Theor Appl Genet (2012) 125:1393–1402
DOI 10.1007/s00122-012-1920-9

ORIGINAL PAPER

Three-point appraisal of genetic linkage maps

W. R. Gilks · S. J. Welham · J. Wang · S. J. Clark · G. J. King

Received: 12 April 2012 / Accepted: 5 June 2012 / Published online: 29 June 2012

Abstract This paper develops a simple diagnostic for the investigation of uncertainty within genetic linkage maps using a Bayesian procedure. The method requires only the genotyping data and the proposed genetic map, and calculates the posterior probability for the possible orders of any set of three markers, accounting for the presence of genotyping error (mistyping) and for missing genotype data. The method uses a Bayesian approach to give insight into conflicts between the order in the proposed map and the genotype scores. The method can also be used to assess the accuracy of a genetic map at different genomic scales and to assess alternative potential marker orders. Simulation and two case studies were used to illustrate the method. In the first case study, the diagnostic revealed conflicts in map ordering for short inter-marker distances that were resolved at a distance of 8–12 cM, except for a set of markers at the end of the linkage group. In the second case study, the ordering did not resolve as distances increase, which could be attributed to regions of the map where many individuals were untyped.

Communicated by M. Sillanpaa.

W. R. Gilks · S. J. Welham · S. J. Clark
Department of Computational and Systems Biology, Rothamsted Research, Harpenden AL5 2JQ, UK

W. R. Gilks (✉)
Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK
e-mail: wally.gilks@maths.leeds.ac.uk

S. J. Welham
VSN International Ltd, 5 The Waterhouse, Waterhouse St, Hemel Hempstead HP1 1ES, UK

J. Wang · G. J. King
Plant Sciences, Rothamsted Research, Harpenden AL5 2JQ, UK

J. Wang
Centre for Haemato-Oncology, Institute of Cancer, Barts and the London School of Medicine and Dentistry, Charterhouse Square, London EC1M 6BQ, UK

G. J. King
Southern Cross Plant Science, Southern Cross University, PO Box 157, Lismore, NSW 2480, Australia

Introduction

A genetic linkage map represents the distribution of recombination within a genome, by allocating genetic markers to linkage groups and ordering and assigning positions to these markers according to the patterns of recombination between them. The relative positioning is based purely on observed recombination events across a structured population, which is dependent upon the distribution of chiasmata along the chromosomes. When genotype scores for a large number of genetic markers are available, the evaluation of all possible marker orders within each linkage group is not practically possible and so the optimal ordering may not be found. Moreover, because of errors in genotyping (mistyping), there are likely to be conflicts between the overall best order at local and larger scales. Mapping programs (e.g. Lander et al. 1987; Lincoln and Lander 1992; Jansen et al. 2001; Van Ooijen 2006; Wu et al. 2008; de Givry et al. 2005; Cartwright et al. 2007; Cheema and Dicks 2009) find a compromise ordering by optimizing with respect to an objective function. Mapping methods based on Bayesian multipoint models have been proposed by Neumann (1991), Stephens and Smith (1993), Schiex and Gaspin (1997), George et al. (1999), Rosa et al. (2002), York et al. (2005), Gasbarra and Sillanpää (2006)

and George (2005). The purpose of this paper is not to address the global problem of finding an ordering of all markers, but to use Bayesian methods to provide a simple diagnostic to examine the local compatibility of the map ordering with the observed data.

Various simple diagnostics are used to validate map orderings. R/qtl (Broman et al. 2003) provides a plot of pairwise recombination fractions for markers in map order that visually highlights individual markers that are clearly out of position. Apparent double recombinations, where a single locus genotype is flanked by loci of alternative genotype, suggest an apparent double cross-over. As with recombinations occurring within very short distances, these are often used as indications of misordering (or mistyping). No probabilistic diagnostic procedures appear to have been developed.

This paper develops a simple probabilistic diagnostic procedure to assess the uncertainty associated with marker order in a genetic linkage map. In some senses, this can be seen as a Bayesian extension of the pair-based method of R/qtl to a triple-based method. We deal with plant populations of bi-parental inbred lines where, at each marker, each individual takes one of two possible genetic states with probability equal to 0.5. This includes populations of doubled haploid (DH), advanced recombinant inbred lines (RILs) or backcross populations. The method could easily be extended for F_2 or early generation RILs, where both homozygous and heterozygous individuals are present (three genetic states). A diagnostic method has been developed that assesses the posterior probability of different orders for triples of markers, and hence determines the uncertainty associated with the proposed order. The marker triples can be chosen to suit different purposes: a set of neighboring markers may be chosen to assess the uncertainty at small scales, or more distant markers may be chosen to assess the reliability of the map at larger scales.

Our approach accommodates markers with a substantial proportion of missing genotype data and also allows for mistyping. The method is similar in spirit to that of Stringham and Boehnke (2001), who also calculated posterior probabilities for different marker orders, although they used their results directly for QTL detection, taking into account uncertainties in map order, rather than for investigation of the linkage map. Neumann (1991) also calculated posterior probabilities of ordering for marker triples, but used the estimated inter-locus genetic distances, and did not take account of missing or mistyped genotype data. Ehm et al. (1996), Douglas et al. (2000) and Sobel et al. (2002) used a model similar to ours, but with the benefit of information from pedigree relationships and with the aim of identifying the occurrence of mistyping; our aim in allowing for mistyping is simply to reflect it in the resulting uncertainty in marker order.

The paper first recaps the background to the problem and introduces notation. Calculation of the posterior probabilities is then described in detail. A small simulation is used to demonstrate the behavior of the method, and the method is applied to the analysis of two existing experimental data sets. Finally, potential uses and extensions of the method are discussed. Software for implementing the methodology is freely available, as described in "Software".

Methods

Notation and assumptions

For simplicity in developing the method, we work with genotype data from a single bi-parental cross for a set of markers from a single study. We assume that there are two possible genetic states, and the marginal probability of an individual taking either genotype at a marker is equal, which is appropriate for DH, RIL or back-cross populations under Mendelian segregation and in the absence of segregation distortion. Generalization of these assumptions is considered in "Discussion". Finally, we assume that a genetic linkage map has been proposed, which we wish to examine for evidence of local inaccuracies.

Consider a triple of markers on the genetic map, labeled A, B and C, each with two possible genetic states labelled as 0 or 1. Let A denotes the true genotype of an individual at marker locus A, where A takes values in $\{0, 1\}$; similarly B and C denote the true genotypes at marker loci B and C, respectively. Let M denotes the unknown true map order of markers A, B, C, so M has three possible states: $M = \text{ABC}$, $M = \text{BCA}$ or $M = \text{CAB}$. We do not need to distinguish orientation, so for example $M = \text{CBA}$ is equivalent to $M = \text{ABC}$. For $M = \text{ABC}$, we assume that recombinations in interval AB are independent of those in BC. Let ρ_1 denotes the probability of recombination between the first and the second of the three markers in map M, and let ρ_2 denotes the probability of recombination between the second and the third of these markers. Thus, for $M = \text{ABC}$, ρ_1 relates to interval AB, and ρ_2 to BC (see Fig. 1).

Each marker may be mistyped or untyped, where we assume that the probability of a marker being mistyped or untyped does not depend on the observations at other markers, on the true genotype of the individual, or on the

Fig. 1 One arrangement of three markers (A, B and C), showing probability of recombination ρ_1 within interval AB and ρ_2 within interval BC

Table 1 Definition of π_{aA} for a mistyping probability of ϵ

	A	
	0	1
a		
−1	1	1
0	$1-\epsilon$	ϵ
1	ϵ	$1-\epsilon$

true map M. Let a denote the *observed* genotype on marker A, taking values in $\{-1, 0, 1\}$, where -1 denotes that the marker is untyped; similarly, b and c denote the *observed* genotypes of markers B and C. We define π_{aA} to be the conditional probability of observing a given that the true genotype is A *and* given that $a \neq -1$. For $a = -1$, as explained below, it is convenient to define:

$$\pi_{-1A} = 1. \tag{1}$$

We similarly define π_{bB} and π_{cC}. Initially, we assume that (mis)typing probabilities π_{aA}, π_{bB}, π_{cC} are all known, but may differ. For a common probability ϵ of mistyping, we would set π_{aA} (and π_{bB}, π_{cC}) as in Table 1. Let n denote the number of individuals in the study. The data are the counts r_{abc} of individuals in each *observed* genotype (a, b, c), combined as a vector $\mathbf{r}$ of length 27.

Posterior marker order probabilities

We use a Bayesian approach to the calculation of posterior probabilities for each map order. The joint probability distribution of the data and parameters can be written as

$$p(M, \rho_1, \rho_2, \epsilon, \mathbf{r}) = \ell(\mathbf{r}; M, \rho_1, \rho_2, \epsilon)p(M)p(\rho_1)p(\rho_2)p(\epsilon),$$

where $\ell(\mathbf{r}; M, \rho_1, \rho_2, \epsilon)$ is the likelihood of the data $\mathbf{r}$ under map M, conditional on the recombination probabilities ρ_1 and ρ_2 and mistyping rate ϵ, $p(M)$ is the prior probability of map M, $p(\rho_1)$ and $p(\rho_2)$ are the prior density functions for the recombination probabilities and $p(\epsilon)$ is the prior density function for the mistyping rate. As we are investigating the map, we do not wish to impose any prior information on either the map order or the recombination frequencies, and so we use equal prior probabilities for the three map orders, $p(M) \equiv 1/3$, and uniform probability density functions on the range [0, 0.5] for the recombination fractions ρ_1 and ρ_2. We assume a prior Beta distribution for the mistyping rate, with

$$p(\epsilon) = \frac{\epsilon^{\alpha-1}(1-\epsilon)^{\beta-1}}{B(\alpha, \beta)}, \tag{2}$$

for $\epsilon \in [0, 1]$ where $B(\alpha, \beta)$ is the Beta function. This distribution allows the prior density to be concentrated at low mistyping rates, as would be expected in most studies, but also allows a wide range of distributional shapes.

The joint posterior distribution of the parameters takes the form

$$p(M, \rho_1, \rho_2, \epsilon|\mathbf{r}) = \frac{p(M, \rho_1, \rho_2, \epsilon, \mathbf{r})}{p(\mathbf{r})} = \frac{p(M, \rho_1, \rho_2, \epsilon, \mathbf{r})}{\sum_M \int_{\rho_1,\rho_2,\epsilon} p(M, \rho_1, \rho_2, \epsilon, \mathbf{r}) \mathrm{d}\rho_1 \mathrm{d}\rho_2 \mathrm{d}\epsilon}$$

and the marginal posterior distribution of map order M can be found by integrating over the other parameters as

$$p(M, \rho_1|\mathbf{r}) = \frac{\int_{\rho_1,\rho_2,\epsilon} p(M, \rho_1, \rho_2, \epsilon, \mathbf{r}) \mathrm{d}\rho_1 \mathrm{d}\rho_2 \mathrm{d}\epsilon}{\sum_M \int_{\rho_1,\rho_2,\epsilon} p(M, \rho_1, \rho_2, \epsilon, \mathbf{r}) \mathrm{d}\rho_1 \mathrm{d}\rho_2 \mathrm{d}\epsilon}. \tag{3}$$

Evaluation of the posterior probabilities of different map orders therefore requires integration of the joint probability distribution, which is a product of the data likelihood and the prior distributions. The constant prior distributions of the recombination probabilities and the map order appear in both the numerator and the denominator of the posterior probabilities and so can be ignored.

Log-likelihood

The counts r_{abc} of individuals in each observed genotype (a, b, c) are multinomially distributed, for which the log-likelihood under map M is:

$$L_M(\rho_1, \rho_2) = f + \sum_{a,b,c} r_{abc} p(a, b, c \mid M), \tag{4}$$

where $p(a, b, c \mid M)$ is the map-specific probability that an individual has observed genotype (a, b, c), and f is the logarithm of a combinatoric term which is constant with respect to (ρ_1, ρ_2) and map M. We assume that the observed genotype at each locus is independent of all other observations, given the true genotype. Let $p(a|A)$ denote the probability of observed genotype a given true genotype A. Similarly, $p(b|B)$ and $p(c|C)$ for markers B and C, respectively. From these assumptions, it follows that

$$p(a, b, c \mid M) = \sum_{A,B,C} p(a|A)p(b|B)p(c|C)\theta^*_{ABC}, \tag{5}$$

where θ^*_{ABC} is the probability of true genotype (A, B, C). The superscript * is used throughout to indicate dependence on the map order M and recombination probabilities, ρ_1 and ρ_2. We assume a population with two possible true genotypes at each marker which occur with equal probability. For map $M = \mathrm{ABC}$ and recombination probabilities (ρ_1, ρ_2), the probability of true genotype (A, B, C) is therefore written as

$$\theta^*_{ABC} = \frac{1}{2}\rho_1^{\delta_{\mathrm{AB}}}(1-\rho_1)^{(1-\delta_{\mathrm{AB}})}\rho_2^{\delta_{\mathrm{BC}}}(1-\rho_2)^{(1-\delta_{\mathrm{BC}})}, \tag{6}$$

where δ_{AB} and δ_{BC} are cross-over indicators for intervals AB and BC:

$$\delta_{AB} = |A - B|; \quad \delta_{BC} = |B - C|.$$

Thus, $\delta_{AB} = 0$ denotes no recombination and $\delta_{AB} = 1$ denotes recombination between markers A and B.

Now suppose for an individual that the probability of marker A being untyped is α, independently of the true genotype of that individual and of the true map M. Given that the individual has been typed at marker A and has true genotype A, we assume that the probability of observing type a is π_{aA}. Then the probability of observing a given true genotype A is:

$$\begin{aligned} p(a|A) &= \begin{cases} \alpha, & a = -1 \\ (1-\alpha)\pi_{aA}, & a = 0, 1 \end{cases} \\ &= \pi_{aA}(1-\alpha)^{I(a\neq -1)}\alpha^{I(a=-1)}, \end{aligned} \tag{7}$$

using definition (1), where $I(\cdot)$ is the indicator function which takes the value 1 when its argument is true and zero otherwise. We derive similar expressions for markers B and C, where β and γ denote the probabilities of markers B and C, respectively, being untyped. Hence,

$$p(b|B) = \pi_{bB}(1-\beta)^{I(b\neq -1)}\beta^{I(b=-1)} \tag{8}$$

$$p(c|C) = \pi_{cC}(1-\gamma)^{I(c\neq -1)}\gamma^{I(c=-1)}. \tag{9}$$

Substituting (7–9) into (5) gives:

$$p(abc \mid M) = e^{g_{abc}}\omega^*_{abc}, \tag{10}$$

where

$$\omega^*_{abc} = \sum_{A,B,C} \pi_{aA}\pi_{bB}\pi_{cC}\theta^*_{ABC}. \tag{11}$$

and g_{abc} is constant with respect to (ρ_1,ρ_2) and map M. Substituting (10) into (4) gives:

$$L_M(\rho_1, \rho_2) = f + h + \sum_{a,b,c} r_{abc} \ln \omega^*_{abc}, \tag{12}$$

where f and $h = \sum_{a,b,c} r_{abc}\, g_{abc}$ are constant with respect to the parameters $(\rho_1, \rho_2, \epsilon)$ and map M. For maps BCA and CAB, similar calculations apply.

Integration

Integration over the joint probability function for the three parameters, ρ_1, ρ_2 and ϵ, was achieved by simple numerical integration, using the trapezium rule with a given number of intervals across the range of each parameter. Intervals of size 0.01 gave stable values of the posterior probabilities to two decimal places when the total number of lines was <1,000. Smaller intervals might be required for larger numbers of lines.

Evaluation of the method by simulation

Simulation is useful to demonstrate the properties of the method under known conditions. Data were generated from the underlying multinomial distribution for a triple of markers (A, B, C) in order M = ABC for populations of size $n = 50$ or $n = 100$ lines. The distances between markers A and B (d_{AB}) and between B and C (d_{BC}) were set equal or unequal (d_{BC} smaller), with distances converted into probabilities of recombination using the Haldane distance function (Haldane 1919). Mistyping was either absent or present with probability $\epsilon = 0.025$ and with genotype scores at each locus either all present ($q = 0$) or missing with probability $q = 0.5$. Mistyping and missingness were applied to genotype scores A, B, C independently at random for each line. On average, the data sets with $n = 100$ lines and 50 % missing scores had all genotype scores present for 1 in 8 triples (denoted full triples), but the same number of genotype scores observed as for $n = 50$ lines with no missing observations. A Beta distribution with parameters $\alpha = 1$ and $\beta = 40$ (mean 0.024, standard deviation 0.024, mode 0) was used as the prior distribution for the mistyping probability. For each parameter combination, 100 data sets were generated. For each data set, the posterior probabilities of map orders were calculated as described above.

Application to genetic linkage maps

The posterior probabilities for triples can be used in several different ways to investigate genetic maps. It is possible to examine specific triples of interest, or all consecutive triples within the map. In the case where a single marker is misplaced, we can predict the pattern of posterior probabilities that will occur. In practice, the compromise ordering generated by a mapping procedure is rarely this straightforward. In addition, examining triples of consecutive markers will result in uninformative posterior probabilities where there are few observed recombinations. It is often more useful to scan triples of markers at a specified distance such that a reasonable number of recombinations would be expected, or across a range of distances. For relatively sparse maps, the use of a minimum distance may be more realistic, and this approach was taken in the first of the two examples below.

Brassica napus BnaTNDH population

The BnaTNDH population (Qiu et al. 2006) is a population of 202 doubled haploid oilseed rape (*Brassica napus*, $n = 19$) derived from an F1 with homozygous parents TapidorDH (derived from a European winter variety) and Ningyou7 (derived from a Chinese semi-winter variety). The population shows segregation for many agronomically important traits (Shi et al. 2009). The genetic map was obtained from the brassica.info website (http://www.brassica.info/CropStore/maps.php, map identifier BnaTNDH_05_2008b) and the

Table 2 Occurrence of inferences on marker order based on 100 simulated data sets

d_{AB}	d_{BC}	n	ϵ	q	Decision made			
					ABC	BCA	CAB	None
5	2	50	0	0	54	1	0	45
5	2	50	0.025	0	51	20	8	21
5	2	100	0	0	84	2	0	14
5	2	100	0.025	0	66	16	1	17
5	2	100	0	0.5	35	12	0	53
5	2	100	0.025	0.5	39	19	13	29
10	4	50	0	0	74	4	1	21
10	4	50	0.025	0	67	13	2	18
10	4	100	0	0	92	3	0	5
10	4	100	0.025	0	72	17	0	11
10	4	100	0	0.5	55	18	3	24
10	4	100	0.025	0.5	49	20	7	24
5	5	50	0	0	79	1	1	19
5	5	50	0.025	0	61	7	6	26
5	5	100	0	0	97	0	0	3
5	5	100	0.025	0	80	5	5	10
5	5	100	0	0.5	53	2	4	41
5	5	100	0.025	0.5	52	12	11	25
10	10	50	0	0	94	0	1	5
10	10	50	0.025	0	76	4	5	15
10	10	100	0	0	99	1	0	0
10	10	100	0.025	0	96	1	2	1
10	10	100	0	0.5	66	7	10	17
10	10	100	0.025	0.5	62	11	12	15

Decision made on marker order when posterior probability >0.5, no decision made when all posterior probabilities ≤0.5. True marker order = ABC, with inter-marker distances d_{AB} and d_{BC} cM. Simulations used n lines, mistyping probability ε and probability q of missing genotype information

genotype scores were obtained from the John Innes Centre website (http://www.jic.ac.uk/staff/ian-bancroft/tapidor-_x.htm). This analysis used a single linkage group, A9, which has 61 markers. A Beta distribution with parameters $\alpha = 1$ and $\beta = 40$ was used as the prior distribution for the mistyping probability. Each marker on the linkage group was evaluated with respect to its nearest neighbors after applying a minimum distance criterion (2, 4, 8 or 12 cM).

Arabidopsis Col × Ler population

The Col × Ler population (Lister and Dean 1993) has been widely used for genetic studies in the model plant organism and is a population of 101 RILs of *Arabidopsis thaliana*, with segregation and map data available from the NASC website (ftp://ftp.arabidopsis.org/Maps/seqviewer_data/sv_marker.data and http://arabidopsis.info/RI_data/full_markers.may2001.xls). This analysis used Chromosome 2 with 80 markers. A Beta distribution with parameters $\alpha = 1$ and $\beta = 40$ was used as the prior distribution for the mistyping probability. In this case, with a higher density of markers on the map, marker triples that satisfied a specific range of inter-marker distances were assessed. For each range, a marker was assessed if it had one neighbor on each side within the specified inter-marker range. If several neighbors on one side fell within that range, then the neighbor with inter-marker distance closest to the range mid-point was selected. The ranges chosen were: 0–4, 4–8, 8–12, 12–16, 16–20, 21–27 and 27–33 cM. These ranges were chosen to enable a reasonable number of triples to be assessed at each distance (min = 49, max = 78). For maps with a denser marker set, we would expect to be able to reduce the ranges and still find sufficient triples satisfying the distance criterion.

Results

Simulation

A set of possible inferences from the posterior probabilities is presented in Table 2 as a summary of the simulations. The following decision rule was used; if the posterior probability for any order was >0.5, this was designated as

Table 3 Mean posterior probabilities and observed recombination fraction (RF) from 100 simulated data sets

d_{AB}	d_{BC}	n	ϵ	q	Mean posterior prob.			Mean RF		
					ABC	BCA	CAB	AB	BC	CA
5	2	50	0	0	0.59	0.30	0.11	0.04	0.02	0.06
5	2	50	0.025	0	0.53	0.32	0.15	0.09	0.06	0.10
5	2	100	0	0	0.74	0.23	0.04	0.05	0.02	0.06
5	2	100	0.025	0	0.65	0.27	0.08	0.09	0.07	0.11
5	2	100	0	0.5	0.46	0.35	0.20	0.05	0.02	0.07
5	2	100	0.025	0.5	0.42	0.33	0.25	0.09	0.07	0.11
10	4	50	0	0	0.72	0.23	0.05	0.08	0.04	0.12
10	4	50	0.025	0	0.67	0.25	0.08	0.13	0.09	0.16
10	4	100	0	0	0.84	0.15	0.01	0.09	0.04	0.12
10	4	100	0.025	0	0.72	0.25	0.02	0.13	0.08	0.16
10	4	100	0	0.5	0.51	0.32	0.17	0.09	0.04	0.13
10	4	100	0.025	0.5	0.49	0.32	0.19	0.12	0.08	0.16
5	5	50	0	0	0.70	0.16	0.14	0.04	0.04	0.08
5	5	50	0.025	0	0.62	0.20	0.18	0.09	0.08	0.12
5	5	100	0	0	0.88	0.06	0.06	0.05	0.05	0.09
5	5	100	0.025	0	0.75	0.14	0.11	0.09	0.09	0.13
5	5	100	0	0.5	0.50	0.24	0.26	0.04	0.05	0.09
5	5	100	0.025	0.5	0.49	0.27	0.25	0.10	0.09	0.15
10	10	50	0	0	0.84	0.07	0.09	0.09	0.09	0.17
10	10	50	0.025	0	0.73	0.12	0.15	0.13	0.14	0.20
10	10	100	0	0	0.95	0.02	0.03	0.09	0.10	0.17
10	10	100	0.025	0	0.88	0.06	0.06	0.13	0.13	0.19
10	10	100	0	0.5	0.58	0.21	0.21	0.09	0.09	0.17
10	10	100	0.025	0.5	0.55	0.22	0.22	0.13	0.12	0.21

True marker order = ABC, with inter-marker distances d_{AB} and d_{BC} cM. Simulations used n lines, mistyping probability ε and probability q of missing genotype information

the inferred order, if all of the posterior probabilities were $\leq$0.5, no decision was be made. Clearly, many other decision rules could be chosen in practice, but this simple rule is adequate to illustrate the performance of the method. The mean posterior probabilities and observed recombination fractions are shown in Table 3.

The correct inference on marker order was made more often, with a higher posterior probability, as the number of full triples increased, as the inter-marker distances increased and in the absence of mistyping. The occurrence of indecision was more common for smaller distances (less observed recombination). For unequally spaced markers, an incorrect decision was more often made in favor of permuting the closer pair of markers (i.e., BCA rather than ABC). Incorrect decisions were evenly spread over both possibilities when the inter-marker distances were equal. Where incorrect decisions occurred, they resulted from a pattern of observed recombination contrary to that expected from the inter-marker distances. Table 3 shows that the presence of mistyping with probability of 0.025 adds 0.03–0.05 to the observed recombination fractions. In combination with missing genotype scores, anomalous patterns of recombination become far more likely, especially when the true recombination probabilities are low. Figure 2 shows the full distribution of posterior probabilities for the distances $d_{AB} = 10$ cM, $d_{BC} = 4$ cM represented in barycentric triangles. The probability of incorrect inference (close to the lower vertices) increases as the number of full triples decreases. Uncertainty, as represented by the area in the centre of the triangle, away from the vertices, also increases as the number of full triples decreases and when mistyping is present.

Instead of exploring inferences that result from given combinations of *true* (expected) recombination fractions, as in the above simulations, we can proceed in the opposite direction by studying the inferences that result from given combinations of *observed* recombination fractions. Table 4 provides posterior distributions on marker order for given combinations of observed recombination fractions, assuming $n = 100$ lines and prior distributions as per the simulations. This shows that uncertainty in marker order arises when an observed recombination rate is high.

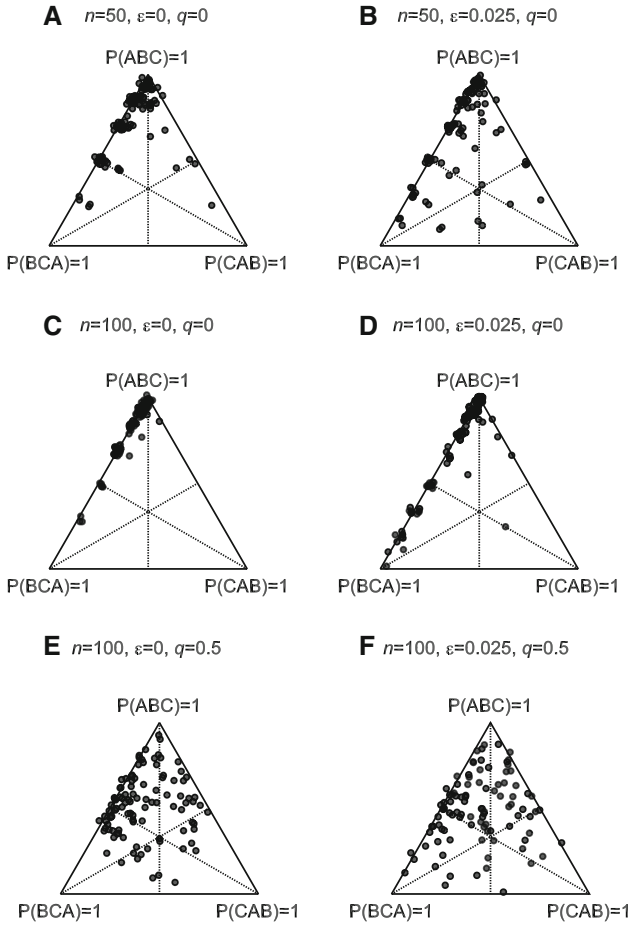

Fig. 2 Posterior probabilities for true marker order ABC with inter-marker distances $d_{AB} = 10$ cM, $d_{BC} = 4$ cM. Points are jittered so that duplicates are visible. Simulations used n lines, mistyping probability ϵ and probability q of missing genotype information

Brassica napus BnaTNDH population

Figure 3 plots the posterior probabilities of the map order against marker position, using a filled circle to indicate that an incorrect order has posterior probability >0.5, in order to distinguish uncertainty from a strong prediction of the incorrect order. Apart from a dubious segment around position 120 cM, this linkage group appears well resolved at an inter-marker distance of 8 cM, although there appeared to be local difficulties at smaller distances, especially around position 70 cM. The anomaly for the markers around position 120 cM is caused by markers at the end of the linkage group (positions 130–139 cM), which form a subset with relatively high recombination with respect to the rest of the linkage group, but with low recombination with a subset between positions 70 and 95 cM. At minimum distance 12 cM, this unexpected pattern of recombination gave high posterior probability to the order CAB. These results suggest that further inspection of this set of markers should be carried out to verify that they have been assigned to the correct linkage group.

Arabidopsis Col × Ler population

With inter-marker distances in a specific range, it is reasonable to examine the distribution of the map order posterior probabilities by genetic distance in order to gain some information on map reliability at different genetic distances. Figure 4 plots the proportion of posterior probabilities for the map order above given threshold values, e.g. 87.2 % of the posterior probabilities for the map order are greater than 0.2 for inter-marker distance 0–4 cM. At inter-marker distances where the genetic map is in agreement with the observed recombination, we would expect high posterior probabilities for the mapped marker order.

For this map, the estimation appears to improve as inter-marker distances are increased up to around 8 cM, but then gets worse at larger distances. This is supported by the data in Fig. 5, which show that some areas of the map appear inconsistent with the observed recombinations across all scales. Further examination of the raw data shows that the positions which consistently show low posterior probabilities correspond to sets of markers with a high probability of being untyped (missing data). The diagnostic results would allow targeted genotyping in these areas to resolve this uncertainty.

Discussion

We have shown that posterior probabilities can be calculated for triples of markers and used to assess the uncertainty associated with different orderings. A certain amount of discrepancy between a given map order and the posterior probabilities may be expected at short distances, due to mistyping errors that cannot be resolved and require compromise to obtain an ordering that is acceptable across the full set of markers. These discrepancies should become less important as the inter-marker distances increase, and consequently uncertainty in marker order should also decrease. We have shown that this can be assessed by plotting posterior probabilities as a function of marker position (BnaTNDH example), or by examining the distribution of posterior probabilities (Col × Ler example), for a given range on inter-marker distances. When integrating genetic maps, the posterior probabilities assist in quantifying uncertainty about different potential positions to be occupied, as markers are incorporated from one map into another. For the Col × Ler population, a sequence pseudochromosome assembly that positions the markers on the genome is also available from the same website. The

Table 4 Posterior distributions of marker order for various combinations of *observed* recombination fractions $\hat{\rho}_{AB}$ and $\hat{\rho}_{BC}$, assuming $n = 100$ lines and prior distributions as per the simulations

$\hat{\rho}_{AB}$	Map M	$\hat{\rho}_{BC}$					
		0.20	0.25	0.30	0.35	0.40	0.45
0.20	ABC	1.00	1.00	0.98	0.91	0.75	0.59
	BCA	0.00	0.00	0.00	0.00	0.00	0.00
	CAB	0.00	0.00	0.02	0.09	0.25	0.41
0.25	ABC	1.00	1.00	0.99	0.95	0.80	0.62
	BCA	0.00	0.00	0.00	0.00	0.00	0.00
	CAB	0.00	0.00	0.01	0.05	0.20	0.38
0.30	ABC	0.98	0.99	1.00	0.97	0.85	0.65
	BCA	0.02	0.01	0.00	0.00	0.00	0.00
	CAB	0.00	0.00	0.00	0.03	0.15	0.35
0.35	ABC	0.91	0.95	0.97	0.97	0.87	0.67
	BCA	0.09	0.05	0.03	0.02	0.01	0.00
	CAB	0.00	0.00	0.00	0.02	0.12	0.33
0.40	ABC	0.75	0.80	0.85	0.87	0.82	0.66
	BCA	0.25	0.20	0.15	0.12	0.09	0.06
	CAB	0.00	0.00	0.00	0.01	0.09	0.29
0.45	ABC	0.59	0.62	0.65	0.67	0.66	0.56
	BCA	0.41	0.38	0.35	0.33	0.29	0.22
	CAB	0.00	0.00	0.00	0.00	0.06	0.22

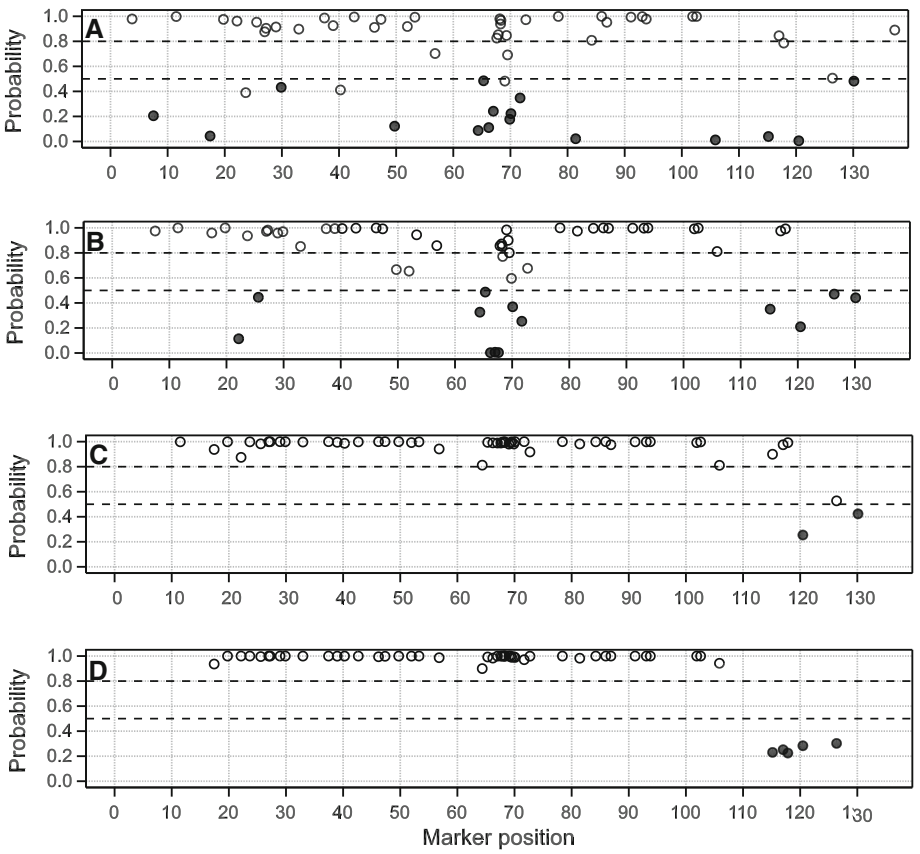

Fig. 3 Posterior probabilities for marker triples on linkage group A9 for the BnaTNDH genetic map, plotted against position of central marker. Minimum inter-marker distances: **a** 2 cM, **b** 4 cM, **c** 8 cM, **d** 12 cM. *Solid circles* Markers with posterior probability >0.5 for a marker order different from that given by the BnaTNDH genetic map

markers can then be assessed with respect to their physical order. This analysis will highlight discrepancies between the physical order and observed recombination.

This approach can easily be extended. The extension to a set of independent studies on the same population, using overlapping sets of markers, is straightforward and results

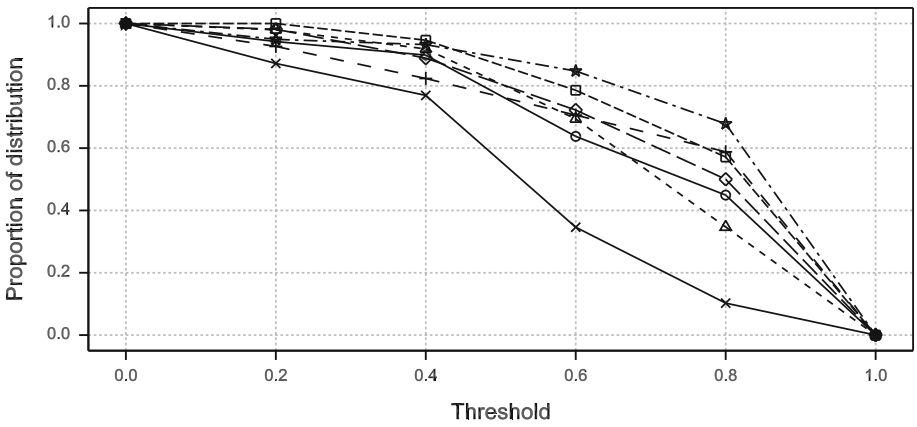

Fig. 4 Proportion of posterior probabilities above a given threshold for a range of inter-marker distances for the Col × Ler genetic map: *crosses* 0–4 cM, *circles* 4–8 cM, *plus symbols* 8–12 cM, *stars* 12–16 cM, *triangles* 16–20 cM, *squares* 21–27 cM, *diamonds* 27–33 cM

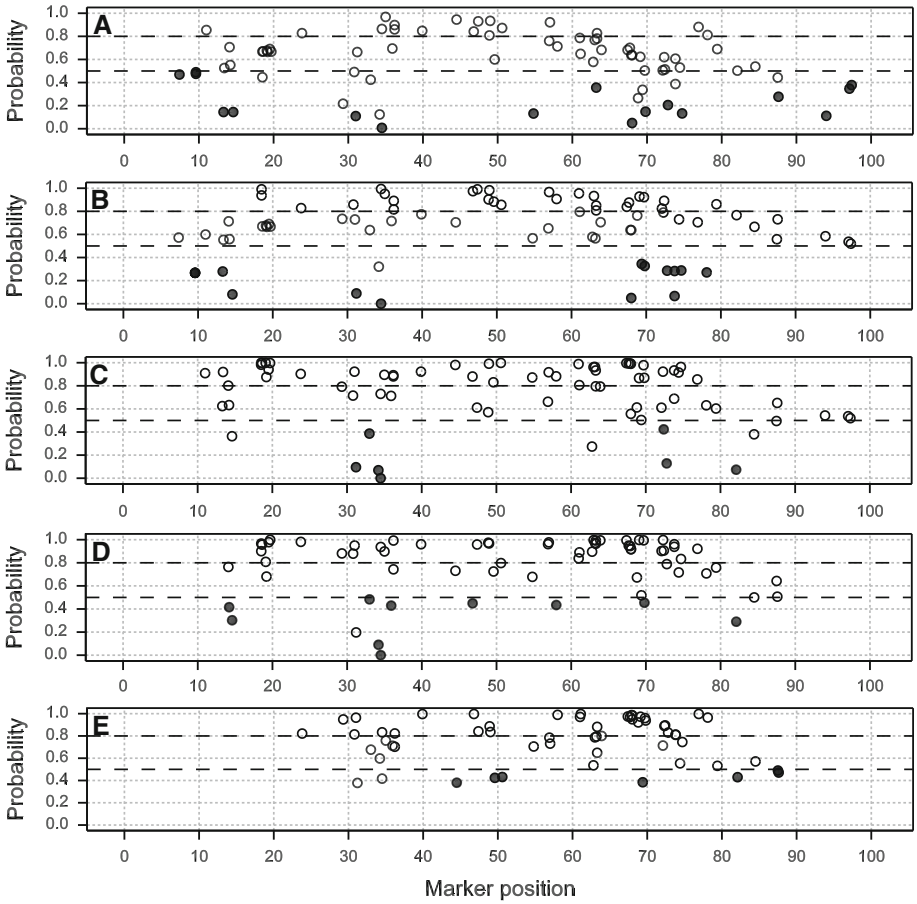

Fig. 5 Posterior probabilities for marker triples on chromosome 2 for the Col × Ler genetic map, plotted against position of central marker. Minimum inter-marker distances: **a** 2 cM, **b** 4 cM, **c** 8 cM, **d** 12 cM, **e** 18 cM. *Solid circles* Markers with posterior probability >0.5 for a marker order different from that given by the Col × Ler genetic map

in the same estimation procedure. The extension to a set of studies on different populations using a common marker set on the same genetic map is also straightforward. In this case, labelling of the parents in different populations can be arbitrary, as it is the presence of recombination rather than the genotype scores that give information on marker order. The method could also be extended to other populations, such as F_2 (or F_n) lines, simply by modifying the number of possible genetic states and the probability of the observed genotype in (6). One area of interest is the application to map integration across different populations (Wang et al. 2011). The method can be used to test or evaluate different marker orders and give insight into the amount of uncertainty present. In this case, a different map order might be required for different populations. For example, a local rearrangement of the markers in one subpopulation may have occurred, or a similar rearrangement may have occurred in a paralogous segment of the same genome. The method can be extended to evaluate the evidence for such genomic rearrangements.

Software

Numerical integration was implemented in a Fortran subroutine, with arguments provided to specify the data, the number of integration points for each parameter and parameters of the prior Beta distribution function for the mistyping rate. The subroutine returns the integrals in Eq. (3) and posterior probabilities derived from it. Interfaces to the Fortran subroutine are available as an R package or GenStat procedures. The R package can be freely downloaded from http://www1.maths.leeds.ac.uk/~wally.gilks/LinkageMapAppraisalR-package/Welcome.html. In this package, function `pp` performs the calculation of posterior probabilities, and functions `triplecheck` and `groupcheck` work with a *cross object*, as defined by package R/qtl (Broman et al. 2003).

Acknowledgments The authors are funded by the UK Biotechnology & Biological Sciences Research Council (BBSRC) with project funding to GJK and JW under BBE0177971, and SJW and GJK from Department for Environment Food and Rural Affairs project IF0144 (OREGIN). We would like to thank Jonathan Myles for wrapping our R functions into a package, and two anonymous referees for comments which led to improvements in the algorithm, paper and interpretation of our method.

References

Broman K, Wu H, Sen S, Churchill G (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Cartwright D, Troggio M, Velasco R, Gutin A (2007) Genetic mapping in the presence of genotyping errors. Genetics 176:2521–2527. doi:10.1534/genetics.106.063982

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

de Givry S, Bouchez M, Chabrier P, Milan D, Schiex T (2005) CARTHAGENE: multipopulation integrated genetic and radiated hybrid mapping. Bioinformatics 21:1703–1704

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

Ehm MG, Kimmel M, Cottingham J R W (1996) Error detection for genetic data, using likelihood methods. Am J Hum Genet 58:225–234

Gasbarra D, Sillanpää M (2006) Constructing the parental linkage phase and the genetic map over distances <1 cM using pooled haploid DNA. Genetics 172:1325–1335. doi:10.1534/genetics.105.044271

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

George A, Mengersen K, Davis G (1999) A Bayesian approach to ordering gene markers. Biometrics 55:419–429

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

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

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

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

Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J 4:745–750

Neumann P (1991) Three-locus linkage analysis using recombinant inbred strains and Bayes' theorem. Genetics 128:631–638

Qiu D, Morgan C, Shi J, Long Y, Liu J, Li R, Zhuang X, Wang Y, Tan X, Dietrich E, Weihmann T, Everett C, Vanstraelen S, Beckett P, Fraser F, Trick M, Barnes S, Wilmer J, Schmidt R, Li J, Li D, Meng J, Bancroft I (2006) A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. Theor Appl Genet 114:67–80

Rosa G, Yandell B, Gianola D (2002) A Bayesian approach for constructing genetic maps when markers are miscoded. Genet Sel Evol 34:353–369. doi:10.1051/gse:2002012

Schiex T, Gaspin C (1997) CarthaGene: constructing and joining maximum likelihood genetic maps. In: Proceedings of the international conference on intelligent systems for molecular biology. AAAI Press (http://www.aaai.org), pp 258–267

Shi J, Li R, Qiu D, Jiang C, Long Y, Morgan C, Bancroft I, Zhao J, Meng J (2009) Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus*. Genetics 182:851–861

Sobel E, Papp J, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 70:496–508

Stephens D, Smith A (1993) Bayesian inference in multipoint gene mapping. Ann Hum Genet 57:65–82

Stringham HM, Boehnke M (2001) Lod scores for gene mapping in the presence of marker map uncertainty. Genet Epidemiol 21:31–39

Van Ooijen J (2006) JoinMap 4. Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, Netherlands

Wang J, Lydiate D, Parkin I, Falentin C, Delourme R, Carion P, King G (2011) Integration of linkage maps for the amphidiploid *Brassica napus* and comparative mapping with Arabidopsis and *Brassica rapa*. BMC Genomics 12:1–20. doi:10.1186/1471-2164-12-101

Wu Y, Bhat P, Close T, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLOS Genetics 4:e1000,212. doi:10.1371/journal.pgen.1000

York T, Durrett R, Tanksley S, Nielsen R (2005) Bayesian and maximum likelihood estimation of genetic maps. Genet Res 85:159–168. doi:10.1017/S0016672305007494