Theor Appl Genet
DOI 10.1007/s00122-016-2761-8

ORIGINAL ARTICLE

A rapid marker ordering approach for high-density genetic linkage maps in experimental autotetraploid populations using multidimensional scaling

K. F. Preedy[1] · C. A. Hackett[1]

Received: 7 April 2016 / Accepted: 2 August 2016

Abstract
Key message **The paper proposes and validates a robust method for rapid construction of high-density linkage maps suitable for autotetraploid species.**
Abstract Modern genotyping techniques are producing increasingly high numbers of genetic markers that can be scored in experimental populations of plants and animals. Ordering these markers to form a reliable linkage map is computationally challenging. There is a wide literature on this topic, but most has focussed on populations derived from diploid, homozygous parents. The challenge of ordering markers in an autotetraploid population has received little attention, and there is currently no method that runs sufficiently rapidly to investigate the effects of omitting problematic markers on map order in larger datasets. Here, we have explored the use of multidimensional scaling (MDS) to order markers from a cross between autotetraploid parents, using simulated data with 74–152 markers on a linkage group and also experimental data from a potato population. We compared different functions of the recombination fraction and LOD score to form the MDS stress function and found that an LOD^2 weighting generally performed well, including when missing values and genotyping errors are present. We conclude that an initial analysis using unconstrained MDS gives a rapid method to detect and remove problematic markers, and that a subsequent analysis using either constrained MDS or principal curve analysis gives reliable marker orders. The latter approach is also particularly rapid, taking less than 10 s on a set of 258 markers compared to 6 days for the JoinMap software. This MDS approach could also be applied to experimental populations of diploid species.

Communicated by H. J. van Eck.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-016-2761-8) contains supplementary material, which is available to authorized users.

✉ K. F. Preedy
katharine.preedy@bioss.ac.uk

[1] Biomathematics and Statistics Scotland, Invergowrie, Dundee DD2 5DA, UK

Introduction

Genetic linkage maps are a vital tool for locating genes responsible for observable traits in plants and animals. The initial steps in constructing a linkage map for an experimental population are clustering the markers into linkage groups corresponding to the chromosomes, and calculating recombination fractions and the strength of the evidence for linkage, measured as the LOD score (logarithm_{10} of the odds ratio) between all pairs of markers within a linkage group. However, ordering the genetic markers into a linkage map is the most computationally demanding part of linkage mapping and this challenge is increasing as modern sequencing technologies produce larger data sets. For a linkage group of m markers, the number of possible orders is $m!/2$ and it is not possible to compare all orders. This ordering problem is a variant on the familiar 'travelling salesman' problem, and there is a large literature on different search methods. There are two aspects to the ordering problem: the choice of a criterion to optimise and the choice of an optimisation algorithm.

In a linkage mapping context, most optimisation criteria are functions of the recombination fraction between pairs of markers and the associated LOD score. An exception is the RECORD software (Van Os et al. 2005), which analyses

Published online: 09 August 2016

the marker genotypes directly to minimise the total number of recombination events: this approach has been shown to order large numbers of markers rapidly and accurately, but is implemented only for crosses from homozygous diploid parents or separate maternal and paternal data from a full sib population from heterozygous parents (a 'pseudo-double-testcross' or CP population, Grattapaglia and Sederoff 1994). Various criteria based on the recombination fraction have been proposed, such as the maximum likelihood of an order (ML), the maximum sum of adjacent LOD scores (SALOD), the minimum sum of adjacent recombination fractions (SARF) or the weighted least squares criterion (WLS). Details of these can be found in Liu (1998) or Van Ooijen and Jansen (2013). For some population types, such as a doubled haploid or backcross population from the F1 generation of homozygous parents, the LOD score is a monotonic function of the recombination fraction and so criteria such as SARF and SALOD are equivalent. In other population types, such as a CP population, some configurations of marker pairs can have the same recombination fraction, but quite different LOD scores. Maliepaard et al. (1997) has discussed this in detail for dominant and codominant markers in diploid CP populations.

The high number of possible configurations of marker pairs, and the variation in precision of estimates of their recombination fraction, is a particular challenge in autotetraploid crops such as potato. Hackett et al. (2013) identified 67 possible configurations for a pair of SNP markers with different dosages and phases in an autotetraploid full sib population and examined the LOD scores for simulated pairs with an expected recombination fraction of 0.1 in a population of size 200. They found that for some configurations linkage would not be detected, with an LOD score as low as 0.6, while for the most informative configurations the LOD score was as high as 43.1. The weighted least squares criterion is a particularly good choice for ordering markers in complex situations such as these, as it takes into account all recombination fractions and LOD scores, rather than just adjacent ones in each ordering. This is useful if nearby markers are in a configuration with a low LOD score and results in this criterion being robust to errors in data (Shields et al. 1991).

There is a wide choice of optimisation algorithms. The older approaches have been reviewed by Hackett et al. (2003) and some newer ones by Van Ooijen and Jansen (2013). Wu et al. (2008) compared several approaches using simulated doubled haploid data and found that their MSTMAP approach, based on a minimum spanning tree algorithm, gave a more accurate marker order than RECORD, while both were faster and better for noisy data than JoinMap's ML algorithm software (Stam and Van Ooijen 1995; Van Ooijen 2006) and CarthaGène (Schiex and Gaspin 1997). Lep-MAP (Rastas et al. 2013) performed better than MSTMAP in large (10,000 SNPs) simulated F2 backcross datasets. However, of these programs, only JoinMap can analyse markers from a diploid CP population without separating the data into male and female maps. Neither JoinMap's ML algorithm nor any of the programs, MSTMAP, RECORD, CarthaGène or Lep-MAP, will calculate recombination fractions and LOD scores correctly for all marker types in an autotetraploid cross.

The JoinMap software has an alternative ordering algorithm to the ML approach, regression mapping, which uses a stepwise approach to optimise the WLS criterion. This can use a pairwise dataset of recombination fractions and LOD scores calculated by a separate program, and so can order markers from any cross, including autotetraploids or higher polyploids. Hackett et al. (2003) used a different optimisation algorithm based on simulated annealing to optimise the same WLS criterion and to order AFLP and SSR markers in an autotetraploid potato population. The simulated annealing search was implemented in their TetraploidMap software (Hackett et al. 2007). Both JoinMap and simulated annealing approaches give a rapid and reliable ordering for a small data set, up to 50 markers, but become unfeasibly slow if there are 200–300 markers on a linkage group, which now occurs regularly with SNP technologies. Hackett et al. (2013) developed a methodology for calculating recombination fractions and LOD scores for data on SNP dosage in autotetraploid potato and ordered the SNPs using JoinMap's regression mapping approach on this pairwise data. Two rounds of the JoinMap algorithm were run, and only markers below a specified threshold for the WLS criterion were retained for their final map. This procedure took 2–3 h for each potato chromosome and meant that some markers that showed an association with the mapped markers were dropped. To place all the markers onto a map requires a third round of the JoinMap algorithm and this can sometimes extend the time from hours to days.

The motivation of the current study is to develop a method for ordering large numbers of markers that can analyse data from autotetraploid species, with particular emphasis on SNPs for which dosage information is available, such as that from the Infinium 8300 potato SNP array (Felcher et al. 2012). It is important that the method is rapid enough to investigate the effect of omitting different sets of markers that appear to be a poor fit to the map, or have other possible errors. We consider an ordering approach using metric multidimensional scaling (MDS) to optimise a criterion similar to the WLS criterion, usually called the stress criterion. This is one method from the general category of ordination techniques, which represent higher-dimensional configurations of points, described by a two-dimensional pairwise distance matrix, in a lower number of dimensions. Ordination techniques have been used occasionally in genetic mapping studies, but not widely.

Lalouel (1977) considered an ordination approach using non-metric MDS, with a criterion based on the ranking of recombination fractions, and applied this to order 25 loci. However, for a larger marker set where many marker pairs are likely to have similarly sized estimates of the recombination fraction but varying precision, ordering markers based on ranking the recombination fractions is less appropriate than using the actual estimates of recombination fractions. A metric MDS approach was used by Newell et al. (1995) and tested on simulated data with up to 50 loci. However, their approach is fundamentally different in that it seeks to reduce the number of pairwise intermarker distances to a subset of the highest quality and then combine short segments into a map, whereas our method utilises all available information, weighting the points according to the precision of the distance estimates and fitting a map to all the data simultaneously. Their DGMAP software is no longer available (Cheema and Dicks 2009). The THREaD Mapper program (Cheema and Dicks 2009; Cheema et al. 2010) is similar to DGMAP, but only analyses data from crosses from homozygous, diploid parents. Their study on the performance of this program and the technical details of its algorithm has not yet been published (Dicks, personal communication).

In the present study, we have tested MDS for ordering larger numbers of markers in an autotetraploid cross. We used either the recombination fraction or Haldane's map distance (Haldane 1919) as a measure of pairwise intermarker distance and considered different weighting schemes based on the LOD scores to construct the stress criterion. Three different ordering methods were investigated: constrained MDS, two-dimensional unconstrained MDS followed by principal curves (PC2) and three-dimensional unconstrained MDS followed by principal curves (PC3). We have applied this to simulated data with 74–152 markers on a linkage group, including missing values and genotyping errors, and to an experimental dataset of 277 SNPs from linkage group I of an autotetraploid potato population.

Methods

The initial step of a linkage mapping study is to cluster the molecular markers into linkage groups. As we have focused on autotetraploid species here, and on markers for which allele dosages are available, the four homologous chromosomes from each parent are connected by duplex, double-simplex and higher-dosage markers into a single linkage group. The 'distance' between each pair of markers was calculated as the significance of a Chi-square test for independent segregation, and the markers were clustered using group average clustering. More details of the clustering for autotetraploids are given in Luo et al. (2001) and Hackett et al. (2013). The recombination fractions and LOD scores were then estimated between all pairs of markers within a linkage group. As there are at least two possible phases for each pair, and often more, the recombination fraction was estimated for all possible phases and the most likely phase was taken as the one with the highest likelihood among the phases with recombination fraction ≤ 0.5. The estimate of the recombination fraction and the LOD score for the most likely phase were then used for ordering. Again, full details can be found in Hackett et al. (2013). The recombination fractions among the m markers in a linkage group were then converted to an $m \times m$ matrix of pairwise map distances using a map function. There are many possible map functions (reviewed by Zhao and Speed 1996) and we used the simplest, the Haldane map function (Haldane 1919). We also considered the recombination fraction as a measure of distance. Other map functions could equally well be used here instead. We have focused here on the step of ordering the markers to form a linkage map and estimating the distances between marker positions on the map. We use the terminology of Stam (1993) and refer to the estimated distances between adjacent markers along the map as direct map distances. In outline, our approach is to analyse the $m \times m$ matrix of pairwise map distances between molecular markers using multidimensional scaling (MDS), weighted by a function of the LOD scores, in the expectation that the MDS configuration in two or three dimensions will reveal the linear structure of the chromosome. It might be expected that a one-dimensional solution would show the structure. However for chromosomes of a realistic length, the markers near opposite ends are inherited almost independently of each other and therefore the recombination fractions between all such pairs are close to 0.5. This generally results in a curved plot, similarly to the 'horseshoe effect' in correspondence analysis. The unconstrained solutions in two and three dimensions were therefore examined to identify and remove any obviously outlying points. To obtain a linkage map, the resulting configuration then has to be mapped onto a line. This can be done by fitting a principal curve (PC) to either the two- or three-dimensional MDS or by constraining the MDS solution to lie on a circle (spherically constrained MDS in two dimensions). The approach was tested on simulated data based on a full sib population from autotetraploid parents, simulating data from the potato genetic maps published by Hackett et al. (2013). We also studied in detail the performance of the PC methods on real data from chromosome I of the potato genome.

Multidimensional scaling

Multidimensional scaling (MDS) refers to a class of ordination techniques designed to display 'distances' among

points in geometrical space. It is generally used to reduce data from many dimensions, m, to fewer, possibly more comprehensible dimensions, n. If there are m observations, then MDS techniques use an $m \times m$ matrix of observed distances (or dissimilarities) between points and the desired number of dimensions, $n < m$, is specified. A configuration of points in n-dimensional space is sought that best preserves the observed distances between points by minimising a loss function L. For a given configuration X, the loss function $L(X)$ is a function of the difference between the observed distances in the m-dimensional configuration (which may be formed using any metric) and the Euclidean distances between points in the n-dimensional configuration

$$L(X) = \sum_{i=1}^{m} ||w_i d_i \cdot \widehat{d}_i(X)||,$$

where $||\cdot||$ is any metric function (i.e. it satisfies the intuitive properties of a distance such as non-negativity, symmetry and the triangle inequality $||x \cdot y|| + ||y \cdot z \cdot || \geq ||x \cdot z||$), d_i is the m-dimensional vector of the observed distances between point i and the other points, w_i is a vector of weights associated with point i and $\widehat{d}_i(X)$ is the m-dimensional vector of distances between point i and the other points in configuration X. In its simplest form, classical multidimensional scaling is also known as principal coordinates analysis and, though the distance matrix may be calculated in a variety of ways, the metric is always Euclidean, $||d_i \cdot \widehat{d}_i|| = \sqrt{\sum_j (d_{ij} - \widehat{d}_{ij})^2}$, and the weights are always equal to one. If the distance matrix is Euclidean, then this is equivalent to the principal components analysis and the function to be minimised reduces to $\sqrt{\sum_{ij} (d_{ij}^2 - \widehat{d}_{ij}^2)}$.

Metric multidimensional scaling (or weighted metric multidimensional scaling) generalises classical multidimensional scaling to allow for different metrics (and weights) and a commonly used loss function in this context is stress, defined as

$$\sigma(X) = \sum_{i<j<m} w_{ij} \left(d_{ij} - \widehat{d}_{ij}(X) \right)^2.$$

There are many ways of minimising $\sigma(X)$ and we used a common method, the stress minimisation by majorisation approach implemented in the smacof R package (de Leeuw and Mair 2009). This minimises $\sigma(X)$ iteratively by minimising at each step a simple function that bounds σ from above, called the majorising function. The method is described in detail in de Leeuw and Mair (2009).

The analysis described above is an unconstrained MDS. It is also possible to constrain the final configuration of points to lie on a circle by imposing a penalty for deviations from that circle, in a constrained MDS. This is done by defining a new point in the centre of the data and constraining all points to be equidistant from it. The variation in distance from the centre point is added to the stress function.

Principal curves

We used the method of principal curves (PC) to map the final MDS configuration of points onto a curve. The same method applies no matter what the dimensionality of the MDS configuration. Formally, principal curves were defined by Hastie and Stuetzle (1989) as self-consistent smooth one-dimensional curves that pass through the middle of a p-dimensional data set providing a nonlinear summary of the data. (In this context, the projection of a data point onto a curve is the closest point on that curve, and for a curve to be self-consistent, any set of data points that project onto the same point, z, on the curve must have point z as their mean.) Fitting a PC is an iterative two-stage process. A summary straight line, such as a principal component, is fitted. Then this summary line is transformed to a smooth curve, using splines, to achieve self-consistency.

Since splines depend on the smoothing constraint, PCs are not unique. We used the algorithm implemented in the R package princurve (Hastie and Weingessel 2013) which uses the first principal component as the initial summary of the data, cubic splines for fitting smooth curves and local averaging to determine self-consistency. The smoothing constraint can be selected by an explicit option or determined automatically by leave-one-out cross-validation.

Algorithm for analysis

The recombination fractions and LOD scores were calculated as described above for each population, and, where map distances were used, the recombination fractions were converted to map distances using Haldane's mapping function.

All pairs of resulting map distances were first analysed using unconstrained MDS implemented in the smacof R package (de Leeuw and Mair 2009) i.e. minimising a stress function. Each pairwise intermarker distance was weighted. Four different combinations of distance and weighting were used in the stress function—the map distance weighted by (1) the LOD score, (2) the square of the LOD score or (3) the square root of the LOD score ($LOD^{0.5}$) and (4) the recombination fraction weighted by the LOD score. Two different approaches were used to obtain the final linkage map of marker positions and direct map distances from the MDS configuration:

(i) Principal curves (PC2 and PC3): the first principal curve was calculated using the princurve R package (Hastie and Weingessel 2013), with the smoothing constraint selected by leave-one-out cross-validation. The projections of markers onto this curve give the marker positions and the direct map distances between them. In PC2, the curve was fitted to a two-dimensional MDS, and in PC3, it was fitted to a three-dimensional MDS.

(ii) Constrained MDS (cMDS): two-dimensional spherically constrained MDS was applied to the data with a penalty for markers deviating from the circle. The penalty was selected to ensure that the stress from the constrained MDS was no more than 10 % greater than that from the unconstrained MDS. The projections onto the circle give the marker positions and the direct map distances between them. The cap of a 10 % increase in stress is somewhat arbitrary, but was chosen by inspection of the simulations as effective in ensuring that key information is retained from the unconstrained configuration whilst allowing projection onto the circle.

A detailed algorithm for the fitting process is provided in the appendix. All the R packages were run in R 3.0.3 (R Core Team 2014). Our R code has been included in the Supplementary Information.

Diagnostics for problem markers

It is common for linkage mapping studies to contain problem markers. Some markers can be generally difficult to score, leading to many genotyping errors, possibly distorted segregation ratios, or many missing values. Such markers need to be identified and eliminated before the final ordering. Other markers will have a lower level of genotyping error, which may be harder to detect, and it is important that the ordering approach is as robust as possible against these. The first diagnostic of problem markers is to inspect the MDS configuration (in the case of PC3, the configuration can be considered in both two and three dimensions), to find clearly outlying points. This is a subjective step and it is usually necessary to carry it out more than once, removing outliers and then recalculating the configuration. Another diagnostic of overall fit from this plot is to superimpose the configurations from the unconstrained MDS and the constrained MDS to see whether there has been a noticeable change in the rank order of markers, in either dimension, particularly if this moves markers from the centre of the configuration to either end of the arc. If this has occurred, it may indicate the inversion of a section of the map. An example is shown in Supplementary Figure 1.

A second diagnostic tool is the nearest neighbour measure (NNfit), derived from the matrix of distances. This is a measure given for each marker and is the sum of the absolute difference between the observed and estimated distance between that marker and the nearest informative neighbours on either side—that is, the nearest neighbours with a non-zero LOD score. (Neighbouring markers where different parents are heterozygous are uninformative about recombination.) For some markers near the ends of the chromosome, there will be a neighbour on only one side. High values of the criterion can be used to identify possible outliers, while the mean NNfit provides a measure of the fit to the original data. It can be used to compare models using the same distance metric (in our case, Haldane map distance or recombination fraction).

A third diagnostic is obtained by examining the ordered marker data, returning to the original genotype scores. Hackett et al. (2013) describe how a hidden Markov model (HMM) can be used to reconstruct the chromosomal states underlying each offspring's genotype scores, using the inferred parental phases for the ordered markers, and this is used as part of their methodology for QTL mapping. In the present study, the HMM was run for each offspring to identify the chromosome configuration most likely to give the observed genotypes. This gave an $m \times o$ matrix of recombination locations, where o is the number of offspring. From this matrix, we then calculated how the total number of recombinations across all offspring was affected by (1) excluding each marker in turn, (2) swapping each marker with the adjacent marker and (3) trying all other possible orderings of the surrounding triplet of markers. (This approach was motivated by the RECORD software (Van Os et al. 2005), and assumes that a badly scored or misplaced marker will have an unusually high number of recombinations in its vicinity in the chromosomal configuration, and that the order can be improved by removing it or by a local swap in ordering.) Orders were compared using HMM_mean, equal to the mean over all offspring of the recombinations that can be removed by excluding that marker.

Finally, a diagnostic for marker genotypes derived from quantitative measurements, such as the theta scores from the Infinium 8300 potato SNP array (Felcher et al. 2012), is to map the quantitative measurements using QTL interval mapping. This approach is discussed in more detail and used by Hackett et al. (2013). It was not appropriate for the simulated data sets here, as they were simulated as 0–4 dosages, but it was applied to the analysis of the experimental data.

Simulation of autotetraploid data

Simulation 1: problem-free simulated data

The autotetraploid simulations were motivated by our work on analysing SNP dosage data in autotetraploid potato and

Table 1 Median rank correlations of the ordering of simulation set 1 with the true order, for chromosomes I–III and for four stress criteria based on Haldane's map distance with LOD, LOD^2 and $LOD^{0.5}$ weightings and recombination fraction with LOD weighting

Weight	Method	Chromosome (no. markers)		
		I (142)	II (120)	III (74)
LOD	cMDS	0.995	0.996	0.991
	PC2	0.995	0.996	0.987
	PC3	0.996	0.996	0.988
LOD^2	cMDS	0.997	0.997	0.995
	PC2	0.997	0.997	0.996
	PC3	0.997	0.997	0.996
$LOD^{0.5}$	cMDS	0.993	0.993	0.979
	PC2	0.994	0.993	0.972
	PC3	0.994	0.993	0.979
Recombination fraction	cMDS	0.995	0.997	0.807
	PC2	0.996	0.997	0.991
	PC3	0.996	0.997	0.992

cMDS is the constrained MDS method. PC2 and PC3 are from the principal curves method applied to two- and three-dimensional unconstrained MDS, respectively

the simulations were based on the maps and parental genotypes for the 12 potato chromosomes published by Hackett et al. (2013). Table 1 and Supplementary Table S1 give details of the number of markers in each case. Ten populations of 200 offspring were simulated in Fortran, and the recombination fractions and LOD scores were calculated using the theory and Fortran program described in detail in Hackett et al. (2013).

Simulation 2: Data with 20 % missing values

Simulation set 2 introduced missing values, generating these randomly and independently for each offspring and genotype with probability 0.2. These were introduced to the ten original simulations for chromosomes I, II and III, choosing these as having a good range of lengths and marker densities.

Simulation 3: data with 5 % random errors

Simulation set 3 introduced errors into the original simulations, selecting the combination of offspring and genotype to be changed randomly and independently with probability 0.05. The genotype of the selected offspring, expressed as its dosage, was changed randomly to a different dosage that was compatible with the given parental dosages under the assumption of random chromosomal selection. (For example, if the parental configuration was a simplex marker AAAB × AAAA, with possible offspring dosages of 0 or 1 B alleles, an offspring with genotype 1 would be changed to 0, but not to 2, which is obtainable only by formation of a multivalent and double reduction, or to 3 or 4, which are not possible from this parental configuration. We assume that errors leading to such inconsistencies would have been detected at an earlier stage of the analysis.) As with the missing values, these random errors were introduced into the ten simulations for chromosomes I, II and III.

Analysis of experimental autotetraploid data

The potato maps of Hackett et al. (2013) were derived using the regression mapping approach of JoinMap 4, and recombination fractions and LOD scores were calculated using the theory in that paper. Due to the slowness of this approach with large datasets, only two rounds of the JoinMap algorithm were run to obtain those maps, and consequently some SNPs that showed an association with the mapped SNPs were not included in the final maps. On chromosome I, 277 non-identical SNPs clustered together initially, but only 142 were placed after two JoinMap mapping rounds. The full set of 277 SNPs were reanalysed here using the MDS approach to see how many extra SNPs could be mapped. The theta scores for these 277 SNPs were then mapped as quantitative traits as a check on their dosage, phase and position, and the distribution of recombinations was explored by fitting an HMM to each offspring's dosages. Linkage maps were drawn using the MapChart 2.2 software (Voorrips 2002).

Results

Various measures were used to assess all the simulations: map length, Spearman's rank correlation with the true order and mean NNfit. The criterion of total swaps from the true order was also investigated, but did not show anything in addition to the first three criteria and so is not discussed below. The process of calculating the HMM_mean (of recombinations that can be excluded by dropping a marker, based on the HMM reconstruction) was harder to automate and so was run for a single simulation of each scenario rather than for all ten simulations, and using the maps from the Haldane map distance with LOD and LOD^2 weighting, estimated by the PC2 and PC3 methods. The first three measures are shown in the text for chromosomes I–III, which we consider in detail, and in Supplementary information for chromosomes IV–XII. The HMM_means are also in the Supplementary information. All three methods ran be used rapidly—of the order of seconds rather

than minutes, but both PC2 and PC3 were of the order of ten times faster than cMDS.

Simulation 1: problem-free simulated data

Figure 1 shows an example of the unconstrained two-dimensional MDS configuration (solid circles) and the projections onto the fitted principal curve (triangles and dashed line) for one of the simulations of chromosome I, using the Haldane map distance and LOD^2 weighting. This arc shape was typical for these configurations. The Spearman rank correlations with the true order for each simulated chromosome are summarised in Table 1 and Supplementary Table S1. All three methods—cMDS, PC2 and PC3—had very high correlations when Haldane map distances and LOD or LOD^2 weights were used, with LOD^2 yielding slightly better estimates of order, while the $LOD^{0.5}$ weighting generally had lower correlations. When the recombination fraction was used as a measure of distance, the correlations were more variable, with some simulations using PC2 and PC3 having as high correlations as the LOD^2 weighting, but the cMDS method having lower mean correlations for some chromosomes, especially LG III. This is because in the two of the ten simulations of LGIII, the unconstrained MDS yielded an S-shaped configuration which had no natural mapping onto an arc. The PC method estimated order well in this case, but the constrained MDS method yielded a poor fit. However, this is visible from the diagnostics and Supplementary Figure 1 gives an example of how to recognise a poor fit. Either relaxing the cMDS penalty or using the PC method gives a better order here. LG III had the fewest markers of the simulated sets and had a central region with larger intermarker distances and neighbouring markers in configurations with low LOD scores, which probably caused the difficulty.

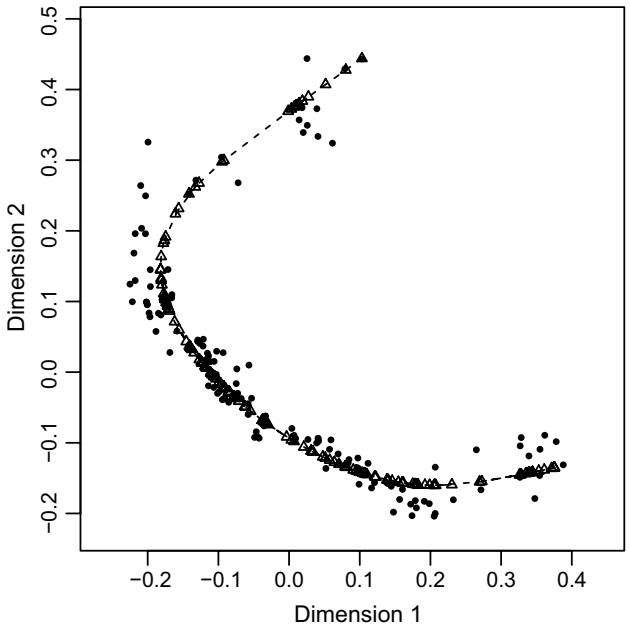

Fig. 1 Final configuration of the unconstrained two-dimensional MDS (*solid circles*) and the projections onto the fitted principal curve (*open triangles* and *dashed line*) using LOD^2 weights for an example from simulation 1—problem-free data simulated from potato chromosome I

For the estimates of length (Fig. 2 and Supplementary Table S2), the LOD or LOD^2 weighting with cMDS yielded better estimates of total chromosome length than PC2 and PC3. However, all methods tended to underestimate the lengths of the chromosomes. The LOD^2 weighting led to consistently greater underestimates of length. When using the cMDS method with $LOD^{0.5}$ weights, the cMDS method tended to overestimate chromosome length slightly. This was less of a problem with the PC methods, but there was still a tendency towards length inflation. For all methods with the recombination fraction as a distance measure, the chromosome length was underestimated to a greater extent than using either LOD or LOD^2-weighted Haldane map distance.

For the mean NNfit (Table 2 and Supplementary Table S3), the LOD^2 weighting performed better than the LOD weighting overall, which in turn was better than the $LOD^{0.5}$ weighting. For some chromosomes, the differences among the three methods were small, while for the others the cMDS method gave a worse fit than PC2 and PC3. The NNfits for the recombination fraction are not comparable with those from the Haldane distances, but again PC2 and PC3 were better than cMDS.

Overall, the $LOD^{0.5}$ weighting did not perform as well as the other weightings for any of the criteria. Using the recombination fraction rather than the map distance performed well in some cases, but underestimated the chromosome length substantially and was worse than the LOD^2 weighting with regard to the rank correlation with the true order for some chromosomes. We therefore focused on LOD and LOD^2 weights and Haldane map distances for the other simulations 2 and 3, the HMM reconstructions and the experimental data.

Simulation 2: data with 20 % missing values

With 20 % missing data, the median rank correlation remained high, although means (not shown) were occasionally lower. The main occurrence of this was for one simulation of chromosome I where the configuration was S-shaped (similar to the example in Supplementary Figure 1) and the cMDS method failed. Either the differences between methods were small, or LOD^2 weights gave better estimates of order (Table 3), and the approaches using PC2 or PC3 were better than cMDS. Both LOD and LOD^2 weights tended to

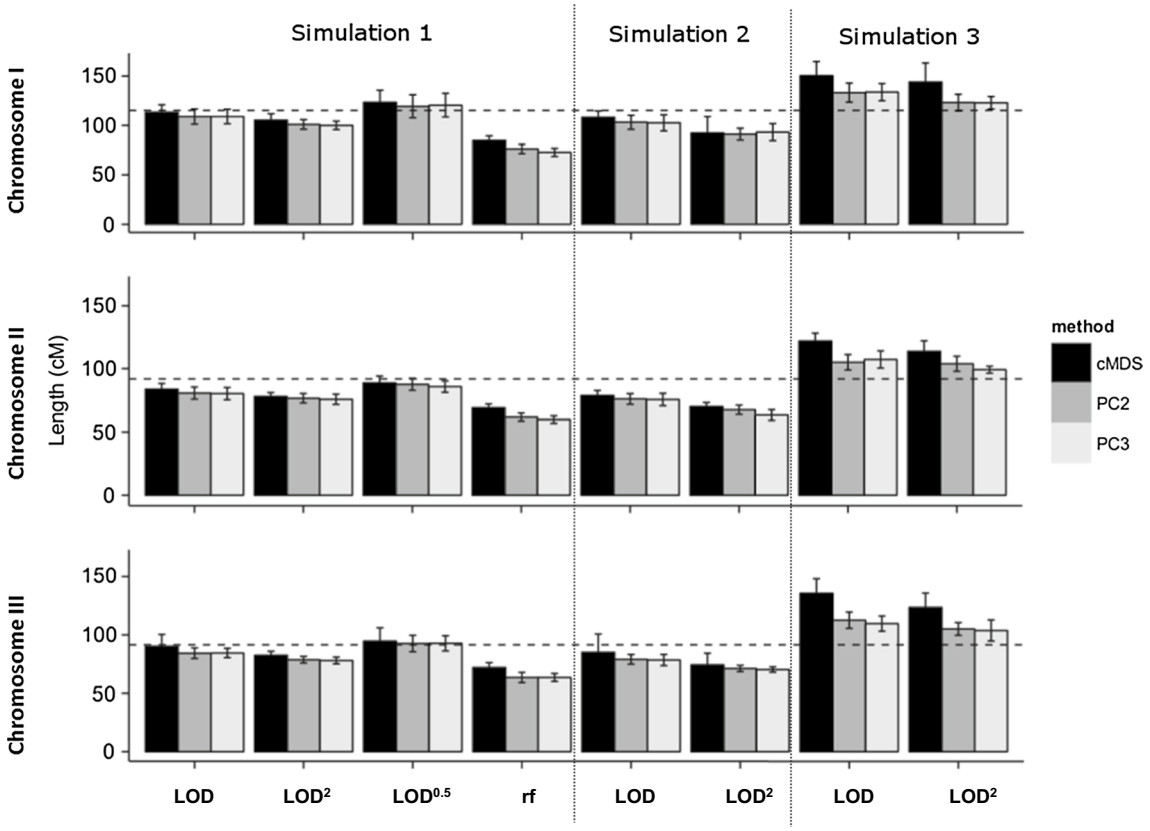

Fig. 2 Mean and standard deviation of lengths from chromosomes I, II and III of simulation set 1 with LOD, LOD^2 and $LOD^{0.5}$ weightings and recombination fraction (rf) with LOD weighting and from simulation sets 2 and 3 with LOD and LOD^2 weighting using cMDS, PC2 and PC3 methods. The *dashed line* is the true length of the chromosome, the height of the *bars* is the mean estimated length and the length of the *error bar* is twice the standard deviation of the estimated lengths over the ten simulated populations of each chromosome

underestimate the length of the chromosome with LOD^2 weights giving greater underestimates (Fig. 2). cMDS with LOD weights consistently gave the best estimates of length, but the PC methods tended to give the best estimates of marker order—for chromosome I PC2 was better whilst for chromosome III PC3 was better. As in the case of the problem-free data, LOD weights gave better estimates of length, whilst LOD^2 weights gave better estimates of marker order. For the mean NNfit (Table 3), the best results were obtained from the LOD^2 weights and either PC2 or PC3; both of these were better than any other combination.

Simulation 3: data with 5 % random errors

With errors in the data, all methods tended to overestimate the chromosome length, but PC3 consistently gave the best estimates of length (Fig. 2) and marker ordering (Table 4). The LOD^2 weightings with PC3 gave the best estimates of length. For chromosomes I and II using the PC3 method with either LOD or LOD^2 weightings gave a better estimate of order, whilst in chromosome III PC3 with LOD^2 weights gave a better estimate. The combination of LOD^2 weights and PC3 had the lowest NNfit for each chromosome.

Recombinations in the HMM reconstructions

A, HMM was used to reconstruct the chromosomal configuration for one simulation of each scenario, using the maps from the PC2 and PC3 methods with the stress function calculated from the Haldane map distance and LOD and LOD^2 weighting. The HMM was also run with the true order for comparison. Supplementary Table 4 summarises the recombinations that were inferred. For the true order and the problem-free simulations (simulation set 1), at most four recombinations could be removed by omitting markers—a proportion of 0.033 per marker. For 10 of the 12 chromosomes, the PC3 method with LOD^2 weighting found the order for which the HMM_mean was lowest. At most, seven recombinations could be removed by omitting a single marker. The results were similar for the simulations with

Table 2 Mean and standard deviation (SD) of the NNfit measure from simulation set 1, chromosomes I–III, with LOD, LOD^2 and $LOD^{0.5}$ weightings and recombination fraction with LOD weighting

Weight	Method	Chromosome		
		I	II	III
LOD	cMDS	2.07 (0.16)	1.64 (0.07)	3.09 (0.26)
	PC2	1.99 (0.16)	1.57 (0.07)	2.97 (0.21)
	PC3	1.98 (0.17)	1.56 (0.08)	2.95 (0.24)
LOD^2	cMDS	1.94 (0.12)	1.50 (0.11)	2.95 (0.14)
	PC2	1.84 (0.08)	1.47 (0.08)	2.86 (0.12)
	PC3	1.82 (0.09)	1.46 (0.07)	2.81 (0.15)
$LOD^{0.5}$	cMDS	2.26 (0.20)	1.79 (0.10)	3.22 (0.41)
	PC2	2.18 (0.21)	1.73 (0.09)	3.16 (0.34)
	PC3	2.18 (0.22)	1.72 (0.10)	3.15 (0.33)
Recombination fraction	cMDS	1.56 (0.09)	1.35 (0.08)	2.49 (0.12)
	PC2	1.41 (0.09)	1.22 (0.07)	2.24 (0.11)
	PC3	1.34 (0.08)	1.18 (0.04)	2.22 (0.15)

cMDS is the constrained MDS method. PC2 and PC3 are from the principal curves method applied to two- and three-dimensional unconstrained MDS, respectively. SD is the standard deviation of the NNfit over the ten simulated populations of each chromosome

Table 3 Comparison of ordering methods for simulation data set 2, with 20 % missing data, for chromosomes I, II and III, by rank correlation and NNfit

Measure	Weight	Method	Chromosome		
			I	II	III
Median rank correlation	LOD	cMDS	0.994	0.994	0.970
		PC2	0.994	0.993	0.985
		PC3	0.994	0.993	0.986
	LOD^2	cMDS	0.995	0.995	0.985
		PC2	0.996	0.995	0.986
		PC3	0.995	0.995	0.990
Mean NNfit (SD)	LOD	cMDS	1.98 (0.14)	1.56 (0.06)	2.91 (0.50)
		PC2	1.71 (0.32)	1.40 (0.06)	2.62 (0.25)
		PC3	1.83 (0.15)	1.49 (0.09)	2.73 (0.29)
	LOD^2	cMDS	1.66 (0.10)	1.33 (0.05)	2.52 (0.19)
		PC2	1.66 (0.10)	1.33 (0.05)	2.52 (0.19)
		PC3	1.68 (0.12)	1.36 (0.08)	2.50 (0.20)

Abbreviations cMDS, PC2 and PC3 are as in previous tables

missing values. For the simulations with 5 % error, all values of HMM_mean were much higher, including for the true order, and it was less clear which approach was best.

Experimental data from potato chromosome I

The MDS methods were used to map 277 non-identical SNPs that were found by Hackett et al. (2013) to cluster together as chromosome I, including 142 that were placed on the final map of chromosome I in that publication. The maps from LOD and LOD^2 weighting were compared, using two- and three-dimensional principal curves. Figure 3 illustrates the first three dimensions using the LOD^2 weighting. SNP 269 is particularly prominent in the plot of dimension 2 against dimension 1, while 174, 232, 112, 18, 123 and 122 are outlying in dimension 3 against dimension 1. The plot of dimension 3 against dimension 2 is less clear, but similar SNPs are outlying. SNP 269, which was placed at the end of the inferred configuration, also had a large value for its NNfit. Two-dimensional MDS generally identified fewer outliers, but in each case SNPs 269 and 174 were well separated from the rest of the data. The SNPs identified as outliers and omitted from the ordering are summarised in Table 5a, which shows a good agreement among the methods and weighting. A review of the SNP scores showed that SNP 269 was a double-duplex (AABB × AABB) SNP with significant ($p < 0.001$) associations with some other SNPs in this group based on a Chi-square test for independent segregation, but the matrix of recombination fractions and associated LOD scores showed that SNP 269 had only a single significant linkage to the rest of the group, with an LOD score of 3.3. A re-examination of the theta scores for this SNP showed it to be difficult to genotype and that it also showed some significant associations with SNPs on LG VIII. Table 5a shows how the mean NNfit statistic is reduced by the first round of dropping outliers and is generally reduced by the second round if one is necessary.

For experimental data, mapping the theta scores as quantitative traits is a very effective means of investigating

Table 4 Comparison of ordering methods for simulation data set 3, with 5 % errors, for chromosomes I, II and III, by rank correlation and NNfit

Measure	Weight	Method	Chromosome		
			I	II	III
Median rank correlation	LOD	cMDS	0.991	0.992	0.845
		PC2	0.991	0.990	0.972
		PC3	0.993	0.992	0.967
	LOD^2	cMDS	0.769	0.986	0.894
		PC2	0.987	0.987	0.967
		PC3	0.992	0.991	0.978
Mean NNfit (SD)	LOD	cMDS	2.76 (0.24)	2.40 (0.12)	4.59 (0.40)
		PC2	2.39 (0.24)	2.10 (0.12)	4.59 (0.40)
		PC3	2.45 (0.20)	2.12 (0.16)	3.82 (0.23)
	LOD^2	cMDS	2.70 (0.39)	2.26 (0.15)	4.17 (0.46)
		PC2	2.26 (0.27)	2.05 (0.13)	4.17 (0.46)
		PC3	2.26 (0.13)	1.95 (0.09)	3.49 (0.20)

Abbreviations cMDS, PC2 and PC3 are as in previous tables

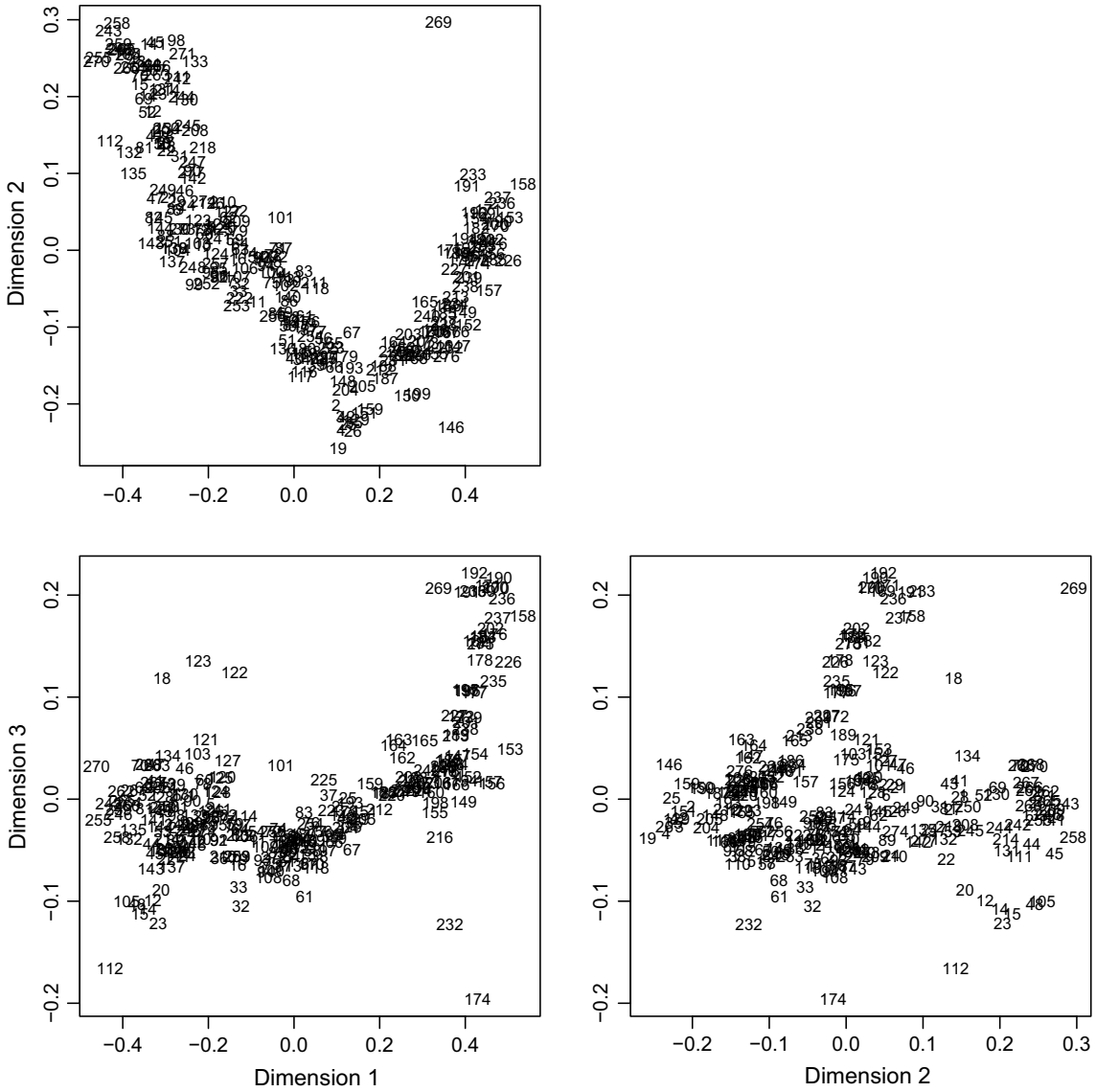

Fig. 3 The three dimensions of the MDS configuration for experimental data from chromosome I, using LOD^2 weighting

the inferred dosage and phase of the parents and whether there are any quality issues with the data. This was carried out after inference of the parental phase, using the approach described by Hackett et al. (2013). The map used was the best configuration from the LOD^2 weighting and two dimensions, excluding only SNPs 269 and 174 as the most extreme outliers. We expect that broad conclusions about the variability and phase from QTL mapping will be robust to small variations among maps. The theta scores from all 277 SNPs were mapped. The theta score for SNP 269 showed little association with this chromosome, with a maximum LOD score of 2.9, confirming that it was wrongly grouped here. (It showed an LOD of 45.5 when mapped to LGVIII.) All other theta scores showed a strong association with LGI, as expected, with the maximum of the LOD profile being at least 46 and the corresponding percentage variance explained being at least 66 %. Of the 12 SNPs listed as outliers, 6 were problematic in the QTL mapping, with either too much variation in the theta scores for the dosage calling to be reliable, or inconsistencies between the parental phase and/or dosage inferred from the SNP and that inferred from mapping the QTL. Problems were also detected for nine further SNPs that were not obvious outliers in the MDS plots.

The second iteration of the MDS analyses used 262 SNPs, excluding all those with problems according to the QTL mapping. Each of the four orders of the 262 markers was then analysed using an HMM to look for areas of high recombinations. For LOD weighting and the two-dimensional configuration, excluding markers could remove a

Table 5 Summary statistics for LGI real data. (a) Mean nearest neighbour fit (NNfit) for the first MDS stage, for each combination of weighting and method. The mean NNfit values for the original 277 markers, after the first round of elimination and after the second round of elimination (if necessary) are shown. The last column shows the markers dropped for each round, separated by /. For the order from the genome sequence, the total NNfit = 225.83, mean 0.911. (b) Mean nearest neighbour fit (NNfit) for the second MDS stage, after removal of 15 SNPs identified as problematic by QTL mapping of theta scores, for each combination of weighting and method. The mean NNfit for the remaining 262 markers and the HMM_mean are shown for the MDS fit, after any local swaps indicated by the HMM, after the first round of elimination and after a further round of local swaps if necessary are shown. The column Rank corr gives the rank correlation of the best order for that method with the true order. The last column shows the markers eliminated. For the order from the genome sequence, the HMM_mean = 1.000

Weight	Method	Initial order NNfit (points)	1st elim.	2nd elim.	Omitted
(a)					
LOD	PC2	1.156 (277)	0.979 (275)	0.954 (273)	269, 174/159, 232
LOD	PC3	1.112 (277)	0.970 (269)	0.982 (267)	269, 174, 232, 101, 274, 123, 133, 112/122, 159
LOD^2	PC2	0.902 (277)	0.876 (275)	–	269, 174
LOD^2	PC3	0.958 (277)	0.863 (270)	0.845 (269)	269, 174, 232, 122, 123, 18, 112/101

Weight	Method	Initial ordering		After swaps		1st elim.		After swaps		Rank corr	Omitted SNPs
		NNfit (points)	HMM mean	NNfit (points)	HMM mean	NNfit (points)	HMM mean	NNfit (points)	HMM mean		
(b)											
LOD	PC2	1.013 (262)	1.168	1.003 (262)	0.935	1.021 (261)	1.019	1.019 (261)	0.866	0.996	248
LOD	PC3	0.981 (262)	1.534	–	–	0.993 (257)	1.501	0.999 (257)	0.914	0.996	146, 112, 159, 83, 248
LOD^2	PC2	0.884 (262)	1.149	0.884 (262)	1.099	–	–	–	–	0.997	None
LOD^2	PC3	0.898 (262)	1.401	–	–	0.881 (258)	1.039	–	–	0.997	146, 112, 18, 83

total of 306 recombinants, (mean 1.168), with one marker (SNP 248) responsible for 32 of these and there were three regions where reordering pairs or triplets of markers could reduce the number of recombinations. After removing SNP 248, repeating the MDS analysis and swapping pairs, the count was reduced to 226 (mean 0.866). Details of the SNPs removed in the second round and the reduction in the NNfit and HMM_mean are given in Table 5b for each method and weighting. The LOD^2 weighting and the three-dimensional configuration gave the lowest value for the NNfit.

Genome positions were known for 248 of the SNPs in this group from the potato reference genome version 4.03 (Potato Genome Sequencing Consortium 2011; Sharma et al. 2013) and as a comparison the HMM was used to calculate recombinations for this order, which totalled 248 (mean 1.000). All of the orders inferred by MDS had high rank correlations with these positions (0.996 for the orders using LOD weighting, 0.997 for the orders using LOD^2 weighting). Figure 4 compares the map from PC3 and the LOD^2 weighting with that of Hackett et al. (2013). Apart from a small inverted region near the top, the orders are similar. Supplementary Table 5 shows the positions from the reference genome where known. The order of the top section found by the MDS analysis is closer to that of the genome sequence than that of Hackett et al. (2013).

The 258 SNPs on the map with the lowest NNfit from the LOD^2 weighting and the three-dimensional configuration were also ordered using JoinMap's regression mapping algorithm. While the MDS analysis ran in less than 10 s once the pairwise data file had been read, the JoinMap analysis took 6 days and 13 min. The rank correlation between the orders was 0.999, and the rank correlation between the JoinMap order and that of the reference genome was 0.997. The positions from JoinMap are shown in Supplementary Table 5.

Discussion

We have shown in this study that MDS can be used to construct linkage maps in autotetraploid species. It reveals markers that are well separated from the rest of the linkage group and gives an order with a high correlation with the correct order in the presence of missing values or genotyping errors. It is sufficiently rapid to run repeatedly to eliminate outlying markers. The JoinMap software also produced a map of the experimental data with a very high correlation with the genome sequence, but took just over 6 days to complete a single linkage group, so is unsuitable for general exploration of large datasets.

The constrained MDS obtained the most accurate estimate of marker position in problem-free data using the Haldane map distance with the best estimates of length being derived from LOD weights and the best estimates of order being derived from LOD^2 weights. However, the PC method (whether in two or three dimensions) did nearly as well and was better than the constrained MDS method when recombination fraction was used as a metric of distance. In addition, the PC method was noticeably faster and also more robust to missing data and errors in the data. There were a small number of simulations where the configuration formed an S-shaped curve and in these cases only the ordering from the PC method was satisfactory. This occurred mainly in the linkage group with the fewest markers and a central region with larger intermarker distances, and is expected to occur less as maps become denser. There was little difference between two and three dimensions with problem-free data sets and the output from the two-dimensional MDS was easier to assess. However, the third dimensions can be very helpful in identifying outliers in more complex data sets. The speed of this approach means that it is feasible to explore the effects of dropping out problematic markers to see their influence on the map. The PC approach has a smoothing parameter, which by default is chosen by leave-one-out cross-validation, but as in the case of THREaD Mapper (Cheema et al. 2010) it is possible to manually specify the smoothing constraint instead to explore different fitted curves. With missing data or data with errors, it is not clear whether LOD or LOD^2 weights give better estimates of length, but LOD^2 weights generally give better estimates of orders, so we recommend the use of unconstrained MDS followed by PC, using this weighting. This is the same weighting used in the WLS criterion of JoinMap.

The MDS analysis is based on estimates of the recombination fraction between all pairs of markers and does not use information on the individual genotypes. However, once an order has been estimated from the pairwise data, the original genotype scores can be inspected using an HMM to reconstruct the most likely inheritance of chromosomes from the parents for each offspring. This was used to identify markers with an unusually large number of recombinations over the population and to see whether excluding a marker or swapping its position locally would improve this. The software RECORD (Van Os et al. 2005) orders markers based on a minimum number of recombinations in diploid populations, but due to the higher number of possible phases in an autotetraploid cross, the information about recombination between neighbouring markers varies in precision and more pairs carry little information. Ordering based on the number of recombinations in an autotetraploid would require an HMM to be run for each order under consideration to establish the chromosome configuration and hence the number of recombinations and would be too slow to be practical. However, the HMM provides another

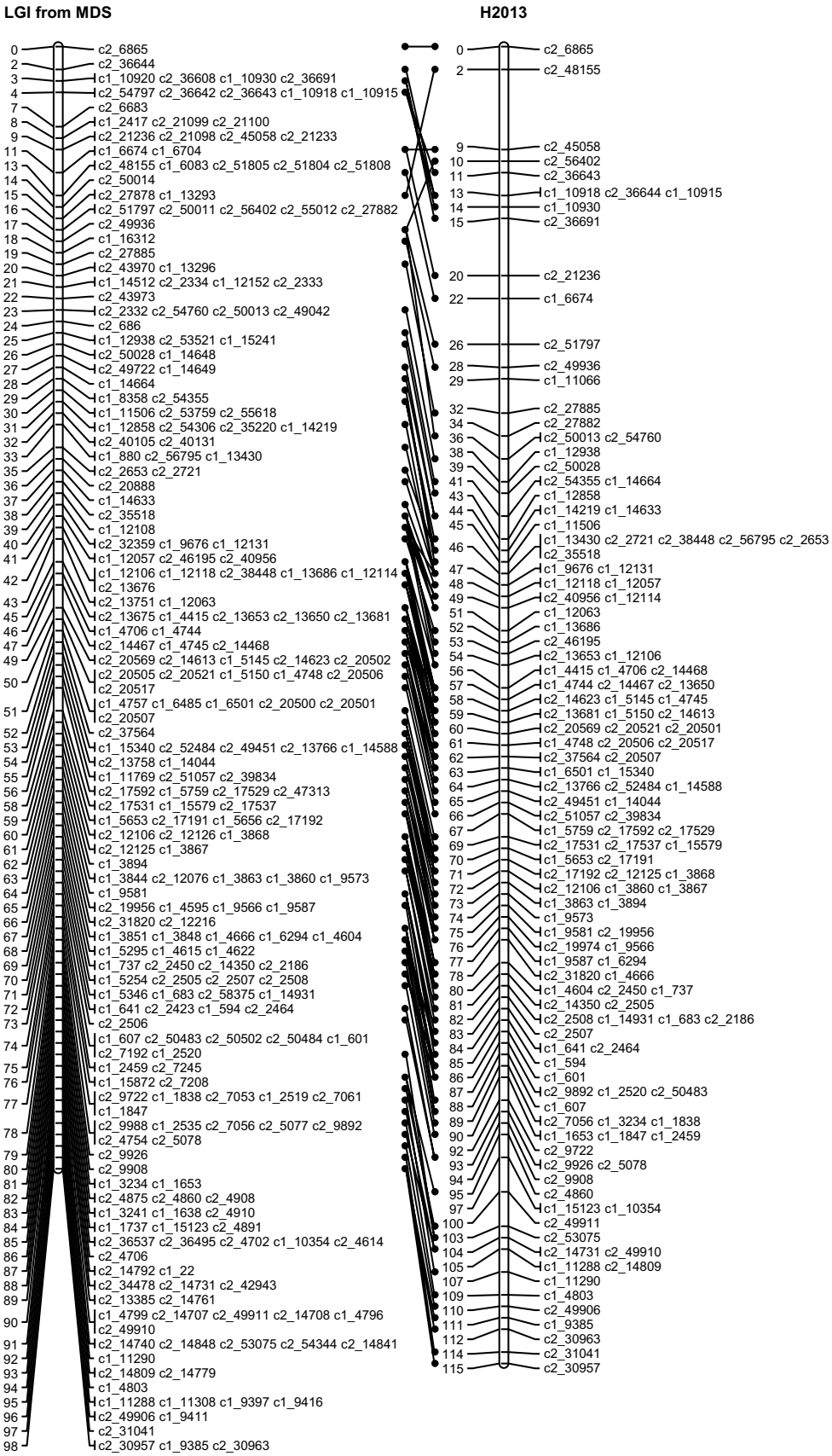

Fig. 4 Comparison of the linkage map based on MDS (LOD^2 weighting, PC3) for potato chromosome I with that from Hackett et al. (2013) (H2013). Marker positions have been rounded to the nearest 1 cM

useful diagnostic that can identify problem markers and suggest where local reordering can improve the map. It also shows which regions of the map have most confidence (i.e. all local swaps increase the count of recombinations) and which have less confidence (some swaps do not change the count of recombinations).

Linkage maps are generally presented in scientific papers as single orders with given marker positions. However, both order and positions are estimated from the data and there is uncertainty in them. Bootstrap sampling from the original population and re-estimation of the map would show the variation (Liu 1998), but this would be very time-consuming, especially for large numbers of markers. JoinMap's maximum likelihood analysis displays 'plausible positions' using a resampling approach, but this cannot currently be applied to autotetraploid populations. If markers are very closely linked, and are informative about the same homologous chromosomes, then it may be useful to consider them together by binning. However if the linkage maps are constructed to map QTLs for phenotypic traits, a single order is necessary for the QTL analysis. The HMM approach developed by Hackett et al. (2013, 2014) for inferring QTL genotype probabilities in an autotetraploid population at a grid of positions along the linkage group uses the dosage information on all markers. If two or more markers are located between recombination events, the information they contribute to the HMM will be invariant to local rearrangements of their order. In Hackett et al. (2013), Table 4 illustrates how an HMM reconstructs the most likely chromosome configuration underlying a short section of four markers and infers a single recombination in the second parent after marker M1; any other permutation of markers M2, M3 and M4 in this example would lead to the same configuration being inferred.

The analysis of the experimental data shows that it is important to select high-quality SNPs for linkage mapping, rather than to call as many SNPs as possible from less clear data. The set of 277 SNPs used here in the map of chromosome I had already been through filtering steps aimed to ensure high-quality genotype calls, as detailed in Hackett et al. (2013). However, some were outliers in the MDS plots and/or were found to have unusually high numbers of recombinations in the HMM. QTL mapping of the theta scores also showed that 15 SNPs among the set had problems that affected the quality of the map. The importance of high-quality markers will increase as even more dense data from genotyping-by-sequencing (GbS) (Elshire et al. 2011) and similar technologies becomes increasingly available in autotetraploid crosses. An alternative approach is the development of ordering methods that can model the uncertainty in the marker dosage estimation.

The methods described in this paper are being integrated into a new version of the TetraploidMap software (Hackett et al. 2007) for use in estimating maps of autotetraploid species. However, these methods can be applied to data on pairwise recombination fractions and LOD scores from any experimental cross and should also be useful in diploid CP populations as a rapid approach for ordering markers without the need to make separate parental maps. We are currently applying these approaches to construct a GbS map of red raspberry (*Rubus idaeus*). The R code for ordering a linkage group using pairwise information on recombination fractions and LOD scores is given in the Supplementary Information, and the MDS and PC methods are also being developed into an R package for general application.

Author contribution statement CAH conceived the project, simulated the data for the autotetraploid simulations and carried out the analyses in Fortran. KFP was responsible for designing the algorithm and implementing the code in R. KFP and CAH wrote the paper together.

Acknowledgments The financial support for this work from the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS) is gratefully acknowledged. We thank Dr. Glenn Bryan, Dr. Karen McLean and colleagues at the James Hutton Institute for use of the potato genotype data and information on the potato reference sequence, and Dr. Herman van Eck and the anonymous reviewers for their constructive comments during the revision of this paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix: Details of the MDS algorithms

Principal curves MDS

1. Use the smacofSym function from the R package smacof (1.4-0) (de Leeuw and Mair 2009) to perform two- or three-dimensional weighted unconstrained MDS on the distance matrix.
2. Plot the final configuration to find potential outliers from Smacofsym plot (see Fig. 1 solid circles for a two-dimensional example and Fig. 3 for a three-dimensional example)
3. Fit the principal curves using the method of Hastie and Stuetzle (1989) implemented in the R package princurve (version 1.1-12) (Hastie and Weingessel 2013).

4. Plot the first principal curve on the final configuration of the unconstrained fit and assess whether it looks reasonable.
5. The projections of the markers onto the first principal curve give the estimated map positions.

Constrained MDS

Steps 1–2 as for principal curve

3. Use the smacofSphere function in two dimensions to constrain the points to approximate to the arc of a circle with a penalty, p, for deviations from the arc.
4. Plot the final configuration from smacofSym and smacofSphere to check for any points which have major changes in rank with respect to either dimension in the final configuration (Supplementary Figure 1A).
5. Check the stress ratio smacofsphere stress/smacofsym stress. This is a metric for the increase in stress (which approximates to a measure of the reduction in fit) caused by forcing the points to lie on an arc and should be below 1.1. If the ratio is above this, return to step 4 and reduce the penalty p.
6. Project the final configuration onto a line to get order and estimated map length.

    (a) Centre sphere on (0, 0).
    (b) Calculate the polar coordinates of each point in the configuration.
    (c) Rotate, so that the mapping starts at the beginning of the arc.
    (d) Radius of the sphere is the median distance of points from (0, 0) rescaled, so that the sum of the configuration is the same as the sum of the observed distances. (We also considered using the mean distance, but this made little difference and the median is less sensitive to outliers and so results are not presented here.)
    (e) Order the markers by increasing the angle.
    (f) Intermarker distances are equal to the radius multiplied by the difference in angle between the points.

References

Cheema J, Dicks J (2009) Computational approaches and software tools for genetic linkage map estimation in plants. Brief Bioinform 10:595–608

Cheema J, Ellis NTH, Dicks J (2010) THREaD Mapper Studio: a novel, visual web server for the estimation of genetic linkage maps. Nucleic Acids Res 38:W188–W193

de Leeuw J, Mair P (2009) Multidimensional scaling using majorization: SMACOF in R. J Stat Softw 31:1–30

Elshire RJ, Glaubitz JC, Qi S, Poland JA, Kawamoto K, Buckler E, Mitchell SE (2011) A robust, simple Genotyping by Sequencing (GbS) approach for high diversity species. PLoS One 6(5):e19379. doi:10.1371/journal.pone.0019379

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP et al (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS One 7:e36347. doi:10.1371/journal.pone.0036347

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137:1121–1137

Hackett CA, Pande B, Bryan GJ (2003) Constructing linkage maps in autotetraploid species using simulated annealing. Theor Appl Genet 106:1107–1115

Hackett CA, Milne I, Bradshaw JE, Luo ZW (2007) TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. J Hered 98:727–729

Hackett CA, McLean K, Bryan GJ (2013) Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. PLoS One 8:e63939

Haldane JBS (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Hastie T, Stuetzle W (1989) Principal curves. J Am Stat Assoc 84:502–516

Hastie T, Weingessel A (2013) Princurve: fits a principal curve in arbitrary dimension. R package version 1.1-12. http://CRAN.R-project.org/package=princurve

Lalouel JM (1977) Linkage mapping from pair-wise recombination data. Heredity 38:61–77

Liu BH (1998) Statistical genomics. CRC Press, Boca Raton

Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne DM (2001) Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics 157:1369–1385

Maliepaard C, Jansen J, Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genet Res 67:55–65

Newell WR, Mott R, Beck S, Lehrach H (1995) Construction of genetic maps using distance geometry. Genomics 30:59–70

Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–197

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P (2013) LepMap: fast and accurate linkage map construction for large SNP datasets. Bioinformatics 29:3128–3134

Schiex T, Gaspin C (1997) CarthaGène: constructing and joining maximum likelihood genetic maps. In: Proceedings of the fifth international conference on intelligent systems for molecular biology, vol 97, pp 258–267

Sharma SK, Bolser D, de Boer J, Sonderkaer M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS, Equiluz M, Guo X, Guzman F, Hackett CA, Hamilton JP, Li G, Li Y, Lozano R, Maass A, Marshall D, Martinez D, McLean K, Mejia N, Milne L, Munive S, Nagy I, Ponce O, Ramirez M, Simon R, Thomson SJ, Torres Y, Waugh R, Zhang Z, Huang S, Visser RGF, Bachem CWB, Sagredo B, Feingold SE, Orjeda G, Veilleux RE, Bonierbale M, Jacobs JME, Milbourne D, Martin DMA, Bryan GJ (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 Genes Genom Genet 3:2031–2047

Shields DC, Collins A, Buetow KH, Morton NE (1991) Error filtration, interference and the human linkage map. Proc Natl Acad Sci USA 88:6501–6505

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™ version 2.0: software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen

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

Van Ooijen JW, Jansen J (2013) Genetic mapping in experimental populations. Cambridge University Press, Cambridge

Van Os H, Stam P, Visser RG, van Eck HJ (2005) RECORD: a novel method for ordering loci on a genetic linkage map. Theor Appl Genet 112:30–40

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from minimum spanning tree of a graph. PLoS Genet 4(10):e1000212. doi:10.1371/journal.pgen.1000212

Zhao H, Speed TP (1996) On genetic map functions. Genetics 142:1369–1377