r)\Pr(m_{ij} = r|g_{ij} \neq r) \\ &= \Pr(g_{ij} = r)\Pr(m_{ij} = r|g_{ij} = r) + \sum_{s=1;s\neq r}^{t} \pi_{rs}\Pr(g_{ij} = s),\end{aligned}$$

where $\pi_{rs} = \Pr(m_{ij} = r|g_{ij} = s)$ represents the miscoding probability of observing a genotype as r, when the actual genotype is s. The miscoding probabilities π_{rs} could be considered, for example, proportional to the distance of each allele in the gel, which reflects the size of each allele (in base pairs). In this case, the miscoding probabilities would be written as:

$$\pi_{rs} = \Pr(m_{ij} = r|g_{ij} = s) = \phi(\text{distance between alleles } r \text{ and } s).$$

The results found in this work suggest that, unless there is strong reason to believe in the absence of ambiguity about genotypes, it may be safer to use the robust model, which would provide a more reliable estimate of the genetic map.

Towards the completion of this research, we became aware of related studies by Keller [13] under the supervision of G.A. Churchill. Their work confirms the utility of our approach to assess miscoding and improve the estimation of map distances. Some of their methods appear as part of the R/QTL software module of Broman [2].

ACKNOWLEDGEMENTS

This work was supported by the Wisconsin Agriculture Experimental Station, by research grant NRICGP/USDA 99-35205-8162.

REFERENCES

[1] Best N., Cowles M.K., Vines K., CODA Manual, Version 0.30. Technical Report, Cambridge, UK MRC Biostatistics Unit, 1995.

[2] Broman K.W., R/QTL Software Module, Version 0.80–3, 2001. (http://biosun01.biostat.jhsph.edu/~kbroman/).

[3] Brzustowicz L.M., Merette C., Xie X., Townsend L., Gilliam T.C., Ott L., Molecular and statistical approaches to the detection and correction of errors in genotype database, Am. J. Hum. Genet. 53 (1993) 1137–1145.

[4] Douglas J.A., Boehnke M., Lange K., A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data, Am. J. Hum. Genet. 66 (2000) 1287–1297.

[5] Ferreira M.E., Satagopan J., Yandell B.S., Williams P.H., Osborn T.C., Mapping loci controlling vernalization requirement and flowering time in *Brassica napus*, Theor. Appl. Genet. 90 (1995) 727–732.

[6] Ferreira M.E., Williams P.H., Osborn T.C., RFLP mapping of *Brassica napus* using double haploid lines, Theor. Appl. Genet. 89 (1995) 615–621.

[7] Gelfand A.E., Smith A.F.M., Sampling based approaches to calculating marginal densities, J. Am. Stat. Assoc. 85 (1990) 398–409.

[8] Geman S., Geman D., Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.
[9] George A.W., Mengersen K.L., Davis G.P., A Bayesian approach to ordering gene markers, Biometrics 55 (1999) 419–429.
[10] Jones H.B., A review of statistical methods for genome mapping, Int. Stat. Rev. 68 (2000) 5–21.
[11] Haines J.L., Chromlook – an interative program for error-detection and mapping in reference linkage data, Genomics 14 (1992) 517–519.
[12] Hastings W.K., Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[13] Keller A.E., Estimation of genetic map distances, detection of genotype errors, and imputation of missing genotypes via Gibbs sampling, M.S. Thesis, Cornell University, 1999.
[14] Lathrop G.M., Lalouel J.M., Julier C., Ott J., Strategies for multilocus linkage analysis in humans, Proc. Nat. Acad. Sci., USA 81 (1984) 3443–3446.
[15] Lincoln S.E., Lander E.S., Systematic detection of errors in genetic linkage data, Genomics 14 (1992) 604–610.
[16] Newton M.A., Raftery A.E., Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), J.R. Stat. Soc. Series B 56 (1984) 3–48.
[17] Ott J., Analysis of Human Genetic Linkage, John Hopkins University Press, Baltimore, 1991.
[18] Raftery A.E., Lewis S.M., How many iterations in the Gibbs sampler?, in: Bernardo J.M., Berger J.O., David A.P., Smith A.F.M. (Eds.), Bayesian Statistics 4, Oxford Univ. Press, 1992, pp. 763–774.
[19] SAS® Institute Inc., SAS/IML Software: Usage and Reference, Version 6. 1st edn., SAS® Institute Inc., Cary, NC, 1989.
[20] Satagopan J.M., Yandell B.S., Newton M.A., Osborn T.C., A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo, Genetics 144 (1996) 805–816.
[21] Smith C.A.B, Stephens D.A., Estimating multipoint recombination fractions, Ann. Hum. Genet. 59 (1995) 307–321.

APPENDIX: PROPOSALS FOR λ

A simplified version of the Metropolis-Hastings step for drawing from the conditional distribution of λ and $\boldsymbol{\theta}$ can be described as follows:

1. Draw λ^* with probability $q(\lambda^*|m) = 2/m!$, from the $m!/2$ different orders;
2. Draw each θ_j from (7);
3. Move from the current state $\boldsymbol{T} = (\lambda, \boldsymbol{\theta})$ to $\boldsymbol{T}^* = (\lambda^*, \boldsymbol{\theta}^*)$ with probability $\pi(T^*, T)$, or stay with $\boldsymbol{T}$ otherwise. In this case, the Metropolis ratio given in (8) is simplified as:

$$\pi(T^*, T) = \min\left[1, \frac{p(\lambda^*, \boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\lambda, \boldsymbol{\theta}|\boldsymbol{G}, \boldsymbol{\tau}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\right].$$

The equally probable process for $q(.)$, as described above, is not an adequate choice for generating candidates for λ. As discussed in Section 3, this process would generate a large number of unlikely (or inconsistent) orders, that would increase the rejection rate of the Metropolis-Hastings step. In order to have a better implementation and mixing of the MCMC, some alternatives for the generation of candidate orders are described below.

A) Switching adjacent loci

In this case, a locus is chosen at random and its position is interchanged with its neighbor, for example, on the right. If the last gene position is chosen, then the two ends of the linkage group are interchanged. This alternative can be schematized as follows:

1. Draw k from $p(k|m) = 1/m$, where $k = 1, 2, \ldots, m$;
2. Define λ^* as equal to λ, except that loci k and $k+1$ are switched, if $k = 1, 2, \ldots, m-1$. If $k = m$, the loci 1 and m have their positions interchanged.

B) Switching two non adjacent loci

This alternative is, in some sense, a generalization of the previous one. Here, two loci are chosen at random and their positions are interchanged. It can be schematized following:

1. Draw k_1 from $p(k_1|m) = 1/m$, where $k_1 = 1, 2, \ldots, m$;
2. Draw k_2 from $p(k_2|m) = 1/(m-1)$, where $k_2 \neq k_1 = 1, 2, \ldots, m$;
3. Define λ^* as equal to λ, except that loci k_1 and k_2 are switched.

C) Rotation of random length segments

In this more general case, a random set of neighbor loci is chosen, and the new order is derived from the old one, with the rotation on this set of genes. It is described as follows:

1. Draw k_1 from $p(k_1|m) = 1/m$, where $k_1 = 1, 2, \ldots, m$;
2. Draw k_2 from $p(k_2|m) = 1/(m-1)$, where $k_2 \neq k_1 = 1, 2, \ldots, m$;
3. Suppose that $k_1 < k_2$ and write λ as:

$$\lambda = (\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(k_1-1)}, \lambda_{(k_1)}, \lambda_{(k_1+1)}, \ldots \\ \ldots, \lambda_{(k_2-1)}, \lambda_{(k_2)}, \lambda_{(k_2+1)}, \ldots, \lambda_{(m)}),$$

where $\lambda_{(j)}$ is the marker at the position j in the linkage group. The new order λ^* is defined as:

$$\lambda^* = (\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(k_1-1)}, \lambda_{(k_2)}, \lambda_{(k_2-1)}, \ldots \\ \ldots, \lambda_{(k_1+1)}, \lambda_{(k_1)}, \lambda_{(k_2+1)}, \ldots, \lambda_{(m)}).$$