The Plant Journal (1993) **3**(5), 739–744

TECHNICAL ADVANCE

Construction of integrated genetic linkage maps by means of a new computer package: JoinMap

Piet Stam*
Centre for Plant Breeding and Reproduction Research CPRO-DLO, PO Box 16, 6700 AA Wageningen, The Netherlands, and Department of Genetics, Wageningen Agricultural University, Dreyenlaan 2, 6703 HA Wageningen, The Netherlands

Summary

A computerized procedure to construct integrated genetic maps is presented. The computer program (JoinMap) can handle raw data from F_2s, backcrosses and recombinant inbred lines, as well as listed pairwise recombination frequencies. The procedure is useful for combining linkage data that have been collected in different experiments; the result is a mathematical alignment of the distinct genetic maps. Data from single experiments can be dealt with as well. In view of the fast growing amount of linkage information for molecular markers, which is often being generated by different research groups, integrated maps provide useful information on the map position of genes and DNA markers.

The procedure performs a sequential build-up of the map and, at each step, a numerical search for the best fitting order of markers. Weighted least squares is used for the estimation of map distances.

Introduction

Genetic maps are a useful tool in various fields of genetic research, both fundamental and applied. The recent developments in molecular genetics, by which large numbers of markers are being generated, have caused a revival of the interest in classic genetic mapping. As a result, linkage analysis and mapping have to a certain extent become computerized. Computer packages such as LINKAGE1 (Suiter *et al.*, 1983), GMENDEL (Echt *et al.*, 1992 and MAPMAKER (Lander *et al.*, 1987; further references to this package will be given as MM) are widely used. The large amount of linkage information which is becoming available for molecular markers in various organisms, has created the need to integrate linkage maps that have been obtained independently. Presently, separate RFLP maps, developed by different groups, are available for some well-documented organisms. Examples of this are the genetic maps of maize (Coe *et al.*, 1990) and *Arabidopsis thaliana* (Chang *et al.*, 1988; Nam *et al.*, 1989). The number of markers on these maps is growing fast; in addition to this, the classical maps are being extended (Koornneef, 1990). Integrated maps would greatly support the current international efforts in 'genome projects', aiming at genetical and physical mapping of complete genomes.

The key factor in the integration of distinct maps is the markers which are common to these maps. Only when a minimum number of common markers are available can distinct maps be tied together. But even if common markers are available, several problems arise. Not only may the precision of estimates of recombination frequencies vary greatly between data sets, the type of information used may also be different. For example, one map might be based on F_2 populations of varying size, while another is based on backcross data and additional information, such as observations on translocation and/or inversion heterozygotes. Yet another map might be the result of a compilation 'by hand' of data taken from the literature. The major problem then is how to weigh these types of information such that the resulting joint map is 'optimal', 'most likely', 'containing the least internal tension', or, generally, satisfying some goodness-of-fit criterion. Obviously this requires a statistical approach.

In this paper I present a computer-implemented procedure to construct integrated genetic maps. The accompanying paper (Hauge *et al.*, 1993) illustrates an application to a large data set of *A. thaliana*. Though this approach is not the final solution to the many problems in the alignment of different maps, the results show their usefulness in combining the most common types of linkage information.

Methods

Mapping procedure

Constructing a linkage map is, essentially, the finding of a linear arrangement of markers from recombination values. In developing the program JoinMap (referred to below as JM) my approach has been the following. If several estimates of the recombination frequency between a

Received 23 October 1992; revised 22 December 1992; accepted 23 December 1992.
*For correspondence at: Centre for Plant Breeding and Reproduction Research CPRO-DLO, PO Box 16, 6700 AA, Wageningen, The Netherlands (fax +31 8370 16513).

given pair of markers are available, these are, after appropriate weighting, replaced by a single one. After assigning weights to the final pairwise data that are available, a numerical search for the best fitting linear arrangement is performed.

If additional information, not contained in the pairwise estimates, is available on the ordering of certain subsets of markers (e.g. from multipoint tests of closely linked markers), this information can be used as a side condition in the searching routine. Thus, in addition to the raw data, the user may list a number of 'fixed' sequences. The program will then produce an ordering which is not contradictory to any of these 'known' sequences, unless the data contradict the forced sequences.

It will be clear that lumping estimates of recombination frequencies obtained in different experiments implies the assumption that the true recombination frequencies are the same in these experiments. Since the rate of recombination itself is known to be a heritable character, in some organisms varying with sex, genetic background and/or environmental conditions, this assumption may in some cases be an oversimplification. Violation of this assumption, however, is not likely to influence the gene order on a combined linkage map.

The map distance between two markers is defined as the mean number of recombination events, involving a given chromatid, in that region per meiosis. This is usually expressed in centimorgans (cM). A map distance of 100 cM thus corresponds to an average of one recombination event per gamete. The relation between map distance and recombination frequency (and vice versa) is expressed by a genetic mapping function (mf). Different mfs correspond to different assumed degrees of interference in crossing over. Interference is the phenomenon that recombination events in adjacent regions of a chromosome are not independent; in a classic three point test it results in a number of double recombinants which differs from what is to be expected on the basis of independence. The most commonly used mfs are those of Haldane (1919) and Kosambi (1944). The mathematical theory of mfs and how this relates to interference is treated in detail by Owen (1950) and Bailey (1961). For some theory on crossing over as a stochastic process and its implications for mfs see Stam (1979); Felsenstein (1979), from a more practical point of view, presented a family of mfs, corresponding to different levels of interference, including negative interference.

JM has two options for mfs, i.e. Haldane's and Kosambi's. Haldane's mf assumes absence of interference, whereas Kosambi's assumes positive interference (fewer double recombinants than expected without interference). The mapping procedure was developed such that an indication is obtained as to which mf the data best fit. For each map produced by JM, a goodness-of-fit criterion, corresponding to the two hypothesized levels of interference, is calculated.

Outline of the JM program

Constructing a linkage map, JM runs through the following steps.

1. Read data.
2. Calculate pairwise recombination frequencies and LOD scores for those pairs on which data are available in the whole data set.
3. Establish linkage groups; if the data encompass several linkage groups, the data file has to be split into separate ones, corresponding to the linkage groups.
4. Sequential build-up of the map.
   (a) Choose first pair of markers and calculate map distance.
   (b) Determine which marker is to be added to the map; this is done on the basis of the total LOD scores of all the markers not yet placed on the map with those that have been placed at an earlier stage.
   (c) Determine the best fitting position of the marker currently to be placed on the map, without changing the order of the ones that were placed earlier.
   (d) Perform a 'reshuffling' of all the current map positions and thus, by trial and error, find the best fitting order.
   (e) Repeat from 4(b).

The criterion used in step 4(b) implies that the marker for which the maximum amount of linkage information with the current map is available, is chosen as the one to be added to the map. The time-consuming search at step 4(d) is included because the information contributed by the marker added in step 4(c) may necessitate a revision of the order of the markers that were placed earlier. The reshuffling is performed by considering all orders within a moving window of triples of markers.

The method advocated by Wilson (1988) for determining a preliminary ordering could not be incorporated because it applies only when estimates for adjacent intervals are available, which may not be the case for many of the intervals when combining data from several sources.

Detailed description of the computations performed by JM

Data types and data files. The data that can be processed by JM are of several types, i.e. raw genotype data from (i) F_2s, (ii) backcrosses, (iii) recombinant inbred lines (RILs) and (iv) estimates of pairwise recombination percentages, the latter together with their standard errors.

A raw data set consists of coded genotypes for all markers that are segregating. The coding as well as the format of raw data files closely resembles the MM format, so that little editing is required to transform an MM data file into a JM data file and vice versa.

In order to deal with non-classic linkage experiments, such as a cross between random individuals of an outbreeding species (in which case the F_1 offspring could be segregating at multiallelic loci), JM requires for each marker an indication of the segregation type. In a raw data set of an F_2 all markers segregate according to the 'F_2' type (i.e. 1:2:1 or 3:1); however in the F_1 offspring of a random pair mating of an outbreeder one marker may segregate as in a backcross (1:1) whereas the next marker may segregate as the F_2 type. (In case of multiallelism the user may lump alleles of either parent in order to obtain a proper coding of genotypes.) Thus various segregation types may occur in a single raw data set. A single data file may comprise several raw data sets, optionally followed by a list of independent estimates, accompanied by their standard errors. A difference with MM is that several raw data sets (e.g. a number of F_2s) need not be filed as if they were a single one.

As mentioned earlier, in addition to these data, 'fixed' orders can be listed in a separate file. An order which would normally be tried by the searching routine, but which is in conflict with any of these fixed orders, is skipped. In terms of computing time, the gain from skipping outweighs the loss of testing for conflicts, and thus the use of fixed orders considerably accelerates the search.

Calculating pairwise recombination frequencies. First, recombination frequencies are calculated per population. The corresponding LOD values are also calculated. (The LOD score is indicative for the likelihood of linkage; LOD means the logarithm of odds, the 'odds' being the ratio of the probability that two loci are linked with a given recombination value over the probability that the two are not linked. A LOD value of over 3.0 means that the chances are greater than 1000:1 that the loci are linked for a given recombination estimate. LOD score decreases with increasing recombination value; it increases with increasing sample size, expressing that, e.g. an estimate of 0.3 from a sample of 40 is less informative than an estimate of 0.3 from a sample of 100. LOD values can be seen as a measure of linkage information in the data.)

Next, the estimates from distinct populations and the independent estimates (if available) are combined into a single one and the corresponding LOD values are recorded. This is done as follows. Both types of estimates (population type and independent ones) are treated as if they had been obtained from binomial samples. The hypothetical binomial sample size is calculated which would have yielded the same LOD value (population data) or the same standard error (independent data). These hypothetical sample sizes are used as weights to calculate the joint estimate and LOD value. Thus the weights assigned to distinct estimates of a certain recombination frequency correspond to the amount of information that is comprised in these estimates.

The estimates of pairwise recombination frequencies from raw data are obtained by maximum likelihood. It is worth mentioning that, contrary to most published procedures for RIL data, any intermediate generation from F_2 to F_∞ can be used. For a particular inbred generation the exact distribution of digenic genotypes for that generation is being used (rather than the usually assumed F_∞ distribution).

Establishing linkage groups. The criterion in assigning markers to linkage groups is the LOD value of the (combined) pairwise estimates. A threshold LOD value, below which linkage is not considered significant, can be set by the user. At any stage in this procedure there is a group of markers which have been assigned to a linkage group and a group of 'free' markers which have not yet been assigned. At each step the following decision is made: if none of the 'free' markers is significantly linked (by LOD value) to one of the existing groups, a new linkage group is created. Otherwise, the first 'free' marker which does show linkage with an existing group is added to that group. (A 'free' marker is linked to an existing group if it is linked (by LOD value) to at least one marker in that group.)

This procedure leads to a unique grouping of markers. The grouping of markers depends, of course, on the user-supplied critical LOD value. Higher critical LOD values will result in more and smaller linkage groups. A critical LOD of 3.0 or more will, in general, prevent incorrect assignment of markers to the same linkage group. When starting from scratch however, trying several other LOD values is not unwise; it will reveal the stability of grouping; it will also indicate which sets of markers form tight linkage groups, which markers are doubtful and which are definitely 'floating'.

Estimating map distances. The core of the program is the estimation of map distances for a given order of the markers. A slightly modified version of the procedure first described by Jensen and Jorgensen (1975) and later by Lalouel (1977) and Weeks and Lange (1987) is used. Essentially, this is a least squares procedure. The following simple example illustrates the idea. Let the order of four markers be *A*, *B*, *C*, *D*, and let the subsequent map distances be *x*, *y* and *z*. Suppose that estimates of the combination frequencies r_{AB}, r_{CD}, r_{AC}, r_{BD} and r_{AD} are available. These recombination estimates are transformed to map distances by the inverse mapping function which

is being used, yielding 'observations' of the distances x, z, $x + y$, $y + z$, and $x + y + z$, denoted as d_{AB}, d_{CD}, d_{AC}, d_{BD} and d_{AD}, respectively. Notice that in this example the distance y can only be estimated indirectly; this situation is typical for combining data from distinct sources. A measure of discrepancy between these observed distances and their expected values is the square of the difference, i.e.

$$(x - d_{AB})^2, (z - d_{CD})^2, (x + y - d_{AC})^2, (y + z - d_{BD})^2, \text{ and } (x + y + z - d_{AD})^2 \quad (1)$$

respectively. Since the d values are not equally accurate, the individual terms have to be given appropriate weights. (After investigating the effects of several possible weights, i.e. the inverse of the squared standard error and LOD, I decided to use LODs as weights.) Denoting these weights by w_{AB}, w_{CD}, etc., the total discrepancy to be minimized is

$$D = w_{AB}(x - d_{AB})^2 + w_{CD}(z - d_{CD})^2 + w_{AC}(x + y - d_{AC})^2 + w_{BD}(y + z - d_{BD})^2 + w_{AD}(x + y + z - d_{AD})^2 \quad (2)$$

Differentiating with respect to x, y and z and setting

$$\delta D/\delta x = 0, \delta D/\delta y = 0 \text{ and } \delta D/\delta z = 0 \quad (3)$$

yields a set of linear equations in x, y and z, which can be solved by standard procedures. As mentioned earlier, the search for the best fitting order essentially is a trial and error procedure. An exhaustive search of the parameter space for maps with over 50 markers soon becomes prohibitive in terms of computing time; therefore, the approach with a sequential build-up, using the most informative markers first, and a reshuffling of the map area around a newly placed marker, was chosen. As an alternative approach to finding the optimal order, I have also investigated the performance of the 'simulated annealing' (SA) algorithm, reported to be useful in combinatory optimization problems (see e.g. Laarhoven *et al.*, 1987). (SA is applied in GMendel and Surveyor; Knapp and Romero-Severson, personal communication.) However, the number of function evaluations (calculating the fitting criterion) with SA turned out to be larger than with the sequential approach. Therefore it was decided not to use SA.

JM reflects a balance between statistical rigour and computational speed; as such it bears the advantages and disadvantages of a compromise.

Example

In the accompanying paper (Hauge *et al.*, 1993) the complete linkage map of *A. thaliana* is presented. For the purpose of illustration here I present part of the results of that study. Three data sets were used, i.e. (i) the crosses examined by Meyerowitz and co-workers (Chang *et al.*, 1988), (ii) the crosses examined by Goodman and co-workers (Nam *et al.*, 1989) and (iii) a set of pairwise estimates compiled by Koornneef (Koornneef, 1990). The Goodman and Meyerowitz crosses involve multiple marker strains, such that the F_2s would, apart from the RFLP markers, segregate for a number of classical markers that had earlier been assigned to the five linkage groups of *Arabidopsis*. Figure 1 shows the three linkage maps for chromosome 2, corresponding to these separate data sets, as well as the integrated map, obtained by using

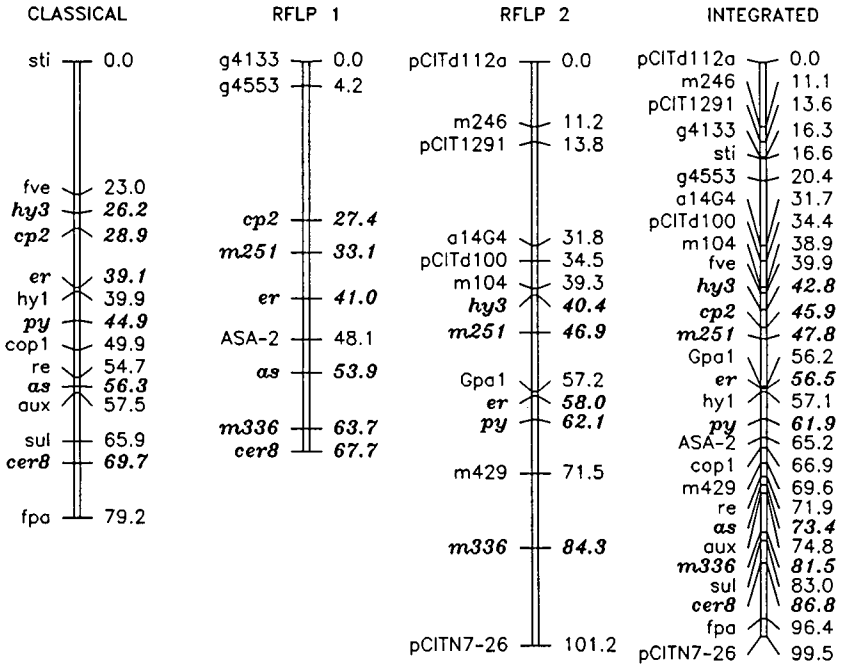

Figure 1. Three separate linkage maps and the integrated map of chromosome 2 of *A. thaliana*.
Markers in bold face occur on more than one of the separate maps. In order not to obscure the picture, only part of the markers that were actually mapped are shown. See text.

the joint data. For the sake of presentation, not all markers that are presently known to map on chromosome 2 are shown. As can be seen, the combined map aligns well with the three separate maps.

The maps shown are based on Kosambi's mf, which turned out to be the mf to which the data as a whole fit best. The maps clearly demonstrate the power of JM: instead of a rough visual alignment, we now have an integrated map which is consistent with the component maps. Any calculated map is, of course, as good as the data allow; additional information will always result in major or minor changes.

Discussion

Several computer packages are presently available for genetic linkage analysis and/or mapping.The most widely used of these are MAPMAKER (MM) (Lander *et al.*, 1987), LINKAGE (Suiter *et al.*, 1983) and GMENDEL (Echt *et al.*, 1992); the package SURVEYOR (Shoemaker *et al.*, 1992) is proprietary software; RI Plant Manager (Manly, 1992) was specially designed for mapping with recombinant inbred lines; CPROP (Letovsky, 1992) assembles genetic maps from distances and ordering constraints derived from a variety of types of information, including crosses, restriction maps and sequence data. None of these packages can use information from different types of raw segregation data, such that the results of distinct crosses can be used to assemble integrated maps.The JM package presented here is new in this respect: raw segregation data of various types, as well as 'independent' estimates of recombination are used to construct integrated linkage maps. Since several features of JM resemble those of the MM package, it is worth highlighting the main differences between the packages.

The original version of MM was to be used interactively, such that in a stepwise build-up of a map the user had to steer the search for the best order, mainly by inspection. When dealing with over 80 markers in a single linkage group, an interactive trial and error search for the best map is not an appealing job, even when the start is not from scratch. (Later versions of MM offer the opportunity to let the program automatically build-up a map (by using intermediate three point data).) JM on the contrary was designed for non-interactive use, leaving no room for the user to steer the search or to sample alternative sequences. On the other hand, large data sets, accompanied by alternative fixed sequences, are easily processed in batch jobs on a main frame computer. Although some of the data may be contradictory, JM continues its search for the best fitting map. Because there is in most cases no *a priori* criterion by which data can be discarded, all data are considered equally valuable initially. However, JM does provide a way to detect 'suspicious' markers *a posteriori*, i.e. by inspection of the chi-square value after each extension of the map. A jump in this goodness-of-fit value indicates that the marker just added to the map causes 'internal friction'. Future versions of JM will have the option of typing error detection, as described by Lincoln and Lander (1992).

Both MM and JM can be given alternative mfs, i.e. Haldane's and Kosambi's. In JM the actual calculations are done on map distances rather than recombination frequencies, and this allows the computation of an (approximate) chi-square for the goodness-of-fit of the calculated map, for either of the alternative mapping functions. MM claims to calculate likelihoods, which are the criterion in searching the best order. However, these likelihood values are not influenced by the choice of the mapping function. This is at variance with the fact that the probability of obtaining a particular multilocus genotype (and in fact the whole data set) should depend on the assumed level of interference, and thus on the mapping function chosen. MM calculations are performed with recombination frequencies for adjacent intervals, assuming independence of these intervals, i.e. assuming no interference, and only afterwards are recombination frequencies translated into map distances (Lincoln, personal communication). As long as Haldane's mf is used this is correct, but applying Kosambi's inverse mf to recombination frequencies that have been estimated under the assumption of no interference is not correct (the discrepancy although depends on the length of adjacent intervals).

The performance of JM was extensively tested with simulated data sets. These tests aimed at answers to the following questions.

What is the variation in both gene order and total map length when using data from a population of variable size?

To what extent are linkage maps influenced by using the 'wrong' mf? This question relates to the phenomenon of map extension, i.e. maps that extend when markers are added between the most telomeric ones. When the correct level of interference is assumed in mapping, such map extensions should not occur.

To what extent will misclassification of marker genotypes lead to incorrect gene orders?

It is beyond the scope of this paper to answer these questions in detail; they will be dicussed elsewhere. It can be stated, however, that the results obtained with JM were in none of the situations investigated at variance with what could be anticipated by common sense and analytical methods.

JM was written in C; four versions of the executable program are presently available: for PCs (MS-DOS), for VAX systems (VMS), for SUN workstations (UNIX) and

for Macintosh computers. For the sake of portability the source code was written in ANSI-C.

The CPU time to generate a map with JM depends, apart from hardware configuration, on the internal consistency of the data; inconsistent data (resulting in, e.g. conflicting three point orders) require a more extensive search than 'smooth' data. Generating a map with 20 markers takes 15–55 CPU-seconds on a 80486 (33 MHz) PC, depending on the 'smoothness' of the data. With 30 markers this increases to 50–200 CPU-seconds.

Some of the matrix routines, as well as the numerical procedure for finding the maximum likelihood estimate of recombination frequency for RIL data, were taken from Press *et al.* (1988). Readers who are interested in receiving JM should contact the author for further details (E-mail: (internet) P.Stam@CPRO.AGRO.NL).

Acknowledgement

I thank Johan van Ooijen for making available the DRAWMAP program to prepare the graphic maps. Elliot Meyerowitz and Maarten Koornneef carefully read the manuscript and supplied suggestions for improvement of the presentation.

References

Bailey, N.T.J. (1961) *Introduction to the Mathematical Theory of Genetic Linkage.* Oxford: Clarendon Press.

Chang, C., Bowman, J.L., DeJohn, A.W., Lander, E.S. and Meyerowitz, E.M. (1988) Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana. Proc. Natl Acad. Sci. USA*, **85**, 6856–6860.

Coe, E., Hoisington, D.A. and Neuffer, M.G. (1990) Linkage map of corn (maize) (*Zea mays* L.) (2N = 10). In *Genetic Maps*, 5th Edn. (O'Brien, S.J., ed.). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, p. 6.39.

Echt, C., Knapp, S. and Liu, B.-H. (1992) Genome mapping with non-inbred crosses using GMendel 2.0. *Maize Genet. Coop. Newslett.* **66**, 27–29.

Felsenstein, J. (1979) A mathematically tractable family of genetic mapping functions with different amounts of interference. *Genetics*, **91**, 769–775.

Haldane, J.B.S. (1919) The combination of linkage values, and the calculation of distance between linked factors. *J. Genet.* **8**, 299–309.

Hauge, B.M., Hanley, S.M., Cartinhour, S.J., *et al.* (1993) An integrated genetic/RFLP map of the *Arabidopsis thaliana* genome. *Plant J.* **3**, 745–754.

Jensen, J. and Jorgensen, J.H. (1975) The barley chromosome 5 linkage map. *Hereditas*, **80**, 5–16.

Koornneef, M., van Eden, J., Hanhart, C.J., Stam, P., Braaksma, F.J. and Feenstra, W.J. (1975) Linkage map of *Arabidopsis thaliana. J. Hered.* **74**, 265–272.

Koornneef, M. (1990) Linkage map of *Arabidopsis thaliana* (2N = 10). In *Genetic Maps*, 5th Edn. (O'Brien, S.J., ed.). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, p. 6.94.

Kosambi, D.D. (1944) The estimation of map distance from recombination values. *Ann. Eugen.* **12**, 172–175.

van Laarhoven, P.J.M. and Aarts, E.H.L. (1987) *Simulated Annealing: Theory and Applications.* Dordrecht: D. Reidel Publishing Company.

Lalouel, J.M. (1977) Linkage mapping from pair-wise recombination data. *Heredity*, **38**, 61–77.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newberg, L. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, **1**, 174–181.

Letovsky, S. (1992) CPROP: A rule-based program for constructing genetic maps. *Genomics*, **12**, 435–446.

Lincoln, S.E. and Lander, E.S. (1992) Systematic detection of errors in genetic linkage data. *Genomics*, **14**, 604–610.

Manly, K. (1992) RI Plant Manager: A microcomputer program for genetic mapping with recombinant inbred strains. *Maize Genet. Coop. Newslett.* **66**, 29.

Nam, H.G., Giraudat, J., den Boer, B., Moonan, F., Loos, W.D.B., Hauge, B.M. and Goodman, H. (1989) Restriction fragment length polymorphism linkage map of *Arabidopsis thaliana. Plant Cell*, **1**, 699–705.

Owen, A.R.G. (1950) The theory of genetical recombination. *Adv. Genet.* **3**, 117–157.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988) *Numerical Recipes in C.* Cambridge University Press.

Shoemaker, J., Zaitlin, D., Horn, J., DeMars, S., Kirschman, J. and Pitas, J. (1992) A comparison of three Agrigenetics maize RFLP linkage maps. *Maize Genet. Coop. Newslett.* **66**, 65–68.

Stam, P. (1979) Interference in genetic crossing over and chromosome mapping. *Genetics*, **92**, 573–599.

Suiter, K.A., Wendel, J.F. and Case, J.S. (1983) LINKAGE-1: a pascal computer program for the detection and analysis of genetic linkage. *J. Hered.* **74**, 203–204.

Weeks, D.E. and Lange, K. (1987) Preliminary ranking procedures for multilocus ordering. *Genomics*, **1**, 236–242.

Wilson, S. (1988) A major simplification in the preliminary ordering of linked loci. *Genet. Epidemiol.* **5**, 75–80.