Theor Appl Genet (2001) 102:1113–1122

J. Jansen · A.G. de Jong · J.W. van Ooijen

Constructing dense genetic linkage maps

Received: 1 June 2000 / Accepted: 21 September 2000

Abstract This paper describes a novel combination of techniques for the construction of dense genetic linkage maps. The construction of such maps is hampered by the occurrence of even small proportions of typing errors. Simulated annealing is used to obtain the best map according to the optimality criterion: the likelihood or the total number of recombination events. Spatial sampling of markers is used to obtain a framework map. The construction of a framework map is essential if the steps used for simulated annealing are required to be simple. For missing-data imputation the Gibbs sampler is used. Map construction using simulated annealing and missing-data imputation are used in an iterative way. In order to obtain some measure of precision of the genetic linkage map obtained, the Metropolis-Hastings algorithm is used to obtain posterior intervals for the positions of markers. The process of map construction is embedded in a framework of pre-mapping and post-mapping diagnostics. The techniques described are illustrated using a practical application.

Keywords Genetic linkage maps · Simulated annealing · Gibbs sampling · Metropolis-Hastings algorithm · Mapping diagnostics · Maximum likelihood · Missing-value imputation

Introduction

Genetic linkage maps play a prominent role in many areas of genetics, e.g. QTL analysis and map-based cloning of genes. They are also important for marker-assisted breeding. In order to obtain dense genetic linkage maps, increasing numbers of molecular genetic markers are typed in many plant species. Currently, the numbers may vary between 1,000 and 2,000 for a whole genome, but larger numbers are expected from new technologies. To cope with these large numbers, powerful methods for map construction are required. Increasing numbers of markers inevitably lead to problems. For dense maps, typing errors lead to unexpected crossovers resulting in large increases of map lengths (Lincoln and Lander 1992). In practice, also non-random typing errors, other laboratory errors and missing observations occur with unforeseen consequences. In these situations the JoinMap algorithm (Stam et al. 1993), based on sequential addition of markers in a systematic way in combination with a computationally demanding optimality criterion, requires large amounts of computer time.

Besides new algorithms, tools are required for establishing the quality of the data and the maps produced. In any mapping program the algorithms for constructing maps should be accompanied by such tools.

In this paper we introduce an algorithm for constructing dense genetic linkage maps based on simulated annealing (Kirkpatrick et al. 1983). In the past simulated annealing has been used extensively for the construction of genetic linkage maps (Lander and Green 1987; Weeks and Lange 1987; Falk 1992). The power of simulated annealing lies in its simplicity. The algorithm considered in this paper consists of simple steps involving only one marker. Furthermore, all steps are computationally equivalent. A major problem with the use of simulated annealing in its simple form is that one linkage group may be divided into two or more groups. This problem is encountered often in situations with dense maps with markers containing typing errors. The problem can be overcome by considering more complex steps, e.g. moving blocks of markers or changing the direction of blocks of markers. The problem can also be avoided by applying a two-stage approach still based on simple steps. In the first stage, a framework map is constructed using a small set of selected markers. This map is then used as the basis for the construction of a map for all markers.

Communicated by J.W. Snape

J. Jansen (✉) · A.G. de Jong · J.W. van Ooijen
Wageningen UR Centre for Biometry,
Plant Research International, P.O. Box 16, 6700 AA Wageningen,
The Netherlands
e-mail: J.Jansen@plant.wag-ur.nl
Fax: +31 317 418094

Missing observations occur in all mapping data. Missing data are dealt with in two stages. Once a map has been obtained further information is available for missing-data imputation. Missing-data imputation is carried out using the EM algorithm, which requires the calculation of conditional expectations. As missing data may occur in a multitude of patterns, the Gibbs sampler provides a simple Monte Carlo implementation of the EM algorithm. An iterative procedure involving the alternate use of simulated annealing and Gibbs sampling is employed. Finally, the Metropolis-Hastings algorithm is used to obtain a set of maps that are not optimal but plausible with regard to the variation in the data. Graphical representations of these plausible maps can gauge the resolution and quality of the map produced. They can also be used to identify whether optimization was successful. A worked application involving a doubled-haploid population of *Brassica napus* is employed to illustrate the methods developed in this paper and to develop a system for the analysis of large sets of molecular marker data in the context of genetic linkage maps.

Materials and method

Data and model

In this paper we will consider populations of a diploid species, where for all segregating loci only two genotypes occur. One example is the backcross population [(*ab*=*aa*×*bb*)×*aa*] with possible genotypes *aa* and *ab*. The methods developed in this paper can also be used for (doubled) haploid populations, and as an approximation to families of recombinant inbred lines (RILs). The data consist of observations on N individuals that are typed for M binary molecular genetic markers. For individual n the marker data will be represented by (y_{n1}, y_{n2} ... y_{nM}), where y_{nm}=0 if for individual n the genotype of marker m is *aa*, and y_{nm}=1 if for individual n the genotype of marker m is *ab*. In practice, the data may contain many missing observations. First, we will assume that all data are present. Typing errors also occur.

For the purpose of this paper we consider pairs of markers. For a pair of markers a backcross may be written as $a_1b_1/a_2b_2\times a_1a_1/a_2a_2$. The genotypes that may be produced in this cross are a_1a_1/a_2a_2, a_1b_1/a_2b_2, a_1b_1/a_2a_2 and a_1a_1/a_2b_2 with probabilities (1-r)/2, (1-r)/2, r/2 and r/2, respectively, where r denotes the recombination frequency. If r=0, only genotypes a_1a_1/a_2a_2, b_1b_1/b_2b_2 occur, each with a probability of $\frac{1}{2}$. In this case the two markers lie on the same position. If r=$\frac{1}{2}$, all four genotypes occur with an equal probability of $\frac{1}{4}$. In this case the two markers are a long-distance apart, possibly on different chromosomes.

The maximum-likelihood estimate $\hat{r}$ of the recombination frequency r is given by:

$$\hat{r} = \frac{R}{N},$$

in which R is the number of individuals for which $y_{n1}\neq y_{n2}$ (n=1, 2 ... N).

In general, a genetic linkage map based on M molecular genetic markers, consists of G straight lines. Each of these straight lines corresponds with a linkage group. Ideally, the number of linkage groups is equal to the number of chromosomes, but in practice it may be smaller or larger. For a particular linkage group the line starts with a marker on the first position at one end of a linkage group (or chromosome) and ends with a marker on the other end.

Table 1 Genotype probabilities and observations in the case of three loci

Genotype		Probability	Observation
$a_1a_1/a_2a_2/a_3a_3$	$a_1b_1/a_2b_2/a_3b_3$	$(1-r_{12})(1-r_{23})$	Y_{00}
$a_1a_1/a_2a_2/a_3b_3$	$a_1b_1/a_2b_2/a_3a_3$	$(1-r_{12})r_{23}$	Y_{01}
$a_1a_1/a_2b_2/a_3b_3$	$a_1b_1/a_2a_2/a_3a_3$	$r_{12}(1-r_{23})$	Y_{10}
$a_1a_1/a_2b_2/a_3a_3$	$a_1b_1/a_2a_2/a_3b_3$	$r_{12}r_{23}$	Y_{11}

The number of markers in linkage group c will be denoted by m_g (g=1, 2 ... G), so that $\sum_{g=1}^{G} m_g = M$.

The mapping problem consists of two separate problems:

(1) the division of M genetic markers into G linkage groups, where the number G has to be determined from the data;
(2) the ordering and positioning of G groups of m_1, m_2 ... m_G markers, respectively.

The total number of potential marker orders is equal to $\sum_{g=1}^{G}\frac{1}{2}m_g!$, which may be incredibly large. In the case where we have 1,000 markers divided into ten linkage groups of 100 markers each, the total number of marker orders is equal to 4.7×10^{158}. In this paper we will only consider the second problem: the ordering and positioning of markers within linkage groups.

Suppose that for a particular linkage group of size three we have markers assigned to positions 1, 2 and 3. In this case the backcross population consists of eight possible genotypes. If we assume that the occurrence of recombination in one interval is independent of the occurrence of recombinations in other intervals, the probabilities of eight genotypes are given in Table 1.

In Table 1, r_{12} and r_{23} are the recombination probabilities between the markers on positions 1 and 2, and 2 and 3, respectively. Furthermore, Y_{11} denotes the number of individuals with a recombination between positions 1 and 2, and between positions 2 and 3; Y_{10} denotes the number of individuals with recombination between positions 1 and 2 only; Y_{01} denotes the number of individuals with recombination between positions 2 and 3 only; Y_{00} denotes the number of individuals with neither recombination between positions 1 and 2 nor between positions 2 and 3.

The likelihood corresponding with this marker order is

$$L = \frac{N!}{Y_{00}!Y_{01}!Y_{10}!Y_{11}!} r_{12}{}^{R_{12}}\left(1-r_{12}\right)^{N-R_{12}} r_{23}{}^{R_{23}}\left(1-r_{23}\right)^{N-R_{23}},$$

where R_{12}=Y_{10}+Y_{11} and R_{23}=Y_{01}+Y_{11} are the number of individuals with a recombination between positions 1 and 2, and 2 and 3, respectively. In the case the three markers are unlinked the likelihood reads

$$L_0 = \frac{N!}{Y_{00}!Y_{01}!Y_{10}!Y_{11}!}\frac{1}{2}^{2N}.$$

The likelihood ratio Λ=L/L_0 will be used as one criterion for finding the mapping order. In general, the maximum value of the likelihood ratio is equal to

$$\Lambda = 2^{(P-1)N} N^{-N} \sum_{p=1}^{P-1}\left(R_{p,p+1}{}^{R_{p,p+1}}\left(N-R_{p,p+1}\right)^{N-R_{p,p+1}}\right),$$

in which P is the number of positions, i.e. markers, in the linkage group. In order to evaluate different mapping orders we only need to know the numbers of recombinant individuals for all pairs of markers.

For dense maps the maximum value of the likelihood ratio may be approximated by

$$\Lambda \approx 2^{(P-1)N} e^{-\sum_{p=1}^{P-1} R_{p,p+1}},$$

i.e. for dense maps finding the map with the largest likelihood ratio is approximately equivalent to finding the map with the smallest total number of recombination events. The total number of recombination events will be used as an alternative to the likelihood ratio. It is computationally much simpler, as it does not require extensive calculation of logarithms.

Searching the best map using simulated annealing

The algorithm that will be described starts from a map on which the markers are placed in random order. From this starting map, a new map is obtained by taking the marker from position s (=1, 2 ... P) and putting it between the markers on positions t (=0, 1, 2 ... P; $t \neq s-1$, s) and t+1. The likelihood ratio corresponding with the step described above is given by

$$LR = \frac{\Lambda_{[s-1][s+1]}}{\Lambda_{[s-1][s]}\Lambda_{[s][s+1]}} \times \frac{\Lambda_{[t][s]}\Lambda_{[s][t+1]}}{\Lambda_{[t][t+1]}},$$

in which $\Lambda_{[u][v]} = R_{[u][v]}{}^{R_{[u][v]}}\left(N - R_{[u][v]}\right)^{N-R_{[u][v]}}$, and u and v are two positions on the chromosomes.

The new map will be accepted if

$$e^{\ell/\gamma} > U,$$

where ℓ=ln(LR), γ is the acceptance control parameter (γ>0) and U is drawn at random from the standard uniform distribution. As a consequence if $\ell \geq 0$ the new map is always accepted, but if $\ell < 0$ the new map is accepted with a probability which decreases with the value of ℓ. The parameter γ is used to control the acceptance of steps with negative values of ℓ. To be able to get around local optima, steps with negative values of ℓ are sometimes allowed.

The acceptance control parameter γ is set to a fixed value in a series of S successive steps, which is called a chain. The annealing algorithm runs through C+1 of such chains for which $\gamma_0 \geq \gamma_1 \geq ... \geq \gamma_C$. If γ becomes close to zero only steps with positive values of ℓ will be accepted.

Chain 0 is used to determine the value of γ_1. The probability that a step is accepted while $\ell < 0$ is equal to $p=e^{\ell/\gamma}$, which itself is a random variable. The expectation π of p is equal to

$$\pi = \mathrm{E}[e^{\ell/\gamma} \mid \ell < 0] \approx e^{\mathrm{E}[\ell|\ell<0]/\gamma},$$

where E[$\ell|\ell<0$] is the expectation of negative outcomes of ℓ. It follows that

$$\gamma \approx \mathrm{E}[\ell|\ell<0]/\ln(\pi).$$

In chain 0, every step is accepted by setting γ_0 to a large value. The parameter γ_1 is set equal to $\ell_{-}/\ln(\pi_0)$, where ℓ_{-} is the average of the negative values of ℓ encountered in chain 0 and π_0 is the initial acceptance probability of upward steps. A simple formula to govern the decrease of γ as the optimisation progresses is given by

$$\gamma_{c+1} = \frac{\gamma_c}{1+\alpha}, (c = 1, 2 \ldots C),$$

in which α, the cooling control parameter, is a positive constant. Optimisation is halted, if in a number of successive chains no improvement of the likelihood criterion is found. One of the failures of map construction using simulated annealing is that it may split up linkage groups into two or more ”unlinked” fragments. Inspection of the resulting linkage map usually indicates that these fragments are unlinked because they have been put in the wrong direction. This is due to the fact that only simple steps are used in the annealing procedure.

To overcome this fragmenting problem optimisation is carried out in two-stages. In the first stage a map is obtained for a small number of markers. This map will be used as the framework map on which all markers will be placed. This is the second stage of the optimisation procedure.

Selection of markers for the first stage of the optimisation procedure is based on the assumption that markers of one linkage group lie approximately on a straight line. Selection of markers is carried out in the following way. The parameter r_0, which will be called the selection radius, plays a crucial role in the selection process. We take one marker at random (marker S_1). All markers with a recombination frequency smaller than r_0 with marker S_1 are deleted from the whole set of markers and put in the set D_1. From the remaining markers we again take one marker at random (marker S_2). All markers with a recombination frequency smaller than the value r_0 with marker S_2 are deleted from the remaining set of markers and put in the set D_2. The sampling process is continued until no markers are left.

Subsequently, a map is obtained for the selected markers S_1, S_2 $\cdots$ using simulated annealing. In the case where the whole set contains 100 markers and the set of selected markers contains ten markers the size of the optimization problem is reduced considerably: the number of possible marker orders is reduced from 4.7×10^{157} to 1.8×10^6. Furthermore, typing errors will have less effect on the resulting marker order. The procedure is continued by putting the markers in set D_1 close to marker S_1, the markers in set D_2 close to marker S_2, and so on. The linkage map obtained in this way provides a good starting point for further optimisation. Simulated annealing is used to find the optimal ordering for the whole set of markers. It may be expected that only local exchanges of markers will improve the optimisation criterion. To avoid excursions over long distances the acceptance control parameter γ is given a small value.

Missing data imputation

The above derivation of the likelihood ratio and the subsequent development of simulated annealing, requires that all marker observations are present. In practice this is hardly ever the case. Instead of the likelihood ratio we use the following approximation

$$\tilde{\Delta} = 2^{(P-1)N} N^{-N} \prod_{p=1}^{P-1} \mathrm{E}\left[R_{p,p+1}\right]^{\mathrm{E}\left[R_{p,p+1}\right]} \left(1 - \mathrm{E}\left[R_{p,p+1}\right]\right)^{N-\mathrm{E}\left[R_{p,p+1}\right]},$$

where E[$R_{p,p+1}$] denotes the conditional expectation of $R_{p,p+1}$ given the map and all available data.

Initially, no map is available. Suppose for the markers at positions p and p+1, pairs of observations are present on N_a individuals and R_a individuals show a recombination. In that case we take

$$\mathrm{E} = \left[R_{p,p+1}\right] = R_a + \frac{1}{2}(N - N_a).$$

In the above formula missing data are replaced with 0 or 1 with equal probability. The recombination coefficient is estimated by

$$\hat{r}_{p,p+1} = \frac{\mathrm{E}\left[R_{p,p+1}\right]}{N}.$$

A direct consequence of the above approach is that if markers have many missing observations their recombination frequencies with other markers will be considerably inflated.

If for an individual the observation of the marker on position p is missing, it is not possible to say whether a recombination occurred between the marker on positions p–1 and p, and the markers on position p and p+1. However, if the recombination coefficients $r_{p-1,p}$ and $r_{p,p+1}$ are known, it is possible to calculate conditional probabilities for the occurrence of recombination.

Given that the markers on positions p–1 and p+1 are not recombinant, then either the markers on positions p–1 and p are not recombinant and also the markers on positions p and p+1 are not recombinant, or the markers on positions p–1 and p are recombinant and also the markers on positions p and p+1 are recombinant.

The conditional probabilities are given by

$$q_{00} = \frac{\left(1 - r_{p-1,p}\right)\left(1 - r_{p,p+1}\right)}{\left(1 - r_{p-1,p}\right)\left(1 - r_{p,p+1}\right) + r_{p-1,p}\, r_{p,p+1}}$$

and q_{11}=1–q_{00}, respectively. If the recombination coefficients $r_{p-1,p}$ and $r_{p,p+1}$ are small, say 0.01, q_{11} is practically zero.

Given that the markers on positions p–1 and p+1 are recombinant, then either the markers on positions p–1 and p are recombinant but the markers on positions p and p+1 are not recombinant, or the markers on positions p–1 and p are not recombinant but the markers on positions p and p+1 are recombinant.

The corresponding conditional probabilities are given by

$$q_{10} = \frac{r_{p-1,p}\left(1 - r_{p,p+1}\right)}{r_{p-1,p}\left(1 - r_{p,p+1}\right) + \left(1 - r_{p-1,p}\right) r_{p,p+1}}$$

and q_{01}=1–q_{10}, respectively. The above approach is the simplest case of three-point linkage analysis, which is often used in more complex situations, e.g. full-sib families of outcrossing species (Ridout et al. 1998).

In practice, it may not be possible to carry out missing-data imputation in the above described way, because observations on neighbouring markers may also be missing. The above conditional probabilities can be used to calculate the conditional expectation of all pairwise numbers of recombinants using the Gibbs sampler. We may start from any ‘completed’ data matrix. Such a matrix can be obtained initially by replacing missing observations with zero or one with a probability of $\frac{1}{2}$. Then we take the missing observations in random order, one at a time, and replace its current value with a zero or a one using the above conditional probabilities. For example, if the observations on the flanking markers are both equal to one, the value of the missing observations becomes one with a probability of q_{00}, and zero with a probability of q_{11}.

Using the above approach we may generate a long sequence of ‘completed’ data matrices with corresponding matrices of the number of recombinants. Successive matrices of recombination coefficients are highly dependent. Consequently, a long sequence of matrices have to be generated from which a sample is taken to calculate a new average matrix of recombination coefficients. This ‘averaging’ coincides with the M-step of an EM algorithm. The new average can be used to refine the previously obtained map and to calculate new conditional probabilities (E-step). This process can be repeated until no further improvement of the likelihood is obtained.

The above approach using the Gibbs sampler is the simplest case of multi-point linkage analysis. It should be noted that estimates of pairwise recombination fractions are only consistent if the mapping order used is the true order.

Combining simulated annealing and Gibbs sampling

Map construction using simulated annealing and missing-data imputation with the Gibbs sampler are combined in the following way.

(1) Obtain an initial matrix of the expected number of recombinants;
(2) select markers using the spatial sampling approach;
(3) construct a framework map for the selected markers using the initial matrix of the expected numbers of recombinants by employing simulated annealing;
(4) construct an initial map for all markers using the framework map for selected markers;
(5) construct a map for all markers using the initial matrix of expected numbers of recombinants by simulated annealing;
(6) obtain with the Gibbs sampler a new matrix of expected numbers based on information from the map obtained in (5);
(7) construct a new map for all markers using the improved matrix of expected numbers of recombinants obtained in (6) using simulated annealing.

Steps (6) and (7) may be continued until no further progress is obtained. In our experience three or four iterations are adequate.

Posterior intervals

Mapping experiments usually result in one map, which is presented without any measure of uncertainty. The Metropolis-Hastings algorithm may be used to obtain a set of maps which, although not optimal, are plausible with regard to the variation in the data. Our approach starts from the best map obtained and the associated matrix of expected numbers of recombinations between markers. The set of plausible maps is obtained by using the simple steps as defined for simulated annealing, i.e. replacing one marker at a time. A new map is accepted if

$e^{\ell}>U$,

in which ℓ=ln(LR) as defined before, and U is drawn at random from the standard uniform distribution. This process allows limited excursions from the optimal map. Successive maps obtained in this way are very similar. Therefore, after a burn-in period, the series of maps is sampled at regular intervals to reduce the dependence between maps. In our current applications, we use a burn-in period of 1,000,000 steps. Then 1,000 samples are drawn at intervals of 1,000 maps.

The sampled maps are used to construct posterior intervals, i.e. for each marker we record the position numbers encountered in the sample of plausible maps. As a summary, these position numbers, which will be called plausible positions, are plotted against the position number on the best map.

Results

The example employed is concerned with a DH population of oilseed rape (*Brassica napus*). The data were obtained from 91 individuals which were characterised by 83 binary markers constituting one linkage group. Some details of the data are given in Fig. 1.

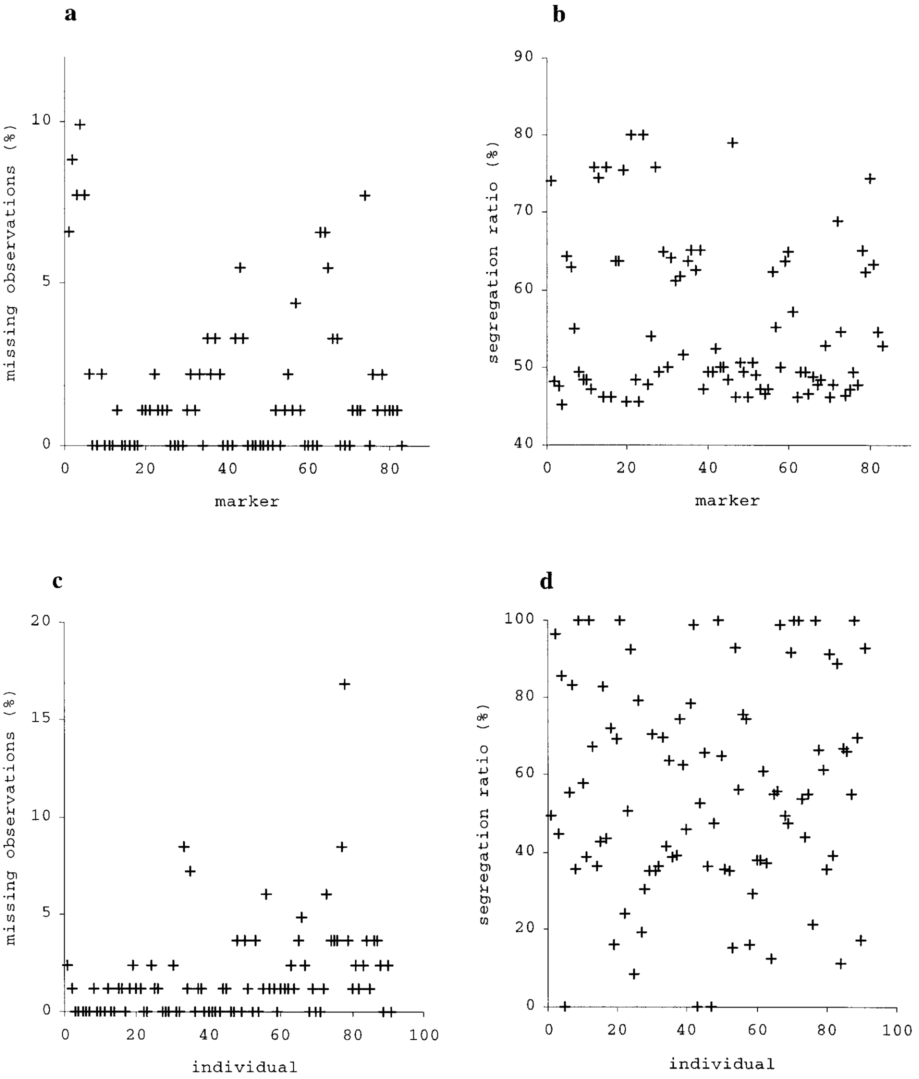
With a few exceptions the percentage of missing observations is smaller than five, both for markers and individuals. The total number of missing observations is equal to 129 (1.7%). For a number of markers the segregation ratio differs clearly from 0.5: one parental haplotype prevails over the other parental haplotype. With regard to the linkage group under study, 11 individuals are non-recombinant: eight individuals consist entirely of one parental haplotype, whereas only three individuals consist entirely of the other parental haplotype. The expected number of non-recombinant individuals depends primarily on the length of the linkage group. As an example, in the case we consider, a 100-cM linkage group of a back-cross or doubled-haploid population with markers at 1-cM intervals, the proportion of non-recombinant individuals is approximately equal to e^{-1}=0.37.

In this example, the initial matrix of expected numbers of recombinants does not contain zeros; values range between 0.5 (i.e. information on one pair of individuals missing and all other pairs of individuals non-recombinant) and 59.5. In this application the total number of recombination events was used as an optimality criterion; the aim is to find the map with the smallest value of this criterion.

Marker selection is employed to obtain a framework map to be used as a starting point for ordering all markers. The selection radius r_0 plays a crucial role in the selection process and the quality of the initial map for all markers. Increasing the value of r_0 has the following effects.

(1) Obviously, it leads to a reduction in the number of markers in the framework map (Fig. 2a). For a given value of r_0 the number of selected markers varies a little. As

Fig. 1a–d Pre-mapping diagnostics for the *Brassica* data. **a** Missing observations (%) versus marker (numbered 1 ... 83). **b** Segregation ratio (%) versus marker (numbered 1 ... 83). **c** Missing observations (%) versus individual (numbered 1 ... 92). **d** Segregation ratio (%) versus individual (numbered 1 ... 92)



a consequence, the size of the computational problem for the framework map is reduced if the value of r_0 is increased. For 83 markers the possible number of marker orderings is equal to 2.0×10^{124}. For ten markers, which is obtained by setting r_0 approximately equal to 0.10, this number is reduced to 1.8×10^{6}.

(2) The length of the map for the selected markers, e.g. the total number of recombinations, becomes smaller (Fig. 2b). The large drop in the number of recombinations is mainly due to the presence of typing errors. In our example the total number of recombinations for the map involving all markers is equal to 248.5. For r_0=0.1 the length of the map obtained for the selected markers is reduced to about 150 recombination events.

(3) The selected markers are more evenly spaced (Fig. 2c). For the map based on all markers the coefficient of variation of the number of recombinations between adjacent markers is more than 0.9; for the map based on a selection of markers obtained with r_0=0.1 this value is reduced to 0.3.

(4) If the value of r_0 is increased, the initial map for all markers becomes longer (Fig. 2d). This is due to the way in which the map for all markers is constructed, i.e. non-selected markers are put in random order around the selected marker they are associated with. It should be noted that if the map for the selected markers is correct, the increase in map length is only due to local imperfections. If all markers are ordered in one go, the search window consists necessarily of the map for all markers. By following the two-stage approach the optimisation process is effectively reduced to a series of local searches with a search window consisting of a small portion of the map, one for each marker.

Our experience is that r_0=0.1 is an appropriate value for obtaining the initial set of markers. This initial set may vary slightly in size. For a typical run, details of the optimisation process are given in Fig. 3a and b. In this run 13 markers were selected; the total number of possible marker orderings is equal to 3.1×10^{9}. The value of the initial acceptance probability of upward steps, π_0, was set equal to 0.25. This resulted in an initial value of γ, the parameter controlling the acceptance of steps with negative values of ℓ equal to 18.6. The cooling control

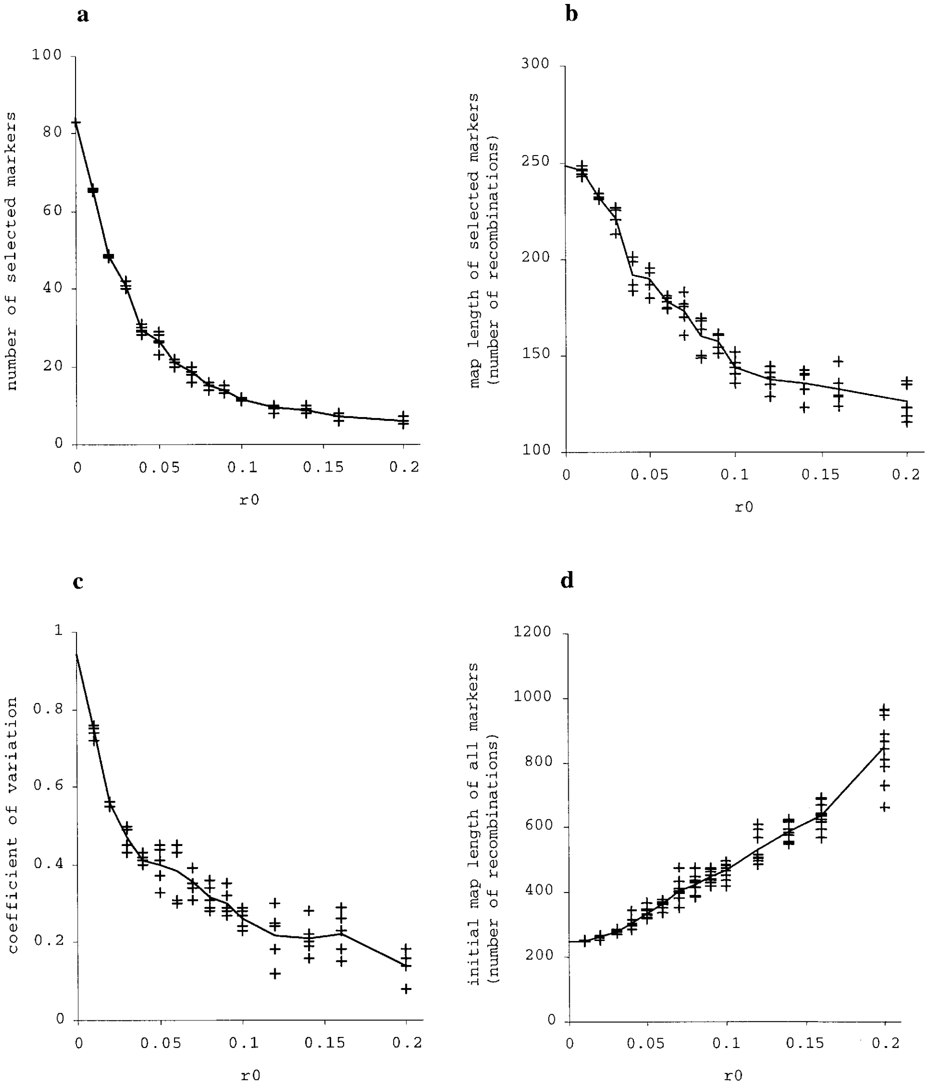


Fig. 2a–d The marker selection process. **a** The number of markers selected for constructing the framework map versus the selection radius r_0. **b** The length (presented as the number of recombinations) of the framework map versus the selection radius r_0. **c** The coefficient of variation of distances (the number of recombinations) between adjacent markers on the framework map versus the selection radius r_0. **d** The initial map length (presented as the number of recombinations) for all markers versus the selection radius r_0.

parameter α was set equal to 0.1, while the chain length was set equal to 1,000. The values given to π_0 and α, as well as the convergence criterion, actually determine the total number of evaluations. In this case we based our check of convergence on the number of times the actual value of the criterion, i.e. the total number of recombinations, is equal to the minimum value of the criterion encountered during optimisation. The map with the 'best' value of the criterion is also stored during optimisation.

If during ten chains no difference was found between the actual value of the criterion and the best value encountered so far, the process was halted. In this case convergence was obtained after 25 chains=25,000 evaluations. It should be noted that the effort required to reach convergence is in sharp contrast with the total number of marker orders. It should also be noted that the map with the minimum number of recombinations was already encountered after six chains=6,000 evaluations.

To illustrate the performance of the procedure for mapping all markers we will continue the previous example. The initial map has 451.5 recombinations. In order to limit extensive excursions of markers during optimisation, the initial value of γ was set equal to 4.0. The cooling control parameter α was set equal to 0.01. Details are given in Fig. 3c and d. The map with the minimum number of recombinations was encountered after 351 chains=351,000 evaluations. In this case the convergence criterion was set equal to 100. Optimisation was halted after 1,388 chains, i.e. 1,388,000 evaluations. In this case the two stages of the mapping process took 13 s on a 200 MHz Pentium II.

Figure 4 shows the graphical genotypes of the 91 individuals based on the map obtained previously. It can be noted that individual haplotypes mainly consist of long stretches of contributions from one parent interchanged with contributions from the other parent. Missing values occur mainly as 'singles' and 'doubles', although one 'string' of five successive missing values occurs in individual 78.

Due to the nature of the estimation procedure many of the missing values are effectively non-missing. Missing observations between two non-recombinant flanking

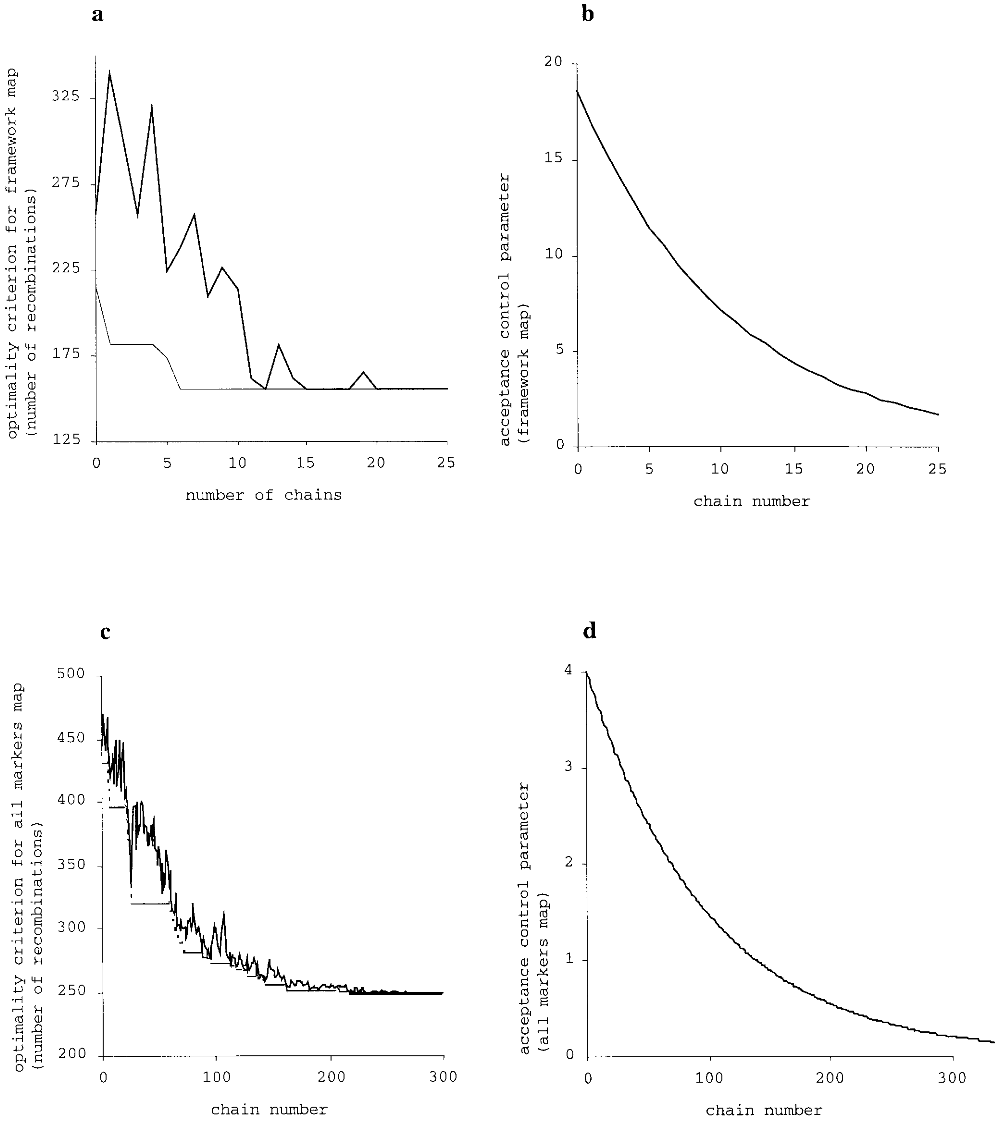


Fig. 3a–d Some details of the optimisation process. **a** The optimality criterion (number of recombinations) for the framework map versus the chain number; the *solid line* represents the current value of the optimality criterion, the *dashed lines* represent the 'best' value of the optimisation criterion encountered so far. **b** The acceptance control parameter (γ) for the framework map versus the chain number. **c** The optimality criterion (the number of recombinations) for the all-markers map versus the chain number; the *solid line* represents the current value of the optimality criterion, the *dashed lines* represent the 'best' value of the optimisation criterion encountered so far. **d** The acceptance control parameter (γ) for the framework map versus the chain number

markers receive the value of the flanking markers with a probability close to unity. Only missing observations between two recombinant flanking markers lead to variation between successive Gibbs samples.

In our application, we start with randomly assigning 0 s and 1 s to missing observations This leads to 234 recombinations. After a few Gibbs samples, the number of recombinations is reduced to values between 144, the minimum number of recombinations, and 149. A burn-in period of 100 Gibbs samples is sufficient for this data. The autocorrelation coefficient between the results of successive Gibbs samples is very small and a lag of 10 is sufficient to guarantee successive samples to be uncorrelated. Five runs with a burn-in period of 100 followed by 100 samples taken every ten Gibbs samples produced average numbers of recombinations equal to 144.9, 145.0, 144.9, 145.1 and 144.9, indicating the low level of variation between different runs.

Missing-value imputation effectively leads to a new, improved estimate of the expected number of recombinations using all available information. This new estimate can be used as a starting point for obtaining a new, improved map. This can be repeated until convergence is attained. After three cycles of simulated annealing and Gibbs sampling, a map was found with an expected number of recombinations equal to 143.1. This value was not further improved in subsequent iterations. Using Haldane's mapping function this is equivalent to 168.0 cM.

Figure 5 shows posterior intervals for the current data. This figure indicates that the effective number of markers is much smaller than 83, i.e. locally many mapping orders are equivalent, or nearly equivalent, with regard to variation in the data. Markers are clustered into approximately 15 'marker groups'. Figure 4 also indicates that positions 81, 82 and 83 are plausible positions for the markers mapped on positions 74 and 75, but not for positions 76 up to 80. This needs further consideration.

Figure 6 shows a number of diagnostic plots that may be useful to check the adequacy of the map obtained.

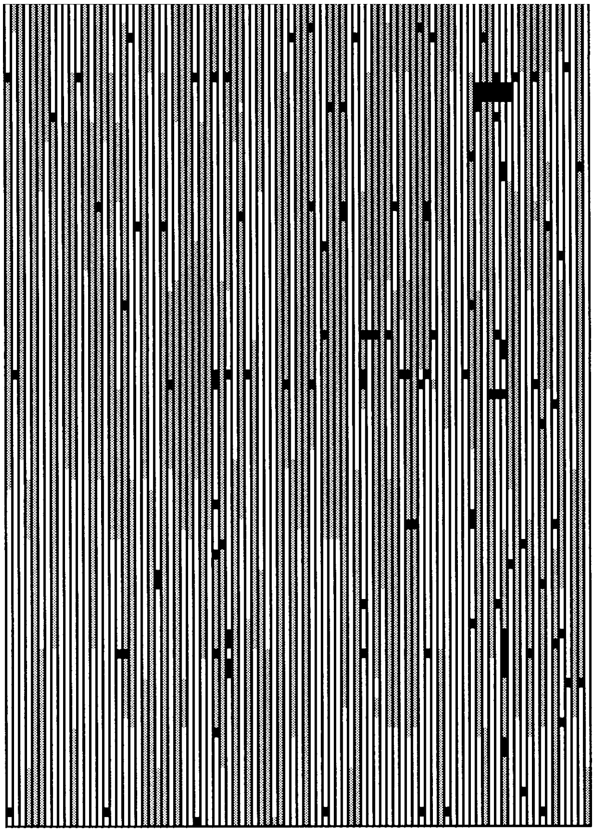


Fig. 4 Graphical genotypes for the *Brassica* data; *rows* represent markers and *columns* represent individuals; *white and grey* represent contributions of the two parents, *black* represents missing observations

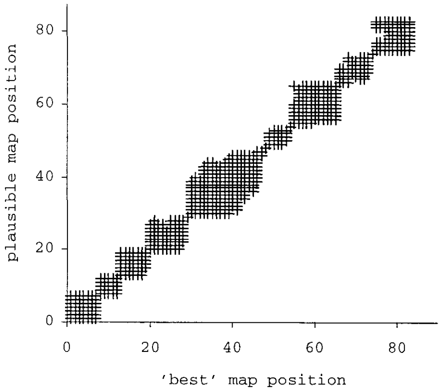


Fig. 5 Posterior intervals for the *Brassica* data; for each marker represented by its 'best' map position (numbered 1 ... 83), plausible map positions are indicated

Figure 6a shows that there is no indication that markers with a large number of missing observations end up at one side of the map. However, in this application the percentages of missing observations are quite small. Figure 6b shows that the segregation ratio slowly rises from 50% to 75% going from left to right on the map. This may be due to a gene favouring one of the parental haplotypes on the right-hand side of the map. Figure 6c shows that number of recombination events are evenly spread across individuals. Large numbers of recombination events may be expected for individuals whose DNA has been mixed-up in one way or another.

Discussion

The two-stage algorithm for genetic linkage mapping as described in this paper is simple and straightforward. It makes full use of the fact that markers are linearly ordered on chromosomes. In the algorithm all markers are treated equally. Approaches, in which markers are sequentially added to the map according to some criterion, e.g. the JoinMap algorithm (Stam 1993), run into computational problems when maps become dense and effects of typing errors on the mapping order become serious. The random nature of the algorithm described in this paper makes it possible to circumvent local optima, many of which may exist in the case where the effects of typing errors become serious. The two-stage approach guarantees that we will get close to the global optimum. This may be the reason that the simple cooling scheme adopted in this paper worked well in practical applications. Even in the first stage of the algorithm when the number of markers is small, simulated annealing is much quicker than an exhaustive search of all possible marker orders (Lander and Green 1987).

When the data are complete, all information about crossing-over in backcross and (doubled) haploid populations is contained in the numbers of recombinations between pairs of markers which can be calculated directly from the molecular-marker data. In F_2 populations and full-sib families of outbreeding species molecular-marker data are usually not completely informative with regard to the numbers of recombinations between pairs of markers (Maliepaard et al. 1997).

However, given the segregation types and observations, the Gibbs sampler can be used to calculate the expected numbers of recombinations between pairs of markers, perhaps even for both parents separately. These expected numbers can then be used in a first iteration of the mapping process. The map obtained from the first iteration can then be used to further improve conditional expectations using the Gibbs sampler. In the current application Gibbs sampling only involves counting integer numbers of recombinations. In these situations most observations have to be treated as 'missing'. As a consequence, the efficiency of missing-data imputation needs careful attention. In the current situation use can be made of the fact that, if the recombination frequencies with both flanking markers are smaller than 0.05, the probability of a double recombination is very close to zero.

Any mapping procedure must be accompanied by tools for checking both data and results. For example, a sub-division of the total number of recombinations or the log-likelihood into individual contributions of plants should be in line with expectations based on the underly-

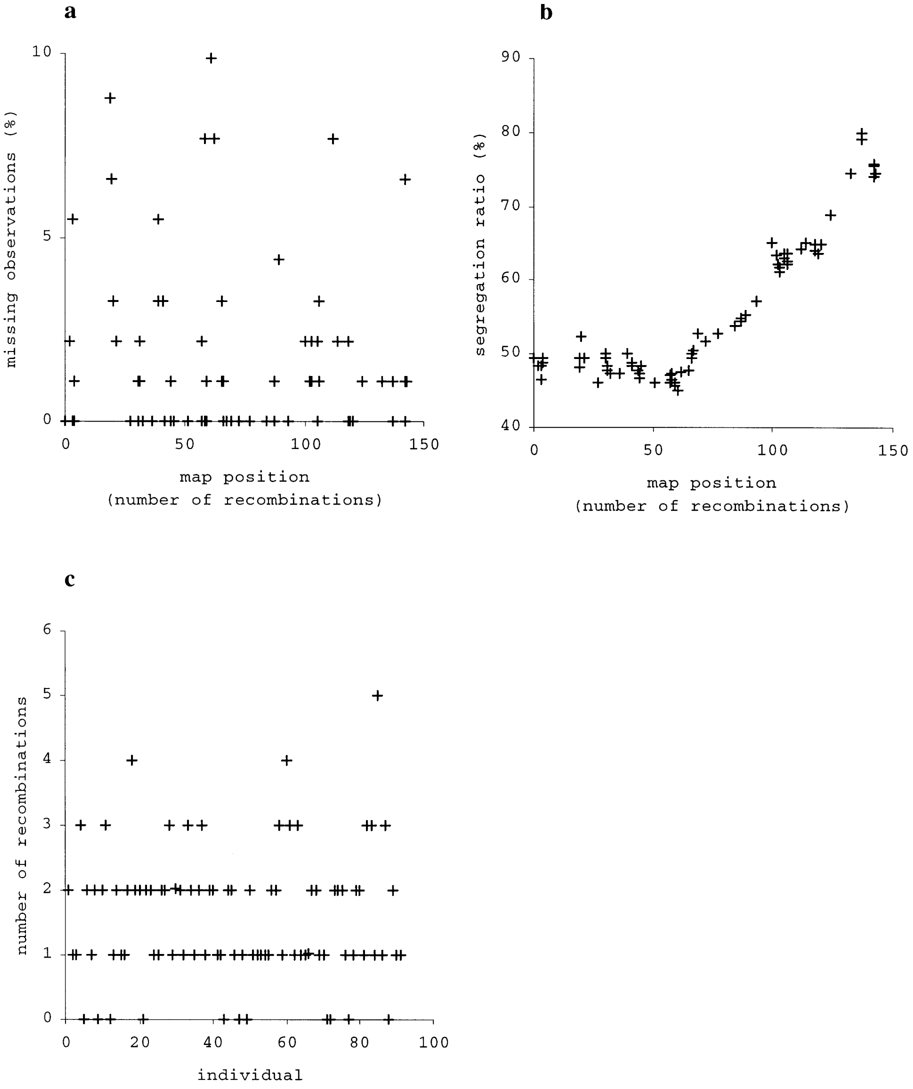


Fig. 6a–c Post-mapping diagnostics. **a** Missing observations (%) versus map position (number of recombinations). **b** Segregation ratio (%) versus map position. **c** Number of recombinations versus individual (numbered 1 ... 92)

ing model. Computer simulation may be helpful in this respect. The ‘posterior’ intervals for marker positions provide information about the resolution the data provide about the order into which the markers are put on the map. They can also be useful in the design stage of mapping experiments. They can be applied to investigate the effects of the numbers of plants, the distribution of markers and the typing errors on the resolution of maps. Segregation ratios differing greatly from 50% may pose a serious problem. Markers with similar segregation ratios (different from 50%) will clump together on the map. They may be produced by selection at the gametic or zygotic level. They are often encountered in interspecific crosses. For applications of maps in QTL analysis this will not be a problem as QTLs will have the same segregation ratios as closely linked markers. They may also be the result of a failure in the process of data production. Therefore, one must be sure that the latter possibility is excluded.

The process of constructing ‘posterior’ intervals is started from the ‘best’ map. The most important argument is that we want to find maps which although not ‘best’ are very ‘plausible’ considering the variation in the data. However, the ‘posterior’ intervals could become too narrow. The reason is that we only allow jumps involving one marker and by doing so are not able to reach all ‘plausible’ maps starting from the ‘best’ map. Thus, if we start from ‘any’ map we would in many cases end up in some local optimum (‘absorbing barrier’) and not be able to reach the ‘best’ map.

Acknowledgements We are grateful to Aventis Crop Science N.V. for providing the *Brassica* data, and to our colleague to Dr. Michiel Jansen who suggested using the Metropolis-Hasting algorithm for obtaining posterior intervals and to Dr. Martin Boer for helpful advice. A computer programme written in C was used to carry out the calculations. The methods will be made available in a future version of the computer package JoinMap .

References

Falk CT (1992) Preliminary ordering of multiple linked loci using pairwise linkage data. Genet Epidemiol 5:75–80

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 2200:671–680

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

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

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding species: overview and consequences for applications. Genet Res 70:237–250

Ridout MS, Tong S, Vowden CJ, Tobutt KR (1998) Three-point linkage analysis in crosses of allogamous plant species. Genet Res 72:111–121

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Stam P, Van Ooijen JW (1995) JoinMap (tm) version 2.0: Software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen

Weeks DE, Lange K (1987) Preliminary ranking procedures for multilocus ordering. Genomics 1:236–242