Euphytica (2009) 165:401–417
DOI 10.1007/s10681-008-9805-3

Ordering dominant markers in F_2 populations

J. Jansen

Received: 17 April 2008 / Accepted: 2 September 2008 / Published online: 26 September 2008

Abstract Ordering dominant markers in F_2 populations is considered a difficult problem. The difficulties arise from the fact that recombinations cannot be observed directly from the data. In general, the multi-point maximum likelihood would be the appropriate criterion for ordering markers. This criterion takes into account all available information present in marker data. However, calculation of multi-point maximum likelihoods is very time-demanding, especially if the number of markers is large. In this paper, ordering markers by minimising the number of recombinations between adjacent markers is used as a simple alternative to multi-point maximum likelihood. Contrary to multi-point maximum likelihood, this method does not involve any probability assumptions about the occurrence of recombination events. Simulated data indicate that the minimum number of recombinations between adjacent markers is approximately a linear function of the map length obtained by multi-point maximum likelihood. As a consequence, it will lead to more or less the same optimum marker orders. Optimisation of marker orders with regard to the number of recombinations between adjacent markers is carried out by a modified form of simulated annealing. The reliability of the resulting marker orders is studied by generating marker orders that are plausible with the data using a Metropolis algorithm.

Keywords Co-dominant marker · Dominant marker · F_2 population · Inheritance vector · Minimum number of recombinations · Multi-point maximum likelihood

Introduction

With dominant markers, construction of linkage maps using data from F_2 populations is considered problematic (Knapp et al. 1994; Liu 1997). In this paper, an F_2 population is considered to originate from the selfing of an individual, which itself is obtained by crossing two pure lines (Allard 1960). A dominant marker is a marker of which the presence of one allele can be observed. In an F_2 population, markers can be divided into two types. For markers of type I, the observable allele is present in the second grandparent of the F_2 pedigree, for markers of type II the observable allele is present in the first grandparent. Markers of the same type are in said to be in coupling phase, whereas markers of different types are in repulsion phase. For a given type, it is usually easy to obtain linkage maps. The problem lies in integrating the linkage maps of the two types into one linkage map for the F_2 population.

Knapp et al. (1994) state that it is not possible to obtain reliable maximum likelihood estimates of the

J. Jansen (✉)
Biometris, P.O. Box 16, 6700 AC Wageningen,
The Netherlands
e-mail: johannes.jansen@wur.nl

recombination frequency between two markers that are in repulsion phase. The maximum likelihood estimate will practically always be zero, if the number of individuals in the F_2 population, which do not carry the observable allele for both markers, is zero. Of course, the chance that this happens depends on the number of individuals, but it may be appreciable, even for large values of the recombination frequency. As a consequence, map construction based on maximum likelihood estimates of the recombination frequency for pairs of markers, the so-called two-point maximum likelihood estimates, does not lead to reliable linkage maps. For the F_2 with dominant markers, Tan and Fu (2007) developed improved two-point estimates by taking averages over three-point maximum likelihood estimates. Three-point estimates of recombination of recombination frequencies were earlier used by Ridout et al. (1998) for crosses of outbreeding species.

Rather than using two-point maximum likelihood estimates of the recombination frequencies, it would be possible to evaluate different marker orders using the multi-point maximum likelihood (Lander and Green 1987). Rather than considering two markers at a time, multi-point maximum likelihood considers the whole sequence of markers for each marker order. A major advantage would be that it takes into account all information present in the data for all markers simultaneously. However, a major disadvantage is that the time required for calculating multi-point maximum likelihood estimates of recombination frequencies becomes prohibitive if the number of markers increases. It should be noticed that in the case of an F_2 with dominant markers on average 75% of the marker data are incomplete.

In this paper, we consider a simple alternative to multi-point maximum likelihood. The method is based on minimising the number of recombinations between adjacent markers using hidden inheritance vectors (Jansen 2005). Inheritance vectors (Lander and Green 1987), also known as segregation indicators (Thompson 1994) provide in the form of a binary code the grandparental origin of alleles. As a consequence, they allow determining numbers of recombinations between adjacent markers by simply counting recombinant individuals. However, in the case of F_2 populations with dominant markers, inheritance vectors are not unique. Therefore, the problem consists of simultaneous minimisation of the number of recombinations with regard to marker order and inheritance vectors. Only inheritance vectors that are consistent with the marker data are being considered in the optimisation process. The major advantage of the proposed method is that it does not require estimates of recombination frequencies during the optimisation process. After finding the optimum order, multi-point maximum likelihood estimates have to be obtained for the optimum order only by the EM algorithm (Dempster et al. 1977) or a stochastic alternative based on the Gibbs sampler (Jansen et al. 2001a, b).

Three examples, two with simulated data and one with real data, will be considered. The first example uses a small set of simulated data to compare three criteria: two-point maximum likelihood, multi-point maximum likelihood and minimum number of recombinations between adjacent markers. The second example uses a large set of simulated data to investigate ordering markers by minimising the number of recombinations between adjacent markers in the case of dense linkage maps. The third example uses real data in a situation where a part of the marker data have been scored dominantly and another part co-dominantly.

Methods

Marker data

A molecular genetic marker (shortly marker) is considered as a point or locus on a pair of homolog chromosomes of a diploid species. A marker is a short segment of DNA, of which different variants or alleles exist. Different alleles arise from differences in DNA sequence. In this paper the emphasis will be on markers of which only one allele can be observed. The alleles that cannot be observed are called null alleles.

In Fig. 1, the segregation of a marker involving a null allele is shown in detail. In the F_2 mating system, the grandparents are homozygous, and as a consequence, only two alleles exist. For type I, the second grandparent is homozygous with regard to the observable allele; the first grandparent is homozygous with regard to the unobservable allele. For type II, the first grandparent is homozygous with regard to the observable allele, and the second grandparent is

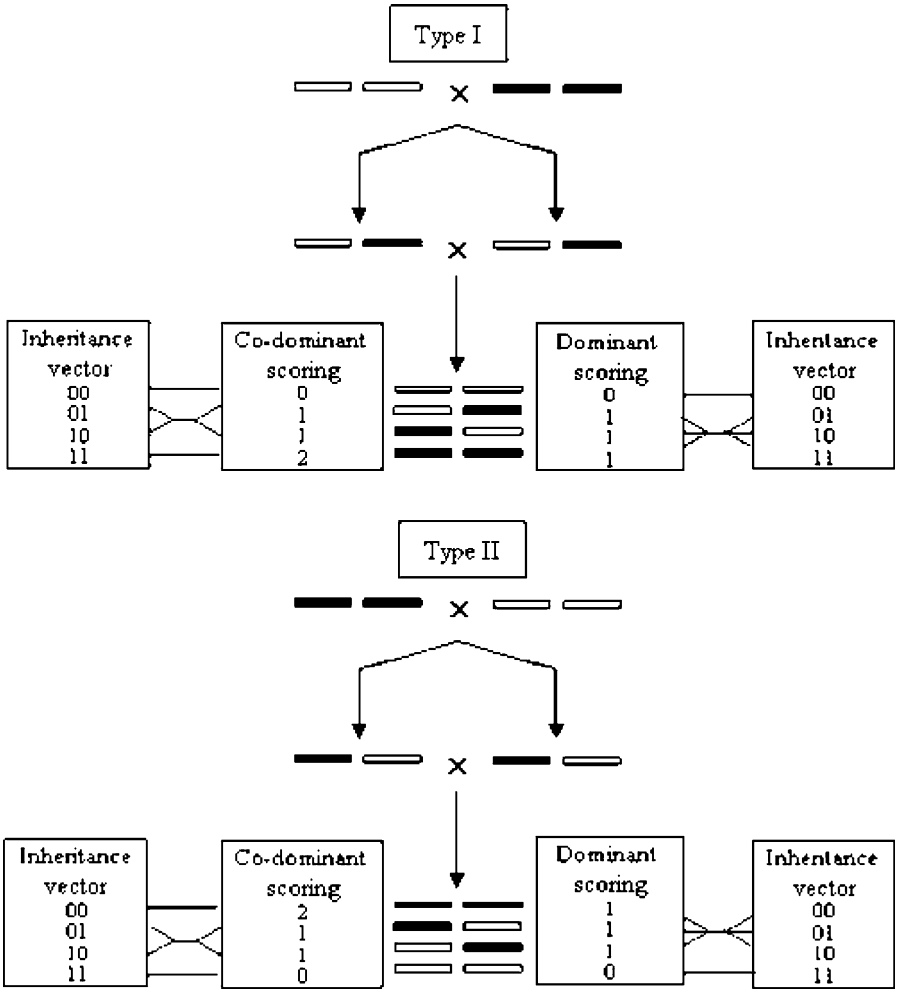

Fig. 1 Segregation of a dominant marker in the F_2: ▬, observable allele; ▭, unobservable allele

homozygous with regard to the unobservable allele. In the F_1 population, obtained by crossing the grandparents, all individuals are identical and heterozygous. The F_2 population is obtained by selfing an F_1 individual.

With regard to the marker under study, the F_2 consists of individuals of four different genotypes. Individuals are either identical to:

1. the first grandparent (with inheritance vector 00), or
2. the second grandparent (with inheritance vector 11), or
3. the F_1 (with inheritance vector 01), or
4. the F_1 with alleles in reversed order (with inheritance vector 10).

A co-dominant scoring system is able to determine the number of copies of the observable allele. In the F_2 population, individuals either carry 0, 1 or 2 copies of the observable allele. For type I, an individual with 0 copies has inheritance vector 00 and an individual with 2 copies has inheritance vector 11. An individual with 1 copy either has inheritance vector 01 or 10. For type II, an individual with 0 copies has inheritance vector 11 and an individual with 2 copies has inheritance vector 00. Again, an individual with 1 copy either has inheritance vector 01 or 10.

A dominant scoring system is able to determine whether the observable is present (coded 1) or not (coded 0). For type I, an individual coded 1 has either inheritance vector 01, 10 or 11. An individual coded 0 has inheritance vector 00. For type II, an individual coded 1 has either inheritance vector 00, 01 or 10. An individual coded 0 has inheritance vector 11. Two markers of type I, or two markers of type II, are said to be in coupling phase; a marker of type I and a marker of type II are in repulsion phase.

If the grandparents are not available anymore, it is not possible to distinguish between type I and type II. In that case, the phase, i.e. the order of the observable and the unobservable allele, in the F_1 is unknown, and has to be determined from the marker data. The problem of phase determination will not be considered in this paper; see Jansen (2005).

Optimisation problem

Marker data obtained from an F_2 population provide incomplete information about the inheritance vectors, even if the marker data are obtained using a co-dominant scoring system. For a dominant scoring system, ~25% of the marker data correspond with exactly one inheritance vector, whereas ~75% of the marker data correspond to three different inheritance vectors. Only missing marker data are worse: they correspond to all four different inheritance vectors.

The multi-point maximum likelihood would be the appropriate criterion for ordering markers. The problem is to find the marker order with the greatest value of the multi-point maximum likelihood. For each marker order, that is proposed in the optimisation process, multi-point maximum likelihood estimates of the recombination frequencies have to be obtained. This has to be done iteratively, e.g. by means of the EM algorithm (Dempster et al. 1977). In this paper, the E-step of the EM algorithm is carried out using Gibbs sampling.

Minimising the total number of hidden recombinations between adjacent markers will be considered as an easy alternative to maximising the multi-point maximum likelihood. A theoretical comparison of the methods is given in “Appendix.” In this approach, only inheritance vectors that are consistent with the marker data are allowed. Given a set of inheritance vectors that are in agreement with the data, it is possible to count the total number of recombinations between adjacent markers. In this case, the problem is to find the combination of marker order and set of inheritance vectors that result in the smallest number of recombinations between adjacent markers summed over all individuals of the F_2 population. A detailed description is given by Jansen (2005).

Optimisation

Optimisation is carried out by a modified form of simulated annealing (Jansen 2005). A general treatment of simulated annealing is given by Kirkpatrick et al. (1983) and van Laarhoven and Aarts (1987). In general, proposals for improving the optimality criterion could be obtained by proposing combinations of changes of the marker order and changes of the inheritance vectors. However, for reasons of simplicity proposals for improving the marker order are treated separately from proposals for improving the inheritance vectors.

The increase of the number of recombinations due to a proposal will be denoted by Δ. A proposal for the marker order will always be accepted if $\Delta \leq 0$. It will also be accepted if $\exp(-\Delta/\gamma) \geq u$, in which γ is the acceptance control parameter and u is a random number from the standard uniform distribution. For marker orders optimisation starts with a large value for γ and during the optimisation process the value of γ is decreased slowly to zero. The algorithm mainly uses proposals involving one marker at a time (cf. Jansen et al. 2001a, b). Close to the optimum also proposals are used that involve changing the order of windows of two up to five adjacent markers. A proposal for changing an inheritance vector is only accepted if the number of recombinations is not increased. The two types of proposals (for orders and inheritance vectors) are applied in an alternating way.

Construction of framework maps

In the case of a dense marker map, it may be worthwhile to construct first a framework map using a small number of markers. These markers should be chosen in such a way that they cover the whole linkage group. Selection of markers for the framework is done using the spatial sampling procedure described by Jansen et al. (2001a, b). An application involving simulated data in which all markers are scored dominantly is presented.

In the case of an F_2 with dominant markers two sets of markers will be selected, one set representing the markers of type I and one set representing the markers of type II. The selection radius that is used for the spatial random sampling procedure is 0.05, i.e. all markers, for which the number of recombinations with selected markers is equal to 0.05 * (2N)

(in which N represents the number of individuals in the offspring population), are temporarily discarded. In this case, the selected markers will lie approximately 10 cM apart.

Subsequently, in order to obtain a framework map the algorithm is applied to the selected markers only. After obtaining an optimum, markers that were first discarded are now placed on the map near the marker they are closely linked to. Finally, the algorithm is applied to all markers allowing only local changes in the marker order.

Calculation of plausible maps

Starting from the optimum configuration, a set of plausible marker orders is obtained by accepting all proposals that do not lead to an increase of the number of recombinations and, if proposals leads to an increase of the number of recombinations, by accepting those proposals for which $\exp(-\Delta) > u$, in which Δ is the increase of the number of recombinations and u is a random number from the standard uniform distribution.

Proposals that are used consist of (1) a change of order of a groups of two up to five successive markers chosen at random, followed by (2) a number K of changes of randomly chosen inheritance vectors, which do not lead to an increase of the number of recombinations. In the current applications the value of K has been set equal to a fixed number (1,000). The results in this paper have been obtained using a burn-in of 1,000 proposals; the sequence of marker orders is sampled at intervals of 1,000 proposals.

Applications

A comparison of methods

Ordering markers by minimising the number of hidden recombinations

In this section, use is made of simulated marker data concerning 11 markers. Marker data were obtained for 100 individuals. The positions of the markers were set equal to 0, 10,…, 100 cM (Haldane mapping function).

First, inheritance vectors were generated as shown in Fig. 2a. They represent exactly the way alleles have been transmitted from grandparents via the parents (F_1) to the individuals of the F_2 population. From the inheritance vectors, the number of recombinations between adjacent markers can be obtained by simply counting recombination events. From the inheritance vectors, it is also possible to distinguish between recombination events in the maternal and in the paternal meiosis. For the true marker order, the number of recombinations between adjacent markers is equal to 193.

Secondly, the inheritance vectors were recoded into marker data that would have been obtained using co-dominant scoring (Fig. 2b). A major consequence is that it is not possible anymore to distinguish between inheritance vectors 01 and 10. It should be noticed that in an F_2 population, 50% of the marker observations have either inheritance vector 01 or 10. In the following, inheritance vector 01 will used to represent inheritance vector 01 and 10. For the true marker order, the minimum number of recombinations between adjacent markers is equal to 183:10 recombinations have been lost due to the fact that in the original inheritance vectors at five positions inheritance vector 01 was adjacent to inheritance vector 10 (e.g. see Fig. 2, individual 96). A possible reconstruction of the inheritance vectors based on the co-dominant scores is shown in Fig. 2c.

Thirdly, the original inheritance vectors were recoded into marker data that would have been obtained using dominant scoring (Fig. 2d). For markers 1, 3,…, 11 the allele of the second grandparent is observable. For these markers no distinction can be made between inheritance vectors 01, 10 and 11. For markers 2, 4,…, 10 the allele of the first grandparent is observable. For these markers no distinction can be made between inheritance vectors 00, 01 and 10.

For the true marker order, the minimum number of recombinations between adjacent markers is equal to 145. So, 48 recombinations have been lost due to dominant scoring. The minimum value of 145 recombinations is not only obtained for the true marker order, but also for many other marker orders. A possible reconstruction of the inheritance vectors for the true marker order with the minimum number of recombinations between adjacent markers is shown in Fig. 2e. It can be observed from Fig. 2e that many valid reconstructions of the inheritance vectors can be obtained for the true marker order with

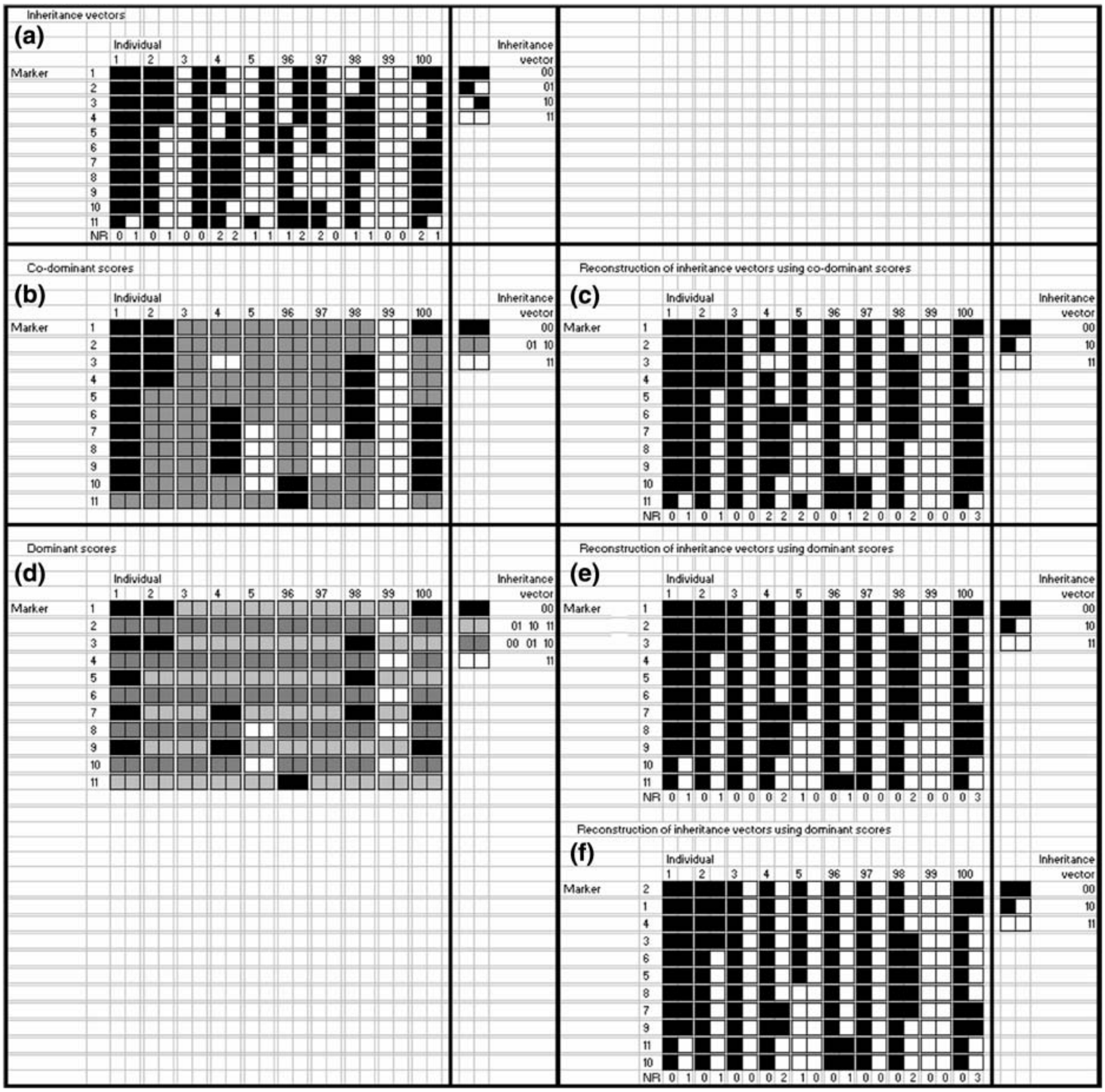

Fig. 2 The effect of co-dominant and dominant scoring on the recovery of the transmission of alleles from the parents (F_1) to the individuals of the F_2 population: (**a**) true inheritance vectors; (**b**) co-dominant scores; (**c**) a reconstruction of the inheritance vectors with minimum NREC; (**d**) dominant scores; (**e**) a reconstruction of the inheritance vectors with minimum NREC; (**f**) an alternative reconstruction of the inheritance vectors with minimum NREC

145 recombinations. In many instances, the dominant marker data do not provide information about the exact location of recombination events. Figure 2f shows a valid reconstruction of the inheritance vectors for a marker order which differs markedly from the true marker order but still has the same minimum number of recombinations between adjacent markers.

Ordering markers using maximum likelihood

For the above situations it is also possible to estimate recombination frequencies between markers. This can be done in two ways: by two-point maximum likelihood or by multi-point maximum likelihood. If the data are complete, as is the case with the initially generated inheritance vectors, two-point maximum likelihood and multi-point maximum likelihood give results that are identical to simply counting of recombination events. In this case, the total number of recombinations between adjacent markers is calculated as 200 (= twice the number of individuals) times the sum of the recombination frequencies between adjacent markers.

For the co-dominantly scored marker data, the total number of recombinations obtained with two-point maximum likelihood and with multi-point maximum likelihood were equal to 192.1 and 192.4, respectively, for the true marker order. However, a major difference occurs if the dominantly scored marker data are considered. Now, for the true marker order, the total number of recombinations between adjacent markers obtained with two-point maximum likelihood is equal to 41.2. The total number of recombinations between adjacent markers obtained with multi-point maximum likelihood is again equal to 192.4. A summary of the results is given in Table 1.

Several conclusions can be drawn from the above example. Firstly, for the true marker order multi-point maximum likelihood is able to recover nearly all recombination events between adjacent markers, even if the markers have been scored dominantly. Secondly, methods based on two-point maximum likelihood are not at all suitable for an F_2 with dominant markers. This is so, because many of the recombination frequencies between markers that are in repulsion phase will be estimated by zero. Thirdly, ordering markers by minimising the number of recombinations between adjacent markers leads to fewer recovered recombinations than multi-point maximum likelihood.

A comparison of methods

In the following, multi-point maximum likelihood and minimum number of recombinations between adjacent markers will be studied in more detail. Apart from the true marker order, it is possible to consider other marker orders. Marker orders that are slightly differing from the true marker order are obtained by changing the order of windows of two, three, four or five successive markers. Marker orders that are greatly differing from the true marker order are obtained by taking random marker orders. In the following 25 random markers orders were used. The effect of changing the marker order was studied both for the simulated data with co-dominant markers and with dominant markers. The results are shown in Fig. 3. In Fig. 3, the minimum number of recombinations between adjacent markers is plotted against the number of recombinations between adjacent

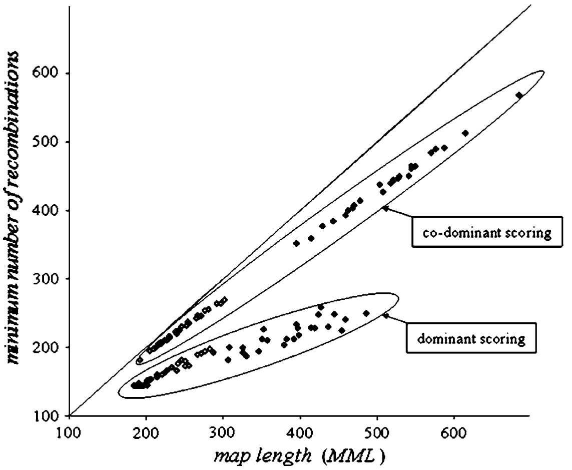

Fig. 3 Minimum number of recombinations versus the map length (defined as twice the number of individuals times the sum of recombination frequencies between adjacent markers) obtained using multi-point maximum likelihood; a. co-dominant markers, b. dominant markers (◇: orders obtained by changing the order of a group of g adjacent markers ($g = 2,\ldots, 5$) relative to the true marker order; ◆: random orders)

Table 1 A summary of results for the simulated data for the comparison of methods

Case	Method	Number of recombinations between adjacent markers
Full information		193
Co-dominant markers	Minimum number of recombinations	183
	Two-point maximum likelihood	192.1
	Multi-point maximum likelihood	192.4
Dominant markers	Minimum number of recombinations	145
	Two-point maximum likelihood	41.2
	Multi-point maximum likelihood	192.4

markers as obtained by multi-point maximum likelihood.

In both cases a more or less linear relationship exists between the total number of recombinations obtained by minimising the number of recombinations and by multi-point maximum likelihood. A major consequence is that both methods will eventually lead to the same 'best marker order' although some deviations may occur, especially if all markers are scored dominantly.

Plausible maps

Table 2a shows a summary of 1,000 optimisations, each starting from a random marker order and a random choice of inheritance vectors. All optimisations resulted in a marker order with 145 recombinations between adjacent markers. For each marker, Table 2a gives the number of times the optimum position was 1, 2,…, 11. It should be noticed that marker numbers are identical to the true positions of the markers.

It can be observed from Table 2a that most of the times the optimum position of a marker is also its true position. However, in quite a number of cases the optimum position of a marker is a position next to the true position.

Table 2b shows a summary of 1000 plausible marker orders obtained as described in "Methods." Table 2b shows that the distributions of plausible marker positions are flatter and wider than the distributions of optimum positions. It can be observed from Table 2b that marker orders which deviate considerably from the true marker order are plausible given the observed marker data.

A dense genetic linkage map with dominant markers

In an F_2 population, the problem of obtaining a dense linkage map using dominant markers only consists of integrating two sets of markers that contain very little information about each other. In order to investigate

Table 2 (a) Summary of the results of 1,000 optimizations (all resulting in a marker order with 145 recombinations); (b) summary of the results of a sample of 1,000 plausible marker orders obtained using a Metropolis algorithm

Marker	1	2	3	4	5	6	7	8	9	10	11
(a) Optimum position											
1	885	115									
2	115	885									
3			887	113							
4			113	884	3						
5				3	890	107					
6					107	859	34				
7						34	950	16			
8							16	660	324		
9								324	676		
10										694	306
11										306	694
(b) Plausible position											
1	569	417	14								
2	418	545	37								
3	11	38	655	296							
4	2		294	635	68	1					
5				69	711	220					
6					220	640	138	2			
7					1	139	804	56			
8							56	535	403	6	
9							2	407	575	10	6
10									11	573	416
11									11	411	578

the construction of a dense genetic linkage map with dominant markers, a data set was simulated with 101 markers numbered 1, 2,…, 101. The distance between adjacent markers was set equal to 1 cM (Haldane mapping function). Marker data were simulated for 100 individuals.

As in the previous section, the marker data were generated using inheritance vectors. For the true marker order, the number of recombinations between adjacent markers is equal to 223. Data for an F_2 with dominant markers were obtained by changing inheritance vectors into the appropriate marker data types of the offspring population: markers 1, 3,…, 101 are of type I (no distinction can be made between inheritance vectors 01, 10 and 11) and markers 2, 4,…, 100 are of type II (no distinction could be made between inheritance vectors 00, 01 and 10). For the true marker order, the number of recombinations between adjacent markers is equal to 210.

In general, simply minimising the number of recombinations between adjacent markers does lead to optimum marker orders with a minimum much greater than 210. For a typical optimum, Fig. 4a shows the relationship between the optimum marker order and the true marker order. In this case, the minimum number of recombinations between adjacent markers is equal to 362. From Fig. 4a it can be observed that sections of the optimum marker order are correct, but also that sections have disintegrated and are ordered in the wrong direction. The results of this optimum will be referred to as the result of the first round of ordering all markers.

The problem of disintegration can be resolved by applying spatial sampling of a subset of markers covering the whole linkage group. In this case, we have to take separate samples from the set of markers of type I and from the set of markers of type II. In this case we used the number of recombinations between all pairs of markers as obtained from the first round as the basis for sampling the two subsets of markers. By taking a selection radius of 0.05, 32 markers were selected: 16 of type I and type II. Markers which were

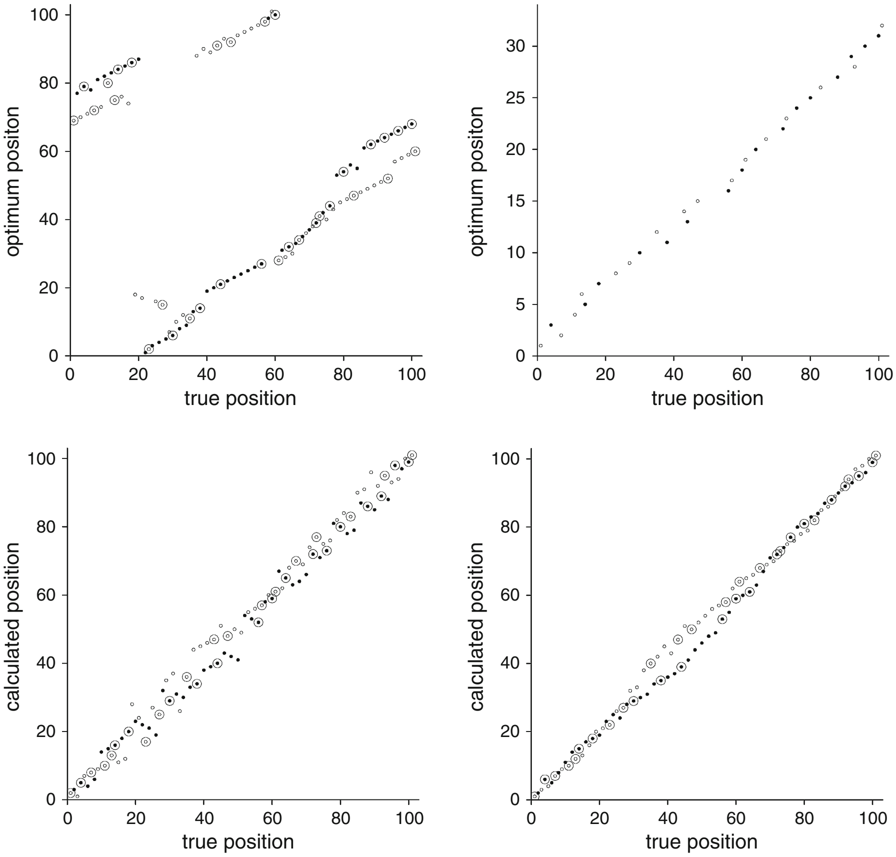

Fig. 4 Graphical representation of ordering 101 dominant markers; •: marker of type I; ○: marker of type II; ◯: marker selected using spatial sampling with a selection radius of 0.05. (**a**) The optimum marker order versus the true marker order after the first round of ordering all markers. (**b**) The optimum marker order versus the true marker order for the markers selected using spatial random sampling. (**c**) The marker order obtained after including the markers that were excluded using spatial random sampling. (**d**) The optimum marker order versus the true marker order after the second round of ordering all markers

Table 3 Results of spatial sampling

Type I							Type II					
1	3						**4**	2	6	8		
7	5	9					**14**	10	12			
11	–						**18**	16	20	22	24	26
13	15	17					**30**	28	32	34		
23	–						**38**	36				
27	19	21	25	33			**44**	40	42	46	48	50
35	29	31					**56**	52	54			
43	37	39	41				**60**	58				
47	45	49	51				**64**	62	66	68	70	
57	53	55					**72**	74				
61	59	63					**76**	–				
67	65	69					**80**	78	82	84		
73	71	75	77				**88**	86	90			
83	79	81					**92**	94				
93	85	87	89	91	95	97	**96**	98				
101	99						**100**	–				

Note: Selected markers (bold) of types I and II and markers associated with selected markers

excluded by being in the selection radius of 0.05 of a selected marker, were stored in a group associated with that selected marker (see Table 3).

It is much easier to order 32 markers than the whole set of 101 markers. Figure 4b shows the relationship between the optimum marker order (within the set of 32 markers) and the true marker order (in the set of 101 markers). Deviations from the true marker order are inevitable due the low information content of the markers.

In Fig. 4c, markers that were excluded using spatial sampling and temporarily stored, are added again. They are placed randomly to the left or to the right of the marker they are associated with. The result is shown in Fig. 4c. This marker order, for which the minimum number of recombinations between adjacent markers is equal to 373, forms the starting point for the second round in which all markers are ordered again. The optimum marker order is shown in Fig. 4d. The optimum marker order shown in Fig. 4d is one of many marker orders with a minimum number of recombinations between adjacent markers equal to 210.

Figure 5a shows plausible positions for the 101 dominant markers. For the current set of data, Fig. 5a shows that positions some five positions away from the optimum position may still be plausible given the information provided by the data. To show the effect of dominant scoring compared to co-dominant scoring on the plausible positions, Fig. 5b shows plausible positions for the 101 markers if they would have been scored co-dominantly. Although the plausible positions for dominant scoring may be acceptable, the results for co-dominant are very much to be preferred.

Capsicum data

The data are obtained from an F_2 population that was obtained by selfing the F_1 between *Capsicum annuum* var. *Jatilaba* and *C. chinense* accession PRI95030. Full details are given by Voorrips et al. (2004). The set of data that will be considered consists of 20 markers of one linkage group scored on 145 individuals. A summary of the data is given in Table 4. The markers comprise one linkage group of the *Capsicum* genome (cf. Voorrips et al. (2004), Fig. 2, linkage group B).

The order of the markers is the one obtained by the JoinMap regression algorithm (Stam 1993; Van Ooijen 2006) using the default settings for the calculation options. Application of JoinMap requires three rounds: only in the third round markers 108 and 161 are placed on the map. The mean *chi*-square contributions for these markers are 4.4 and 3.6, respectively, which is much greater than the expected contribution of one. The JoinMap regression algorithm provides just a single map.

Constructing a map using the new algorithm does not proceed without difficulty

Repeated application of the new algorithm gives many different 'optimum' map orders. Typical orders are given in Table 5. Orders with 191 recombinations occur most often. It is clear from Table 5 that these orders may be very different. Very worrying is the order with 196 recombinations, in which markers 108 and 103 appear next to each other. In the other orders, with only a few recombinations less, they appear at the outer ends of the map. The map order obtained by the JoinMap regression algorithm (though with an entirely different optimality criterion) has 204 recombinations. A conclusion: the maps obtained from these data cannot be relied upon, but what is causing the problems?

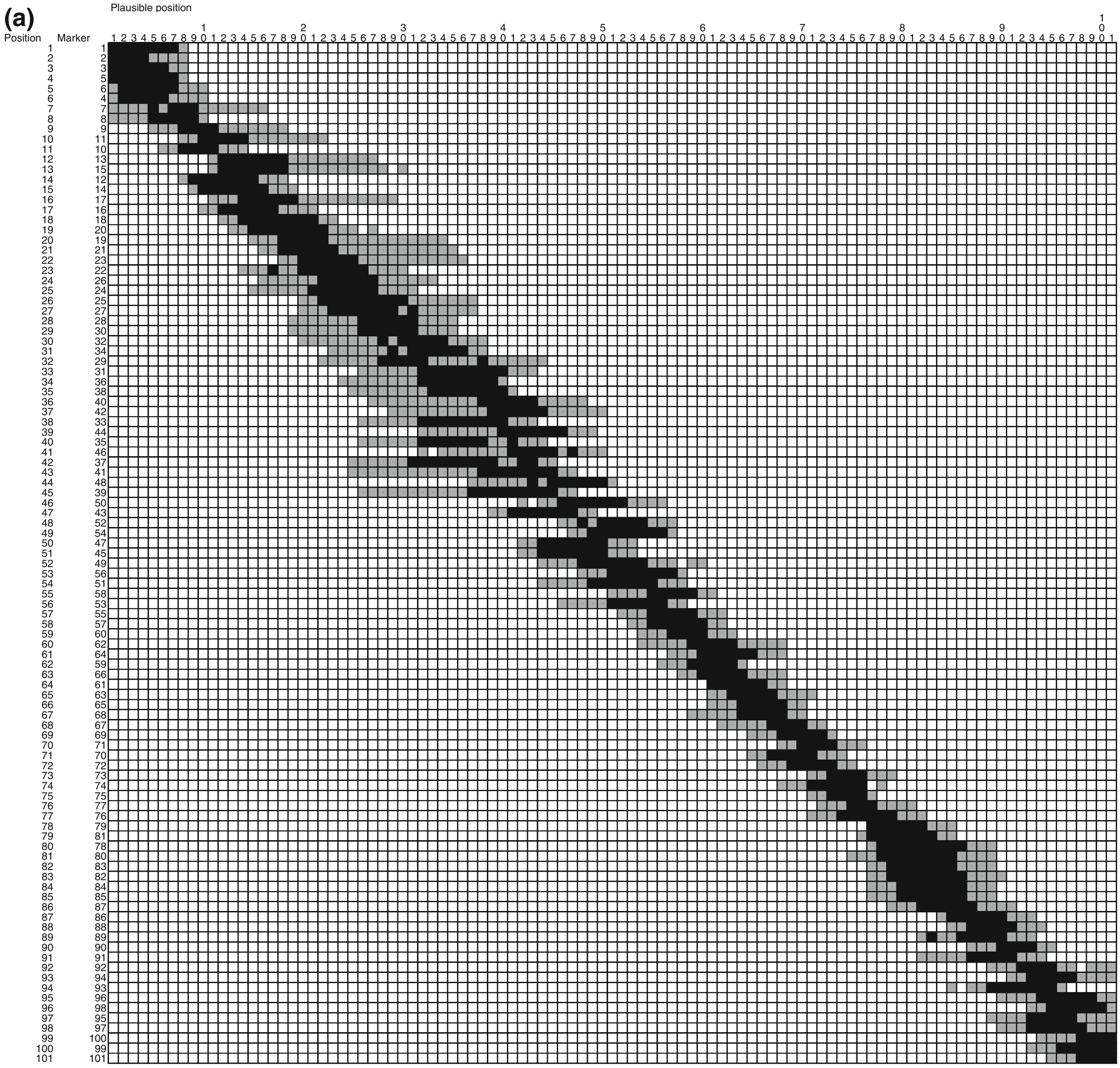

Fig. 5 Plausible positions for the 101 dominant markers (sample size = 1,000; white = no occurrences; grey = number of occurrences between 1 and 50; black = number of occurrences greater than 50)

One way to proceed is to obtain two new marker data sets. In the first set (name *ACU*) all marker data are recoded into either *A* (=*A*), *C* (=*H*, *B*, *C*) or *U* (=*D*, *U*). In the second set (named *BDU*) all marker data are recoded into either *B* (=*B*), *D* (=*A*, *H*, *D*) or *U* (=*C*, *U*). Purely dominant markers will have only missing data in one of the marker data sets, *ACU* or *BDU*, and have been removed from that set. Application of the new algorithm gives optimum map orders with 68 recombinations for the first set of marker data (*ACU*; 13 markers) and optimum map orders with 105 recombinations for the second set (*BDU*; 15 markers). However, for each of the marker data sets the optimum number of recombinations may be constant, the map orders that are produced may be very different.

Table 6 shows the effect of deleting markers in either of the two data sets. Especially, markers 108

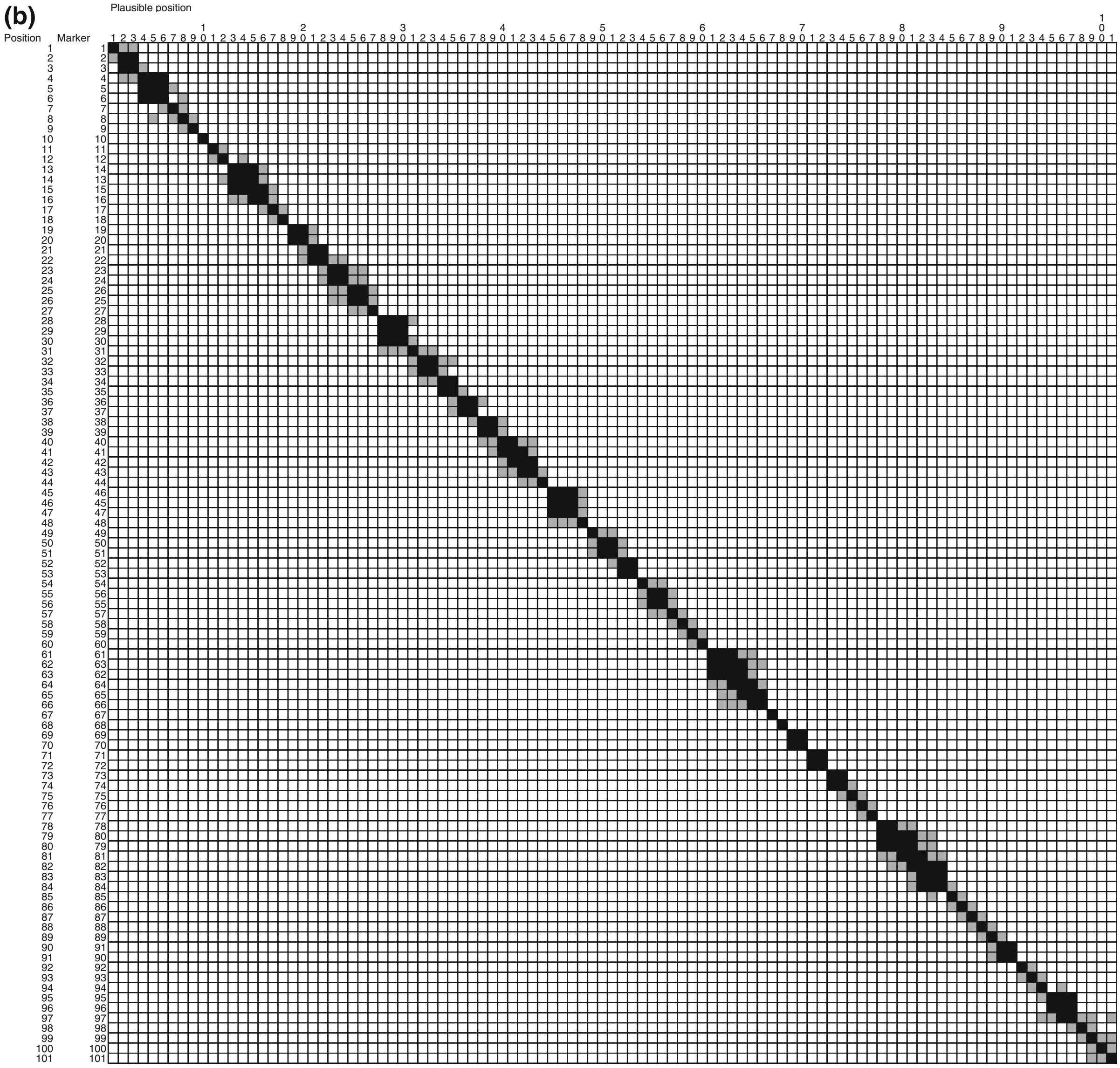

Fig. 5 continued

and 103 lead to a large reduction in number of recombinations. However, their likely position is at the ends of the map and large reductions may be expected. However, deletion of markers 44, 285, 284, 268 and 78 also leads to large reductions in number of recombinations. Marker 44 requires special attention with no reduction in marker data set *ACU* and a reduction of 8 recombinations in marker data set *BDU*.

Figure 6 shows typical graphical genotypes for marker data sets *ACU* and *BDU*. Figure 6a shows that marker 108 has an unexpected distribution of scores *A* over the individuals. It should be noticed that *A* means: no band on an electrophoretic gel. In both Fig. 6a and b, a large number of 'singletons' are present.

Discussion

This paper describes an algorithm that can be used for ordering markers with marker data obtained from an

Table 4 Summary of the *Capsicum* data

Position	Nr	Locus	*A*	*H*	*B*	*C*	*D*	*U*
1	108	P14M58_96	35			101		9
2	17	CA-MS12	14	39	35			57
3	342	E37M51_184	22	67	51			5
4	327	P11M50_380	24			112		9
5	44	P14M49_334	18	53	37		14	23
6	161	P11M47_201	25			116		4
7	81	P11M49_355	19	62	39			25
8	346	E37M51_251			41		80	24
9	80	P11M49_352	22	63	38			22
10	218	P11M48_367	25	72	44			4
11	347	E37M51_266			41		84	20
12	60	P11M51_362			43		92	10
13	285	P11M61_160	21			102		22
14	157	P11M47_185	24	68	49			4
15	336	E37M51_135			34		87	24
16	284	P11M61_159			40		82	23
17	268	P14M50_304	20	69	51		4	1
18	345	E37M51_213	21			112		12
19	78	P11M49_214			50		72	23
20	103	P14M58_55			56		88	1

F_2 population with dominant or mixed dominant/co-dominant marker scoring. The present algorithm is simple in the sense that proposals for changing the marker order and for changing the genotype configuration in the form of hidden inheritance vectors are used alternately. As an alternative, for optimising the marker order as well as for obtaining plausible maps, the cross-entropy method (Rubinstein and Kroese 2004) could be useful. Methods for constructing linkage maps by minimising the number of recombinations in hidden inheritance vectors will become available in a future version of JoinMap (van Ooijen 2006).

Marker data obtained from an F_2 population using dominant scoring contain very little information about the occurrence of recombination events. On average 75% of the data is incomplete, i.e. they refer to more than one inheritance vector. The algorithm described in this paper can be used to obtain optimum marker orders even if the markers are large in number and lie close together on the linkage map. A further advantage is that it provides a means of establishing the reliability of the map by producing a set of plausible maps. A further study of the reliability of maps based on the multipoint likelihood, e.g. in a Bayesian framework (George 2005), could be useful. The results of using the algorithm described in this paper could be a starting point for such a study. In every mapping study, the reliability of the linkage maps obtained, should get a prominent role. It follows from the current study (see Fig. 5), that whenever possible, the number of dominant marker scores should be kept to a minimum in the case of an F_2 population.

In the case some markers are scored dominantly and others co-dominantly, the algorithm described in this paper does not necessarily require a framework map of co-dominant markers as an initial step in map construction (cf. Mester et al. 2003). Approaches requiring such a framework map implicitly assume that the co-dominant markers contain no or perhaps a few errors. Our practical experience is that we cannot always rely on such assumptions.

The random nature of the algorithm provides a means to investigate the quality of the data. If alternative runs of the algorithm leads to entirely different markers with nearly the same optimum number of recombinations the conclusion must be that the marker data contain major imperfections. Inspection of the marker orders obtained and of associated reconstructions of inheritance vectors (also known as graphical genotypes) may assist in finding the source of the problems. Elimination of errors in the data is an essential part of obtaining reliable linkage maps.

In this paper it is assumed that the pure lines that form the basis of the F_2 are available. If they are not available, it is unknown whether a band comes from the grandmother or from the grandfather. This leads to the so-called phase problem. The phase problem can be solved by adding a third ‘dimension’ to the optimisation problem: minimise the number of recombinations between adjacent markers with regard to marker order, genotype configuration and the phases of the markers. This optimisation problem is discussed in Jansen (2005) for full sib families of out-breeding species.

The method described in this paper may be very suitable in situations where mixed dominant/co-dominant scoring of AFLP® is applied (Jansen et al. 2001a, b; Piepho and Koch 2000). It may also be used in a situation where some markers have been scored dominantly and other markers have been

Table 5 Optimum solutions provided by the new algorithm

Position	[a]	Mapping orders obtained by the new algorithm									
		1	2	3	4	5	6	7	8	9	10
1	108	108	108	108	108	108	108	108	108	108	108
2	17	285	17	17	284	284	284	284	284	284	103
3	342	17	342	342	285	285	347	285	285	347	78
4	327	342	327	327	347	60	285	218	60	60	345
5	44	327	44	44	60	347	60	347	347	218	268
6	161	44	81	81	218	218	218	60	218	80	336
7	81	81	80	80	80	346	80	346	346	346	157
8	346	80	346	346	81	80	81	17	80	81	161
9	80	346	347	347	44	81	44	342	81	44	60
10	218	218	218	218	327	44	327	327	44	327	285
11	347	347	284	284	342	347	342	44	327	342	284
12	60	284	285	285	17	17	17	81	342	17	347
13	285	60	60	60	161	342	161	80	17	285	218
14	157	161	161	161	346	161	346	161	161	161	346
15	336	157	157	157	157	157	157	157	157	157	80
16	284	336	336	336	336	336	336	336	336	336	81
17	268	268	268	268	268	268	268	268	268	268	44
18	345	78	345	78	78	78	345	345	345	78	327
19	78	345	78	345	345	345	78	78	78	345	342
20	103	103	103	103	103	103	103	103	103	103	17
[b]	204	187	188	188	191	191	191	191	191	193	196

[a] Mapping order obtained by JoinMap

[b] Number of recombinations

scored co-dominantly. Mester et al. (2003) propose a method for obtaining an integrated map using the co-dominant markers as anchors.

Acknowledgements Comments on earlier versions made by Martin Boer, Johan van Ooijen, Roeland Voorrips and Piet Stam of the manuscript are gratefully acknowledged.

Appendix

Relationship between maximum likelihood and minimum number of hidden recombinations

For marker order v, the likelihood function L will be denoted by

$$L = p(\mathbf{Y}|v),$$

in which $\mathbf{Y}$ represents the marker data. In addition, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N)$, in which $\mathbf{Y}_i$ represents the marker data of individual i(=1, 2,…, N). It will be assumed that the marker data of different individuals are distributed independently. As a consequence,

$$L = \prod_i p(\mathbf{Y}_i|v).$$

Suppose, the problem concerns the ordering of M markers. Let $\mathbf{H}_i = (\mathbf{h}_{i1}, \mathbf{h}_{i2}, \ldots, \mathbf{h}_{iM})$ denote a set of inheritance vectors for individual i, i.e. $\mathbf{h}_{ij}$ is either (00), (01), (10) or (11). For individual i the number of sets of inheritance vectors is equal to 2^{2M}; for all individuals the number of sets inheritance vectors is equal to 2^{2MN}.

The probability $p(\mathbf{Y}_i|v)$ may be written as

$$p(\mathbf{Y}_i|v) = \sum_{\mathbf{H}_i} p(\mathbf{Y}_i|\mathbf{H}_i)p(\mathbf{H}_i|v)$$

in which the summation takes place over all sets of inheritance vectors $\mathbf{H}_i$ The first probability $p(\mathbf{Y}_i|\mathbf{H}_i)$ is either 0 or 1:1 if the marker data $\mathbf{Y}_i$ are in agreement with the transmission of alleles given by the set of inheritance vectors $\mathbf{H}_i$, and 0, otherwise. The second probability depends on the set of inheritance vectors $\mathbf{H}_i$, marker order v and the recombination frequencies between adjacent markers for marker order v. As a consequence, $p(\mathbf{Y}_i|\mathbf{H})$ may be written as

Table 6 Reduction in numbers of recombinations due to deletion of markers

Position	[a]	[b]	
		ACU	*BDU*
1	108	24	
2	17	0	1
3	342	0	3
4	327	2	
5	44	0	8
6	161	2	
7	81	0	0
8	346		3
9	80	0	0
10	218	0	0
11	347		0
12	60		4
13	285	7	
14	157	4	0
15	336		3
16	284		12
17	268	7	3
18	345	4	
19	78		9
20	103		23

[a] Mapping order obtained by JoinMap

[b] Reduction in numbers of recombinations due to deletion of marker

$$p(\mathbf{Y}_i|v) = \sum_{\mathbf{A_i}} p(\mathbf{A}_i|v),$$

in which $\mathbf{A}_i$ represents sets of inheritance vectors $\mathbf{H}_i$ that are in agreement with marker data $\mathbf{Y}_i$.

The likelihood equations are obtained by differentiating $\ell = \ln(L)$ with respect to $\mathbf{r}_v$, the set of recombination frequencies associated with marker order v, and by setting the derivative equal to zero. The contribution of individual i to the likelihood equations can be written as

$$\frac{\partial \ln(p(\mathbf{Y}_i|v))}{\partial r_v} = \sum_{\mathbf{A}_i} p(\mathbf{A}_i|\mathbf{Y}_i, v)\frac{\mathrm{d}\ln(p(\mathbf{A}_i|v))}{\partial r_v}$$

in which $p(\mathbf{A}_i|\mathbf{Y}_i, v)$ denotes the conditional probability of the set of inheritance vectors $\mathbf{A}_i$ given the marker observations $\mathbf{Y}_i$ and marker order v. It can be shown that

$$\frac{\mathrm{d}\ln(p(\mathbf{A}_i|v))}{\partial r_{v[m][m+1]}} \propto 2r_{v[m][m+1]} - R_{[\mathbf{A}_i][m][m+1]}$$

in which, for marker order v, $r_{v[m][m+1]}$ denotes the recombination frequency between the markers at positions m and $m+1$, respectively, and $R_{[\mathbf{A}_i][m][m+1]}$ denotes the number of recombinations between the markers at positions m and $m+1$. As a consequence,

$$r_{v[m][m+1]} = \sum_i \sum_{\mathbf{A}_i} w(\mathbf{A}_i)\frac{R_{[\mathbf{A}_i][m][m+1]}}{2N}$$

($m = 1, 2,\ldots, M-1$), in which $w(\mathbf{A}_i) = p(\mathbf{A}_i|\mathbf{Y}_i, v)$. The computing scheme for estimating $r_{v[m][m+1]}$ reads as follows:

calculate for each ($\mathbf{A}_1$, $\mathbf{A}_2$,…, $\mathbf{A}_N$)

1. the number of recombinations between the markers at position m and $m+1$;
2. calculate the weights $w(\mathbf{A}_i) = p(\mathrm{A}_i|\mathbf{Y}_i, v)$, $i = 1$, 2,…, N.

Estimate the recombination frequency as the weighted sum of recombination frequencies.

For the method based on minimising the number of recombinations, the approach may be described as follows. For $i = 1$, 2,…, N, determine all $\mathbf{A}_i$ for which $\sum_{m=1}^{M-1} R_{[\mathbf{A}_i][M][M+1]}$ is a minimum, $\min(R_{[\mathbf{A}_i]})$, say. Suppose the number of sets of inheritance vectors $\mathbf{A}_i$, for which the minimum attained is K_i. Now all inheritance vectors for which the minimum is attained get equal weight, $w(\mathbf{A}_i) = 1/K_i$, and all others get weight zero, $w(\mathbf{A}_i) = 0$. In multi-point maximum likelihood, the weights would be different and depend on the current values of the recombination frequencies between adjacent markers. Finally, estimate the recombination frequency according to the above formula. The final result is equal to

$$r_{v[m][m+1]} = \sum_i \frac{1}{K_i}\sum_{\mathbf{A}_i}\frac{R_{[\mathbf{A}_i][m][m+1]}}{2N}.$$

Differences between the two methods are due to assigning different values to $w(\mathbf{A}_i)$.If all marker observations are complete, only one set of inheritance vectors $\mathbf{A}_i$ is in agreement with the marker observations. As a consequence, $w(\mathbf{A}_i) = p(\mathbf{A}_i|\mathbf{Y}_i, v) = p(\mathbf{A}_i|\mathbf{Y}_i) = 1$ for that set of inheritance vectors, and 0 otherwise. The method based on minimizing the number of recombinations assigns the same weights

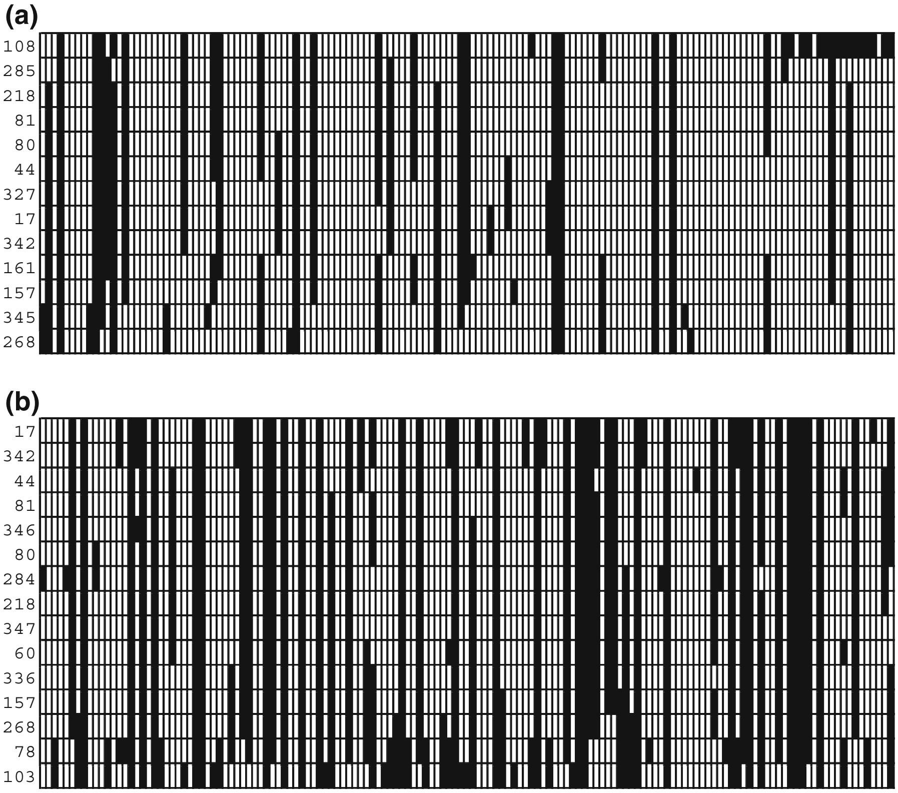

Fig. 6 Graphical genotypes for marker data sets *ACU* and *BDU*. (**a**) *ACU* (black = *A* or *U*; white = *C* or *U*). (**b**) *BDU* (black = *B* or *U*; white = *D* or *U*)

to sets of inheritance vectors $\mathbf{A}_i$. Consequently, both methods lead to same result.

In general,

$$p(\mathbf{A}_i|\mathbf{Y}_i,v)=\frac{p(\mathbf{A}_i|v)}{\sum_{\mathbf{A}_i^*}p(\mathbf{A}_i^*|v)},$$

in which the numerator may be written as

$$p(\mathbf{A}_i|v)=\prod_{m=1}^{M-1} r_{v[m][m+1]}{}^{R_{\mathbf{A}_i[m][m+1]}}\left(1-r_{v[m][m+1]}\right)^{2-R_{\mathbf{A}_i[m][m+1]}}.$$

If the recombination coefficients between adjacent markers are small and differences between recombination coefficients are not important, the following approximation can be used

$$p(\mathbf{A}_i|v)\approx r^{\sum\limits_{m=1}^{M-1} R_{\mathbf{A}_i[m][m+1]}}\leq r^{\min(R_{\mathbf{A}_i})}.$$

The maximum value is found for K_{0i} sets of inheritance vectors. The value of $p(\mathbf{A}_i|v)$decreases exponentially with the number of recombinations.

The denominator may be written as

$$\begin{aligned}\sum_{\mathbf{A}_i^*}p(\mathbf{A}_i^*|v)&=K_{0i}r^{\min(R_{\mathbf{A}_i})}+K_{1i}r^{\min(R_{\mathbf{A}_i})+1}\\&\quad+K_{2i}r^{\min(R_{\mathbf{A}_i})+2}+\cdots\\&=K_{0i}r^{\min(R_{\mathbf{A}_i})}\left(1+\frac{K_{1i}}{K_{0i}}r+\frac{K_{2i}}{K_{0i}}r^2+\cdots\right)\\&=CK_{0i}r^{\min(R_{\mathbf{A}_i})}\end{aligned}$$

As a consequence, weights assigned to sets of inheritance matrices are equal to r^x/CK_{0i}, in which x denotes the excess number of recombinations above the minimum.

References

Allard RW (1960) Principles of plant breeding. Wiley, New York

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

George AW (2005) A novel Markov chain Monte Carlo approach for constructing accurate meiotic maps. Genetics 171:791–801. doi:10.1534/genetics.105.042705

Jansen J (2005) Construction of linkage maps in full-sib families of diploid outbreeding species by minimizing the number of recombinations in hidden inheritance vectors. Genetics 170:2013–2025. doi:10.1534/genetics.105.041822

Jansen J, de Jong AG, van Ooijen JW (2001a) Constructing dense genetic linkage maps. Theor Appl Genet 102:1113–1122. doi:10.1007/s001220000489

Jansen RC, Geerlings H, van Oeveren AJ, van Schaik RC (2001b) A comment on codominant scoring of AFLP markers. Genetics 158:925–926

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680. doi:10.1126/science.220.4598.671

Knapp SJ, Holloway JL, Bridges WC, Liu B-H (1994) Mapping dominant markers using F_2 populations. Theor Appl Genet 91:74–81

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367. doi:10.1073/pnas.84.8.2363

Liu B-H (1997) Statistical genomics: linkage, mapping and QTL analysis. CRC, Boca Raton

Mester DI, Ronin YI, Hu Y, Peng J, Nevo E, Korol AB (2003) Efficient multipoint mapping: making use of dominant repulsion-phase markers. Theor Appl Genet 107:1102–1112. doi:10.1007/s00122-003-1305-1

Piepho H-P, Koch G (2000) Codominant analysis of banding data from a dominant marker system by normal mixtures. Genetics 155:1459–1468

Ridout MS, Tong S, Vowden CJ, Tobutt KR (1998) Three-point linkage analysis in crosses of allogamous plant species. Genet Res 72:111–121. doi:10.1017/S0016672398003371

Rubinstein RY, Kroese DP (2004) The cross-entropy method. A unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning. Springer, New York

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

Tan Y-D, Fu Y-X (2007) A new strategy for estimating recombination fractions between dominant markers from an F_2 population. Genetics 175:923–931. doi:10.1534/genetics.106.064030

Thompson EA (1994) Monte Carlo likelihood in genetic mapping. Stat Sci 9:355–366. doi:10.1214/ss/1177010381

van Laarhoven PJ, Aarts EHL (1987) Simulated annealing: theory and applications. Reidel, Dordrecht

van Ooijen JW (2006) JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B·V, Wageningen

Voorrips RE, Finkers R, Sanjaya L, Groenwold R (2004) QTL mapping of anthracnose (*Colletotrichum* spp.) resistance in a cross between *Capsicum annuum* and *C. chinense*. Theor Appl Genet 109:1275–1282. doi:10.1007/s00122-004-1738-1