Theor Appl Genet (2010) 120:1653–1672
DOI 10.1007/s00122-010-1283-z

ORIGINAL PAPER

Bayesian estimation of marker dosage in sugarcane and other autopolyploids

Peter Baker · Phillip Jackson · Karen Aitken

Received: 30 June 2008 / Accepted: 24 January 2010 / Published online: 25 February 2010

Abstract In sugarcane or other autopolyploids, after generating the data, the first step in constructing molecular marker maps is to determine marker dosage. Improved methods for correctly allocating marker dosage will result in more accurate maps and increased efficiency of QTL linkage detection. When employing dominant markers like AFLPs, single-dose markers represent alleles present as one copy in one parent and null in the other parent, double-dose markers are those present as two copies in one parent and null in the other parent and so on. Observed segregation ratios in the offspring are employed to infer marker dosage in the parent from which the marker was inherited. Commonly, for each marker, a χ^2 test is used to assign dosage. Such an approach does not address important practical considerations such as multiple testing and departures from theoretical assumptions. In particular, extra-binomial variation or overdispersion has been observed in sugarcane studies and standard methods may result in fewer correct dosage allocations than the data warrant. To address these shortcomings, a Bayesian mixture model is proposed where all markers are considered simultaneously. Since analytic solutions are not available, Markov chain Monte Carlo methods are employed. Marker dosage allocation for each individual marker employs the estimated posterior probability of each dosage. For a sugarcane study these methods resulted in more markers being allocated a dosage than by standard approaches. Simulation studies demonstrated that, in general, not only are more markers classified but that more markers are also correctly classified, particularly if overdispersion is present.

Introduction

Commercially important polyploid crops include potato, alfalfa, sweet potato and sugarcane. Polyploidy or the presence of more than two genomes per cell is an important method of species formation in plants (Stebbins 1950). Indeed, Soltis and Soltis (2000) argued that polyploid plant species are successful due to their inherent genetic characteristics since compared to their diploid progenitors, polyploids generally exhibit higher levels of heterozygosity, less inbreeding depression and higher levels of selfing as well as higher genetic diversity and the potential for duplicated genes and genomes to evolve new functions.

Autopolyploids may be regarded as being derived from a single species and are expected to have polysomic inheritance due to the ability of each homolog to pair with any other homolog during meiosis. On the other hand, allopolyploids which can be regarded as being derived

Communicated by J. Bradshaw.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1283-z) contains supplementary material, which is available to authorized users.

P. Baker
CSIRO Mathematics, Informatics and Statistics,
St Lucia, QLD, Australia

Present Address:
P. Baker (✉)
School of Population Health, The University of Queensland,
Herston, QLD 4006, Australia
e-mail: p.baker1@uq.edu.au

P. Jackson
CSIRO Plant Industry, Private Mail Bag PO Aitkenvale,
Aitkenvale, QLD 4814, Australia

K. Aitken
CSIRO Plant Industry,
306 Carmody Rd, St Lucia, QLD 4067, Australia

from several distinct species, exhibit disomic inheritance. It may be that some polyploids are not purely auto or allopolyploid. Instead, some may possess varying levels of preferential pairing. Methods to assess levels of preferential pairing are the subject of ongoing research and controversy (Janoo et al. 2004; Sybenga 1994, 1995, 1996; Ripol et al. 1999; Xie and Xu 2000; Soltis and Soltis 2000; Qu and Hancock 2001; Hackett 2001; Wu et al. 2001a, b, 2002; Luo et al. 2004). The methods presented here assume perfect autopolyploidy without preferential pairing. Implications of departures from these assumptions are discussed.

In diploids, informative dominant markers occur when a marker is present in one parent but not the other. In polyploids, a similar situation will occur if one parent possesses one or multiple doses of a marker. Single-dose markers are commonly called simplex markers, double dose are referred to as duplex, and triple-dose markers are known as triplex markers. Like in diploids, markers may be linked in coupling if the dominant allele is physically located on the same chromosome or linked in repulsion if the dominant alleles are on different chromosomes in the same homology group. Ripol et al. (1999) provide general expressions for linkage between single- and multiple-dose markers in various coupling and repulsion configurations. For the purposes of marker mapping and QTL linkage analysis, single-dose markers in coupling may be treated as diploid and so many methods are readily available. Typically, in marker map construction, a framework map is constructed using simplex markers in coupling and then simplex markers in repulsion and multi-dose markers are added using methods like maximum likelihood.

Statistical methods for genomic mapping are relatively well developed for diploids (Lander and Botstein 1989; Haley and Knott 1992). However, methods for more complex polyploids are much less developed due to issues such as the unknown number of gene copies or dosage (Burner 1997) and unknown allelic configuration (Luo et al. 2001). Nevertheless, DNA linkage maps have been constructed for autotetraploids (De Winton and Haldane 1931; Wu et al. 1992), autohexaploids (Ukoskit and Thompson 1997) and autooctoploids (Ripol et al. 1999; Aitken et al. 2005, 2007). While a unified approach is available for autotetraploids via the *TetraploidMap* software (Hackett and Luo 2003), more ad hoc methods are often employed for species with higher ploidy levels. Ukoskit and Thompson (1997) noted that although simplex markers can be mapped by the approach of da Silva et al. (1993) and Al-Janabi et al. (1993), other polysomic segregations cannot be mapped with diploid methods. Subsequently, da Silva et al. (1995) showed that double- and triple-dose markers may be added to a framework linkage map consisting of simplex markers in autooctoploids and Meyer et al. (1998) added duplex and double simplex markers for QTL analysis in tetraploid potato. More complicated procedures have also been employed for marker mapping (Ripol et al. 1999; Wu et al. 2001a) but all methods rely on identifying the dosage for each marker.

Segregation ratios, or ratios of the number of offspring exhibiting the marker to those which do not, play an important part in inferring marker dosage in the parent. Haldane (1930) first tabulated expected segregation ratios in the offspring where one parent is nulliplex for a range of ploidy levels. The proportion of markers observed in the offspring is called the *segregation proportion* in this article.

Segregation distortion occurs when there are departures from these theoretical distributions.

Like other plants, sugarcane commonly exhibits segregation distortion as seen in Fig. 1. Since modern sugarcane cultivars originally arose from the hybridization of multiple genomes, they are highly heterozygous and subject to meiotic irregularities (Grivet and Arruda 2002). Potential causes include chromosome loss, translocations, rearrangements and double reduction. Commonly maps were constructed assuming such irregularities are not present (da Silva et al. 1993; Ming et al. 1998; Aitken et al. 2005) since methods to incorporate them are not well established. While Aitken et al. 2007 incorporated distorted markers into the linkage analysis, most of the distorted markers were unlinked or unmapped. However, incorporating distorted markers had a considerable effect on map lengths and several chromosome rearrangements such as chromosome breakages, fusions and translocations were detected.

Routinely, for assessing marker dosage, a standard χ^2 test is employed (Mather 1951), usually at the 5% level or perhaps at the 1% level, to test all markers for a segregation

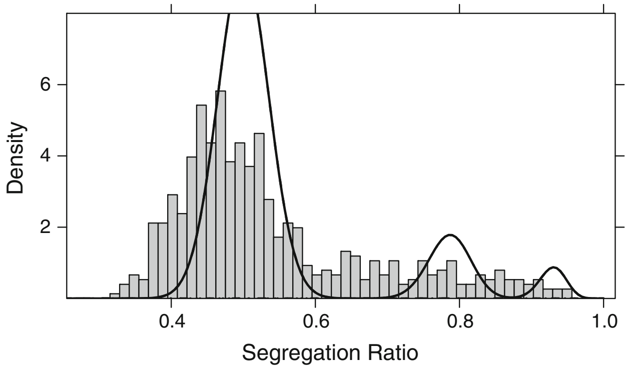

Fig. 1 *Histogram* of observed segregation ratios for 566 AFLP markers inherited from KQ99-1410 in a sugarcane trial at Ayr, Queensland. The *solid line* is the expected theoretical distribution obtained as a mixture of single, double and triple dose markers with proportions of each type of marker estimated by the χ^2 test. Similar proportions were estimated by a mixture model fitted on the logit scale. There is a shift to the left which could be due to aneuploidy or chromosome loss

ratio of 1:1 which is expected for simplex markers. Markers are also tested for alternate dosages based on the appropriate expected segregation ratio. An alternative method employing the binomial distribution (Ripol et al. 1999) should produce similar dosage allocations as the sample size increases since the χ^2 and binomial distributions are asymptotically equivalent (Mood et al. 1974).

These methods appear to possess inherent limitations such as ignoring multiple testing in the case of the χ^2 test or not allowing for measurement error in the binomial confidence interval method. However, of more practical importance is that marker dosage may be wrongly allocated if the underlying assumptions are not met or if significance levels are too stringent. For instance, if extra-binomial variation or overdispersion is induced by measurement errors or correlations between markers then the resulting distributions may well be wider than the binomial and so markers will not be allocated a dosage even when their segregation ratio is clearly part of a wider dosage distribution. Also, note that for the standard χ^2 test, the aim of allocating marker dosage to the null distribution is not equivalent to the usual aim of rejecting the null hypothesis. The result is that markers may not be allocated a dosage even if their segregation ratio clearly belongs to a particular null distribution. A similar problem arises when the segregation ratio of a marker may yield non-significant tests for more than one dosage and so is not assigned either dosage.

In order to circumvent these limitations and to estimate Bayesian posterior probabilities of possible marker dosages, a mixture model approach is proposed. Markers are allocated to empirical distributions estimated from a combination of the data and theoretical parameter values. Each component of the mixture distribution corresponds to a specific marker dosage. While a direct mixture of binomials is feasible, it is more straight forward to consider a model where the segregation ratios are considered to be a finite mixture of normals on the logit scale. This also has the added advantage of allowing for overdispersion by inducing a hierarchical model (Gelman et al. 2004) and so each component can have a wider distribution than the standard binomial which has a fixed variance $p(1 - p)/N$, where N is the sample size and p is the binomial proportion.

Employing a Bayesian mixture model approach not only allows prior genetic theory be incorporated, but also provides estimates of the uncertainty of marker dosage classification. The posterior probability distribution of belonging to each dosage component is estimated for each marker.

Results from sugarcane and simulated marker data indicate that, in general, more markers are allocated a dosage by an appropriate mixture model than by standard methods. Simulation studies also reveal that more markers are correctly allocated, particularly if overdispersion is present.

This article is organised as follows. The motivating example of a sugarcane cross, current approaches and proposed methods are outlined in "Materials and methods". Both the results of simulation studies which led to a recommended model and its application to the sugarcane study are described in "Results". Finally, implications for marker dosage allocation of both the new and previous methods are discussed in "Discussion".

Materials and methods

Case study: a sugarcane AFLP data set

While the methods outlined here may be employed for any autopolyploid, they were motivated by considering amplified fragment length polymorphism (AFLP) and simple sequence repeats (SSR) markers in a sugarcane backcross population. The sugarcane cultivar KQ99.1410 was produced by crossing a *Saccharum officinarum* IJ76.514 with the commercial sugarcane variety Q165. The markers considered here are from 200 progeny generated by crossing KQ99.1410 with an elite sugarcane variety MIDA at Ayr, (147.40E, −19.50S) Queensland. These data were part of a larger experiment which will be reported elsewhere.

In all, 1,725 AFLP markers were generated by the methods of Aitken et al. (2005) for the KQ99.1410 by MIDA cross. Of these, 469 were present in both parents and 596 were inherited from KQ99.1410 only. The latter markers are used to explore standard and mixture model approaches to estimate marker dosage. In order to remove the effects of segregation distortion, 30 markers with a segregation ratio below 0.325 were eliminated from this study. This threshold was set below the 99th percentile of the appropriate binomial distribution in order to allow for the possibility of overdispersion.

The observed segregation proportions and theoretical distributions for 566 AFLP markers inherited from KQ99.1410 are shown in Fig. 1. The theoretical distributions for each dosage are binomial with parameters n and p. The probability parameters p were set as the appropriate segregation proportions obtained from Eq. 1 assuming octoploidy while the n were taken to be the numbers of markers in each dosage class obtained from the χ^2 test. Not all markers could be assigned a dosage and so observed numbers were scaled up to ensure the correct area under the estimated theoretical distribution. A similar number of markers in each class were obtained by application of the fitted mixture models. It is clear that the observed data do not conform to the expected theoretical distribution.

Single-dose markers appear to have lower segregation ratios than expected. This may be due to aneuploidy which is commonly seen in sugarcane.

Marker dosage in autopolyploids

Polyploids have multiple copies of the basic chromosome set, whereas diploids have two copies. The number of chromosomes in a gamete is typically denoted n, while the monoploid number or size of the basic chromosome set is written as x. The ploidy m is the number of homologous chromosomes in each somatic cell and so $2n = mx$. Until the end of the nineteenth century, sugarcane cultivars were predominantly clones of *S. officinarum* which is a high sugar content octoploid with $2n = 80$ (Grivet and Arruda 2002) Early breeders improved yield and disease resistance by crossing *S. officinarum* to *S. spontaneum*. Modern sugarcane cultivars appear to have between $2n = 80$ and $2n = 140$ chromosomes, indicating that aneuploidy is common (Burner 1997).

Expected segregation ratios when one parent is nulliplex

Haldane (1930) derived the expected ratios of offspring for various parental configurations of autopolyploids with an even number of sets of chromosomes. Expected gametic series for polyploids of various sizes were produced, along with expected ratios of gametic series for crosses and selfing and the equilibrium distribution under random mating. Haldane provided expected gametic series for polyploids up to order 16 (heccaidecaploid). Haldane denoted a gamete containing x copies of A and y alleles a as $A^x a^y$. For example, consider an autooctoploid with A being the dominant marker allele and a is an alternate allele. Crossing a parent with simplex marker Aa^7 with a nulliplex a^8 results in one nulliplex in two in the offspring since the gametic series produced in the parent Aa^7 is $1.Aa^3$:$1.a^4$. Thus, offspring are obtained from the cross $(1.Aa^3 + 1.a^4)(1.a^4 + 1.a^4)$ with resultant ratios $1.Aa^7$:$1.a^8$.

The general formula of Ripol et al. (1999) is adopted for P_k or $P(k)$ or the expected segregation proportion for markers with dosage k which is

$$P(k|m) = 1 - \frac{\binom{m-k}{m/2}}{\binom{m}{m/2}}, \; k = 0, \ldots, m/2 \qquad (1)$$

where m is the ploidy level or number of homologous chromosomes. Note that for diploids $m = 2$, tetraploids $m = 4$ and for octoploids $m = 8$.

From Table 1, it is clear that despite increasing ploidy levels that expected theoretical segregation proportions are similar for dosage k when $m \geq 6$. Thus, in sugarcane, which is an irregular autopolyploid possessing from $2n = 80$ to $2n = 140$ chromosomes with monoploid number $x = 8$ or $x = 10$ (Burner 1997), segregation proportions for a particular marker dosage are relatively constant despite changes in ploidy.

Table 1 Expected segregation proportions for ploidy levels from 2 to 12

Dosage	Ploidy					
	2	4	6	8	10	12
Simplex	0.50	0.50	0.50	0.50	0.50	0.50
Duplex		0.83	0.80	0.79	0.78	0.77
Triplex			0.95	0.93	0.92	0.91
Quad				0.99	0.98	0.97
5					1.00	0.99
6						1.00

Expected segregation ratios in Eq. 1 are for autopolyploids. If preferential pairing or double reduction are present then the segregation ratios will be somewhat altered. Consider the case of preferential pairing. For pure allopolyploids then the segregation proportion will be 0.5 for simplex markers. However, for duplex and other multi-dose markers, segregation ratios will be slightly different. For example, in pure allo-octoploids duplex, triplex and quadraplex markers will have segregation proportions 0.75, 0.88 and 0.94, respectively. It is expected that segregation proportions would be intermediate between the corresponding allo and auto values in Table 1. In any case, since the corresponding segregation proportions are so similar, very large sample sizes would be required to see any difference between allo and autopolyploids.

Expected segregation ratios when the marker is inherited from both parents

In diploids only heterozygous markers are informative. In polyploids, even if both parents possess the dominant marker allele some offspring may not inherit a copy. In such cases, there must be less than $m/2$ copies in at least one parent. For instance, crossing two genetically similar autooctoploid lines Aa^7 results in one nulliplex in four since $(1.Aa^3 + 1.a^4)^2$ is simply $(1.A^2a^6 + 2.\mathrm{A}a^7 + 1.a^8)$. Such markers are often employed to help construct homology groups but are not as informative as simplex markers where one parent is nulliplex.

If both parents contain at least one copy of the dominant marker allele then we deduce a general equation similar to Eq. 1 for the expected segregation proportion P_{jk} or $P(j,k)$ to be

$$P(k,j|m) = 1 - \frac{\binom{m-k}{m/2}\binom{m-j}{m/2}}{\binom{m}{m/2}^2}, j,k = 0,\ldots,m/2 \tag{2}$$

where m is defined in Eq. 1 and one parent possesses j doses of the marker whereas the other parent has k copies.

For autooctoploids, Eq. 2 yields $P_{11} = 0.75$, $P_{12} = 0.89$, $P_{22} = 0.95$, $P_{23} = 0.99$. Therefore, it is clear that while it may be possible to readily identify simplex markers present in both parents, other parental dosage combinations are essentially impossible to distinguish.

Current methods of assessing dosage

The two most commonly used approaches for allocating dosage in autopolyploid marker map construction are the χ^2 and binomial methods. For sample sizes typically employed, similar results are expected due to the large sample properties of these approaches. However, rather than rely on asymptotic theory, these methods were compared via simulation studies as described in "Simulation study".

The χ^2 test

The most widely used test for assessing marker dosage is the standard χ^2 test. Following Mather (1951), this test is often employed to compare the observed segregation ratio against its expected value. For instance, to test for simplex markers a standard χ^2 test on 1 *df* is employed since the expected segregation ratio is 1:1 and so

$$X^2 = \frac{(a_1 - a_0)^2}{a_0 + a_1} \sim \chi_1^2, \tag{3}$$

where a_1 is the number of lines with an AFLP band a_0 is the number of lines without the band.

Power calculations, such as those of Wu et al. (1992), have often been employed since non-significant tests are critical in that these lead to allocating marker dosage. However, such calculations may not be entirely relevant since these calculations are designed for testing one marker at one dosage rather than for the hundreds or thousands of simultaneous significance tests typically conducted.

The binomial confidence interval method

As an alternative to the χ^2 test outlined in "The χ^2 test", Ripol et al. (1999) proposed comparing the observed segregation ratio to the expected distribution of segregation ratios assuming a binomial distribution with size N being the number of offspring and parameter P_k set to the segregation ratio in Eq. 1. To allow for missing values from an experiment with 90 plants from the species *Saccharum spontaneum*, Ripol et al. set the binomial parameter N to 80 and then compared the observed segregation ratio for each marker to the region containing 99% of the distribution in the interval between the 0.5 and 99.5 percentiles. Simplex, duplex, triplex and quadruplex markers were all allocated using this approach but no quadruplex markers were found.

Measurement error is ignored with this approach and so the binomial distribution may provide confidence intervals (CIs) that are too narrow, particularly if overdispersion is present.

A mixture model for assessing dosage

Instead of employing the approaches outlined in "Current methods of assessing dosage", we propose to model the observed segregation ratios of markers in the offspring with a finite mixture distribution. For instance, in the case of autooctoploids, the segregation ratios of each marker are postulated to come from one of four binomial Bin(N,P_k) distributions, where $P_k, k = 1,\ldots,K$ is the theoretical segregation ratio P_k in Table 1, N is the number of progeny and $K = 3$ or 4 is the number of different marker dosages.

While it is possible to fit mixtures of binomial distributions, the methods are not well developed (Rufo et al. 2007; Mao 2007) and in any case would not address limitations due to measurement error outlined in "The binomial confidence interval method". However, by logit transforming the observed segregation ratios to be approximately normal, and at the same time employing a more standard finite normal mixture model, these concerns are addressed. Unlike the binomial model, where each component has fixed variance which depends on the segregation ratio and number of progeny, the variance of each normal component may be estimated from the data. Also, the mixture model induces a hierarchical model with the resultant component distributions being more variable. Figure 2 shows simulated segregation ratios for an autohexaploid with some overdispersion.

A Bayesian approach is adopted since it allows the strong prior knowledge of segregation ratios under given ploidy levels to be incorporated explicitly into the modelling process. This is particularly useful for higher doses where there are likely fewer markers. Importantly, in addition to obtaining the posterior distribution of model parameters, the posterior probabilities of various marker dosages may be estimated for each marker.

In other words, this model provides not only estimates of the proportion of markers belonging to each dosage class or component, but also for each individual marker we can

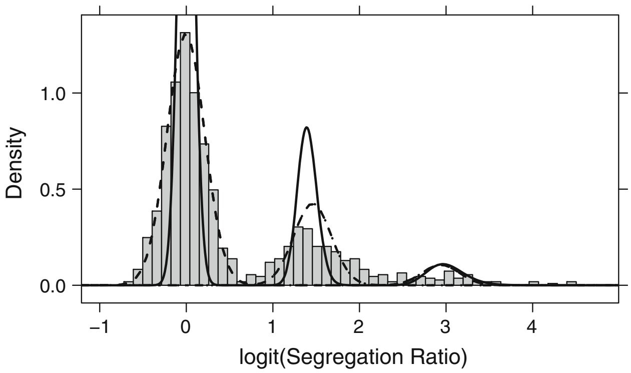

Fig. 2 Simulated data with a small amount of overdispersion, as described in "Simulation study", from 500 autohexaploid progeny with 1,000 markers. The *solid curve* is the theoretical distribution for non-overdispersed data, while the *dashed curve* represents the fitted mixture distribution. The model employed strong priors with equal variance. The fitted distribution better fits the data for single and double-dose markers. However, since there is only a small amount of data for the triple-dose component, the theoretical (prior) and fitted distributions are similar. The 99% binomial CI method correctly allocated 67% of markers as opposed to 98% by mixture method

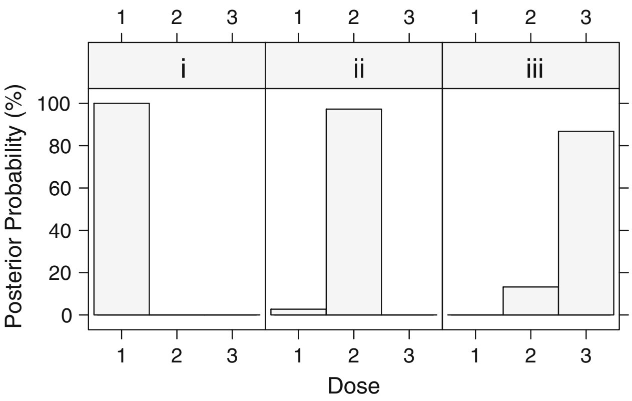

Fig. 3 Posterior probabilities of marker dosages for three markers selected from 566 AFLP markers inherited from the sugarcane KQ99-1410. MCMC runs had a burn in of 5,000 and a sample of 1,0000. Markers *i* and *ii* are clearly simplex and duplex, respectively, while there is slightly weaker evidence that *iii* is a triplex with a posterior probability of around 0.85

obtain the posterior probability of all possible doses as shown in Fig. 3.

Bayesian data analysis may be divided into three steps: (1) setting up the full probability model which depends jointly on the likelihood of the data and the prior distribution, (2) calculating the posterior distribution conditional on the data and (3) evaluating the fit of the model and the implications of the resulting posterior distributions (Gelman et al. 2004).

The joint probability distribution function (pdf) $f(\theta,y)$ of parameters of interest θ and data y can be written as the product of two densities, namely the prior distribution $f(\theta)$ and sampling or data distribution $f(y|\theta)$

$$f(\theta, y) = f(\theta)f(y|\theta)$$

We are interested in the posterior distributions of the parameters given the data

$$f(\theta|y) = \frac{f(\theta)f(y|\theta)}{\int f(\theta)f(y|\theta)\mathrm{d}\theta} \tag{4}$$

where $\int f(\theta)f(y|\theta)\mathrm{d}\theta$ is integrated over all possible values of θ. For fixed y the integral is therefore constant and so Eq. 4 is commonly written as

$$f(\theta|y) \propto f(\theta)f(y|\theta). \tag{5}$$

The models for $f(y|\theta)$ are described in "Models", choosing priors $f(\theta)$ and their implications are discussed in "Priors" and calculating posterior distributions $f(\theta|y)$ via Markov chain Monte Carlo (MCMC) is outlined in "Model fitting and parameter estimation".

Models

For the jth marker $j = 1,\ldots,n$, we assume the observed number r_j of lines with dominant markers out of N_j lines follows a binomial distribution denoted Bin(N_j,P_k), where if we knew the dosage k then P_k could be obtained from Eq. 1. Note that N_j is simply the number of lines if no marker data are missing.

Since the dosage k is unknown, we rely on the missing data representation of Dempster et al. (1977) and Tanner and Wong (1987) which is commonly adopted for MCMC computation in finite mixture models. An indicator variable z_j corresponding to unknown marker dosage class k is introduced where $z_j = k$ if the marker has dose k. For the K components with $K \leq m/2$, consider the logit transformation of the true segregation proportions P_k for dose k, $k = 1,\ldots,k$. The logit transformed segregation ratio ω_k is then

$$\omega_k = \log\left(\frac{P_k}{1 - P_k}\right). \tag{6}$$

Let $z = (z_1,\ldots,z_n)^T$ be a vector of unknown dosages (labelled $1, 2,\ldots,k$ corresponding to simplex, duplex, triplex markers and so on). $r_j|z_j$ is binomially distributed with known size parameter N_j and unknown proportion parameter ω_{Z_j}, where $\omega_{Z_j}$is the segregation ratio for marker dosage z_j. Hence, given marker dosage z_j then

$$r_j|z_j \sim \mathrm{Bin}\left(N_j, \omega_{Z_j}\right), \tag{7}$$

where

$$\mathrm{logit}(\omega_{Z_j}) = \log\left(\frac{\omega_{Z_j}}{1 - \omega_{Z_j}}\right) \sim N\left(\mu_{Z_j}, \tau_{Z_j}^{-1}\right)$$

where μ_k and τ_k are the mean and precision $\left(t_k = 1/s_k^2\right)$ of marker dosage class k on the logit scale.

Since z_j is unknown, then $\mathrm{logit}(\omega_{z_k})$ can be modelled as a finite mixture of K normals

$$\mathrm{logit}(\omega_{z_j}) \sim \pi_1 N(\mu_1, \tau_1^{-1}) + \pi_2 N(\mu_2, \tau_2^{-1}) + \cdots + \pi_k N(\mu_k, \tau_k^{-1}) \quad (8)$$

where μ_k is the mean and τ_k is the precision of component k on the logit scale, and π_k are the mixing proportions of the three components with $\sum_{k=1}^{k} \pi_k = 1$. The probability density function $f(x)$ of $\mathrm{logit}(\omega_k)$ is

$$f(x) = \sum_{k=1}^{k} \pi_k \phi(x|\mu_k, \tau_k^{-1}) \quad (9)$$

where ϕ is the normal cumulative distribution function with parameters mean μ_k and variance $\sigma_k^2 = \tau_k^{-1}$.

Priors

It is our observation that, in general, vague or non-informative priors are often employed in Bayesian data analysis since little a priori information is available. However, in light of genetic theory and current practice, informative prior distributions are warranted here. For the segregation ratios considered here, genetic theory as outlined in Haldane (1930) determines these ratios exactly, albeit under idealised conditions. In contrast, while a Bayesian analysis can assign uncertainty to segregation ratios via a prior distribution, the χ^2 test and binomial distribution outlined in "Current methods of assessing dosage" assume exact ratios.

In any case, for simplex and duplex markers a considerable amount of data are available to estimate the posterior distribution of segregation ratios and so the estimated distributions should not be unduly influenced by the priors. If for instance, the variance of the prior were set to be roughly the same as that observed in the data then the prior would basically only have the influence of a single observation (Gelman et al. 2004, Sect. 2.6). However, for triplex (and quadruplex markers if included) a strong prior may prove useful to obtain realistic posterior distributions as well as to aid MCMC convergence. This is clearly seen for simplex, duplex and triplex marker distributions in Fig. 2.

Results may be compared with those obtained assuming vague priors in order to assess sensitivity to prior specification.

Priors on the means μ_k

Vague priors are typically set for the means μ_k to include the range of the data by specifying $\mu_k \sim N(0, 1/\sqrt{0.1})$.

An informative prior distribution would set the mean as the logit of the theoretical segregation ratio and by specifying a narrow range or variance of the distribution. Thus, $\mu_k \sim N(\mathrm{logit}(P_k), T_k^{-1})$ where P_k is specified in Eq. 1 and T_k^{-1} may be approximated from the specified range of the prior distribution. For instance, in the single-dose marker category $P_1 = 0.5$ and so logit $(P_1 = 0)$. To approximate the variance of the prior distribution we could set the 95% confidence region to be between 0.45 and 0.55. On the logit scale the 2.5 and 97.5 percentiles are −0.2 and 0.2. The size of this interval can be equated to four standard deviations and so the standard deviation is 0.1 and the precision $T_1 = 1/0.1^2 = 100$ Smaller intervals, and hence larger precisions T_k, could be set for higher doses with segregation ratios near 1 if required, either because fewer data points are observed or if the range becomes too large due to the nature of the logit transformation.

Priors of the precisions τ_k on the logit scale

For convenience we can choose vague prior distributions on the precision τ_k or inverse variance by setting parameters of the conjugate Gamma (A, B) distribution as $A = B = 0.1$.

Specifying informative priors on τ_k is more complicated than for the means μ_k. The mean and the approximate variance of the prior on the precision can be obtained by considering the standard deviation of the theoretical binomial distribution transformed to the logit scale. To this end, the mean $\hat{\tau}_k$ of the prior was set as the precision of the theoretical binomial distribution obtained by a process similar to calculating T_K when specifying priors on the means μ_k. Limits were set on the prior distribution by specifying them for τ_k or $s_k = 1/\sqrt{\tau_k}$. This allows the variance to be ascertained and subsequently hyperparameters A and B to be set. If $\hat{\tau}_k(1 \pm x)$ is a set as a 95% confidence region for the prior distribution on the logit scale, then $A = 4/(x^2)$ and $B = 4/(x^2\hat{\tau}_k)$. Alternatively, narrower priors may be obtained by setting the 95% confidence region for s_k as $\tilde{s}_k(1 \pm \mathrm{x})$ which yields $A_k = C.\hat{\tau}_k^4$ and $B_k = C\hat{\tau}_k^3$ where $C = x^2/(1+x)^4(1-x)^4$. For details see Appendix 1.

Mixing proportions π_k of each marker dosage

Genetic theory does not provide any prior information on the proportions of markers expected for each dosage. While in sugarcane, the observed proportions of dosage classes often appear to roughly conform to those expected by da Silva (1993) and Al-Janabi et al. (1993), Qu and Hancock (2002) showed that these proportions cannot be predicted in advance. Qu and Hancock argued that the proportion of each dose category is completely independent of both the inheritance patterns and the segregation ratio of a marker. Hence, we adopt the vague but proper prior

$$(\pi_1, \pi_2, \ldots, \pi_k) \sim \text{Dirichlet}(1, 1, \ldots, 1). \quad (10)$$

Model fitting and parameter estimation

We adopt MCMC as the computational method because analytic solutions for obtaining posterior distributions are not available. The principles of MCMC are simple although implementation may be quite complicated. For further details see Tanner (1993), Smith and Roberts (1993), Besag et al. (1995) and Gilks et al. (1996).

A potential problem with fitting mixtures via MCMC sampling is that chains may inappropriately converge to one of the components. Robert (1996) showed that this problem may be circumvented by reparameterising the means in Eq. 8 as

$$\mu_{k+1} = \mu + \lambda_k, k = 1\ldots(k-1), \quad (11)$$

where the mean shift $\lambda_k > 0$. Note that this corresponds to the natural ordering of segregation ratios which increase with marker dosage. Note also that the parametrisation adopted here, should potentially reduce the possibility of poor posterior sampling due to label-switching which is more likely to occur if the means are not ordered (see Stephens 2000).

Robert (1996) also noted that forcing one observation into each class may be required. In the case where highly informative priors are adopted, convergence should be improved and further reparameterisation, such as that of Mengersen and Robert (1996) on the variances should prove unnecessary.

Gibbs sampling was carried out using JAGS (Plummer 2005) which is similar to BUGS (Spiegelhalter et al. 1995). Given the form of the priors and the likelihood, as well as the nature of the MCMC samples, the MCMC chains are known to converge at least in polynomial time (Tweedie and Mengersen 1996). Computation of theoretical rates of convergence are both extremely complex to compute and very conservative. Hence, we adopted the practical alternative of assessing convergence through diagnostics, such as those available in CODA (Best et al. 1995). These included the diagnostics of Geweke (1992), Raftery and Lewis (1992) and Heidelberger and Welch (1983).

Marker dosage allocation

For each marker, the indicator variable z_j along with the other parameters, were recorded at each iteration of the MCMC sampler to obtain an empirical posterior probability distribution of belonging to each marker dosage class.

While the maximum posterior probability could be used to allocate dosage, this is likely to be too optimistic if a marker is on the boundary between two marker dosage components. Instead, a marker was allocated to a dosage class if its posterior probability was over 0.8 of belonging to that component. In some sense this analogous to the common statistical practises of choosing a sample size in order to attain a power of 80% or setting a 20% false discovery rate. Simulation studies outlined in "Simulation study" confirmed this value to be a reasonable choice.

Model assessment

Spiegelhalter et al. (2002) introduced the deviance information criterion (DIC) for model assessment and comparison for linear and generalized linear models. The DIC is a Bayesian model selection criterion devised to work both as a measure of fit and a measure of complexity with similar aims to the commonly used AIC (Akaike 1974).

Celeux et al. (2006) and others have noted that possible inconsistencies may arise in defining the DIC for missing data models such as for the mixture models employed here. While Celeux et al. compared a number of alternative formulations of the DIC these proved less than satisfactory here since the methods are aimed at parameter estimation rather than marker dose classification. Instead, different models were compared for accuracy of classifying marker dosage and percentage of markers classified by simulation studies as outlined in the following section.

Simulation study

Marker data were simulated for a range of ploidy levels, sample sizes, marker numbers and severity of overdispersion in order to assess the performance of current and proposed methods of dosage allocation.

Five simulated data sets of 200, 500, 1,000, 1,500 markers for 50, 100, 200 and 500 progeny were constructed with a ploidy of $m = 4$, 6, 8 and 10. The proportions of markers in each marker dosage class were set to be similar to those often seen in practice, namely $(0.8,0.2)^T$ for $m = 4$, $(0.7, 0.25, 0.05)^T$ for $m = 6$ and $(0.7, 0.15, 0.1, 0.05)^T$ for $m = 8$ and $m = 10$.

Simulation was carried out as follows. Given, the number of progeny, number of markers and proportions of simplex, duplex, triplex and quadruplex markers, the corresponding numbers of markers for each dosage were generated from a multinomial distribution. Individual marker data were then generated from the binomial distribution with appropriate expected segregation proportions. Three further marker data sets were generated with varying amounts of overdispersion which was introduced by generating markers from a beta-binomial distribution as outlined in Appendix 2. The first shape parameter α was set to 50, 15 and 5 which corresponded to slight, moderate and severe overdispersion, respectively.

The standard methods of allocating marker dosage, namely the χ^2 and binomial methods were employed at the 0.05 and 0.01 α level were compared. Comparisons of percentage of correctly classified markers and also the percentage classified in total were made via Tukey mean difference plots as described in Altman and Bland (1983) and Bland and Altman (1995).

Four mixture model approaches were similarly compared. These employed either informative or uninformative priors and either common or separate variances for each marker dosage component. For the data sets with a ploidy of 8 or 10, due to concerns about convergence and stability of estimates, mixture models with both three and four components were fitted.

Finally, the best overall standard and overall mixture model methods were compared.

Results

Simulation study

Marker dosage was allocated for 960 data sets by the χ^2 and binomial CI methods at the 1% and 5% levels and by the four mixture model methods. Additionally, for those data sets with ploidy 8 or 10, both a three and four component mixture model was fitted which resulted in 1,440 sets of four mixture model results for comparison.

Comparison of current methods

The percentage of correctly classified markers by various methods was compared via Tukey mean difference plots as shown in Fig. 4. These plots indicate that tests conducted at the 1% level correctly classify dosage for more markers than at the 5% level. Note that a difference above zero

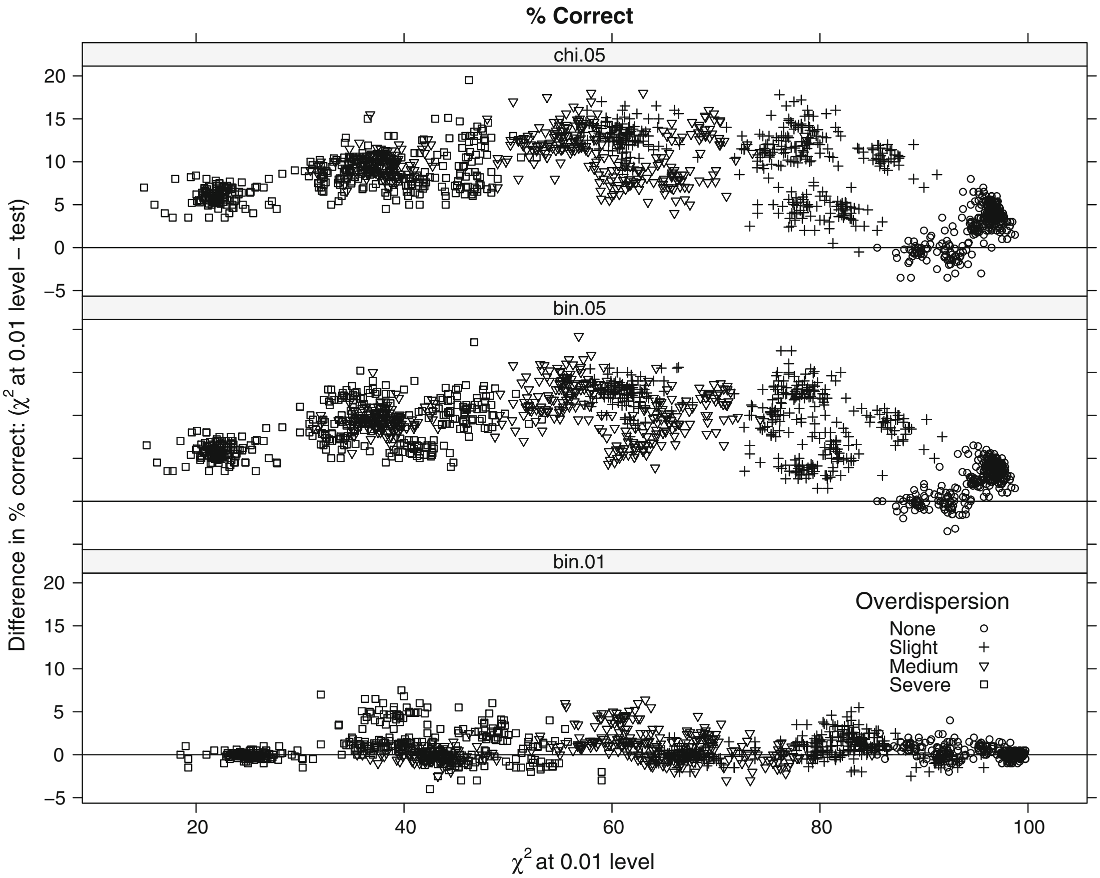

Fig. 4 Tukey mean difference plots of three standard methods for marker dosage allocation, namely binomial 95% CI (*bin.05*), χ^2 at the 5% level (*chi.05*) and the binomial 99% CI (*bin.01*) versus the χ^2 test at the 1% significance level. Overdispersion was set to None (*empty square*), Slight (*inverted triangle*), Moderate (*plus sign*) and Severe (*empty circle*). The *y* axis is the difference in the percentage of markers allocated correctly via the χ^2 test at the 1% level minus the percentage obtained by the other test. The *horizontal line* is set at a zero difference. The plots indicate that the χ^2 test and binomial CI at the 1% level result in similar percentages of markers correctly allocated, whereas both methods at the 5% level result in fewer markers being correctly allocated. It is also clear that all tests result in fewer markers being correctly allocated as overdispersion increases

indicates that the χ^2 test at the 1% level classified more markers than the alternate test and vice versa for a negative difference. The results are as expected since 95% CIs are narrower than 99% CIs and so more markers are classified by methods at the 1% level. Indeed, Tukey mean difference plots of the percentage of markers classified look very similar to the corresponding percent correct in Fig. 4 (not shown).

One concern is that there may be a trade off between correctly classifying more markers and misclassification rate. However, increasing the percentage of correct marker dose allocation by using a smaller α value did not induce a higher classification rate. Instead, in general, the methods applied at the 1% level have a slightly lower misclassification rate than at the 5% level. All four methods have a low misclassification rate when there is no overdispersion but this rate increases as overdispersion increases.

It is clear from Fig. 4 that for all methods fewer markers will be classified as overdispersion increases. This is because the range of segregation ratios widens as the overdispersion increases while the same CI or test value is used, irrespective of overdispersion.

Figure 5 shows clearly a marked increased misclassification of marker dosage with increasing overdispersion. A small increase in misclassification rate is seen with increasing ploidy level although this is always in combination with overdispersion. Interestingly, the percentage correct was independent of the ploidy level $(2, 4, \ldots, 10)$ or number of markers (see Fig. 5. Similarly, the percentage of markers classified was not affected by differing ploidy levels or numbers of markers.

For the χ^2 test at the 1% level, from Fig. 6 it is clear that in the absence of overdispersion, the percentage of markers correctly allocated is very high and increases with more markers. However, when overdispersion is present then more progeny result in less markers being correctly allocated. This is due to the confidence intervals for each marker dosage being narrower as the number of progeny increases.

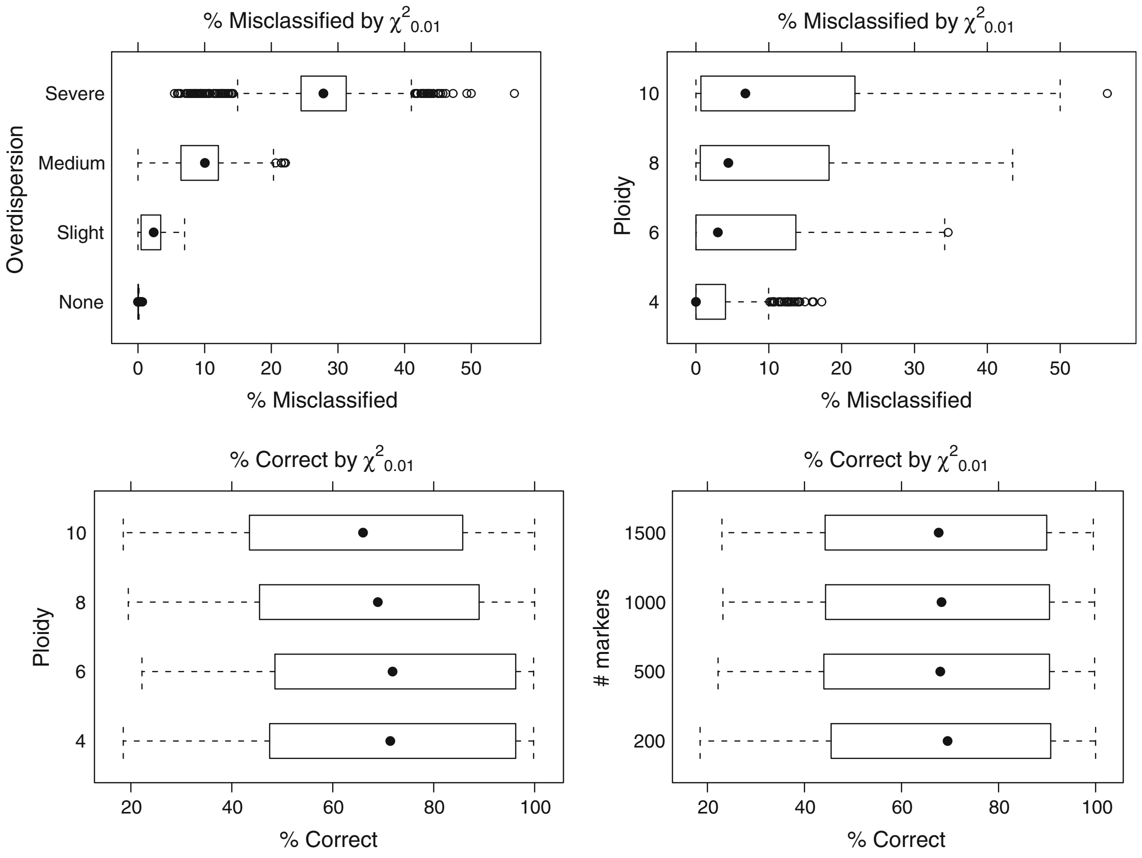

Fig. 5 Box plots of the percentage of markers misclassified with (1) increasing overdispersion and (2) ploidy level along with the percentage of markers where dosage was correctly allocated by χ^2 tests at the 1% level at varying (3) ploidy levels and (4) numbers of markers. Five simulated data sets were produced for a range of ploidy levels $(m = 4, \ldots, 10)$ and overdispersion (*None*, *Slight*, *Medium* and *Severe*), 200–1,500 markers for 100, 200 or 500 progeny. The *top plots* clearly show the (1) effect of overdispersion on misclassification and (2) the smaller effect of increasing ploidy level. The *bottom plots* indicate (3) no difference in correctly allocating dosage for varying ploidy levels 4, 6, 8 or 10 or (4) for a change of 200–1,500 markers

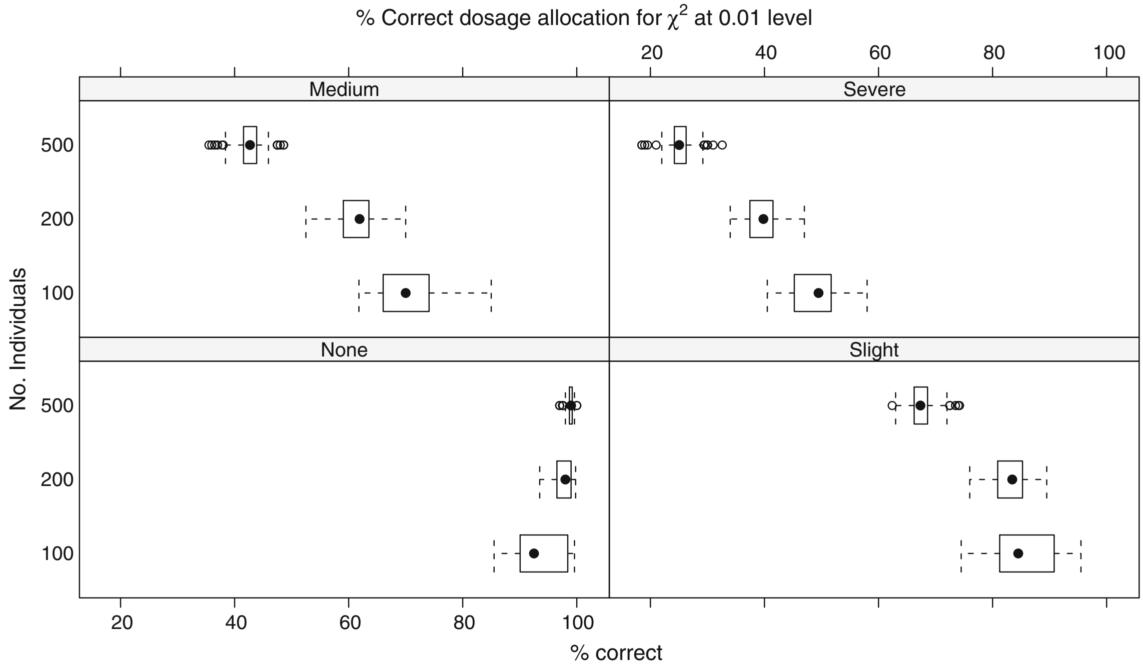

Fig. 6 Box plots of the percentage of markers with dosage correctly allocated by the χ^2 tests at the 0.01 level. Five simulated data sets were produced for a range of ploidy levels ($m = 4, \ldots, 10$) and overdispersion (*None*, *Slight*, *Medium* and *Severe*), 200–1,500 markers for 100, 200 or 500 progeny. Nearly all markers were correctly allocated when there was no overdispersion. Fewer markers were correctly allocated with increasing overdispersion or larger sample size

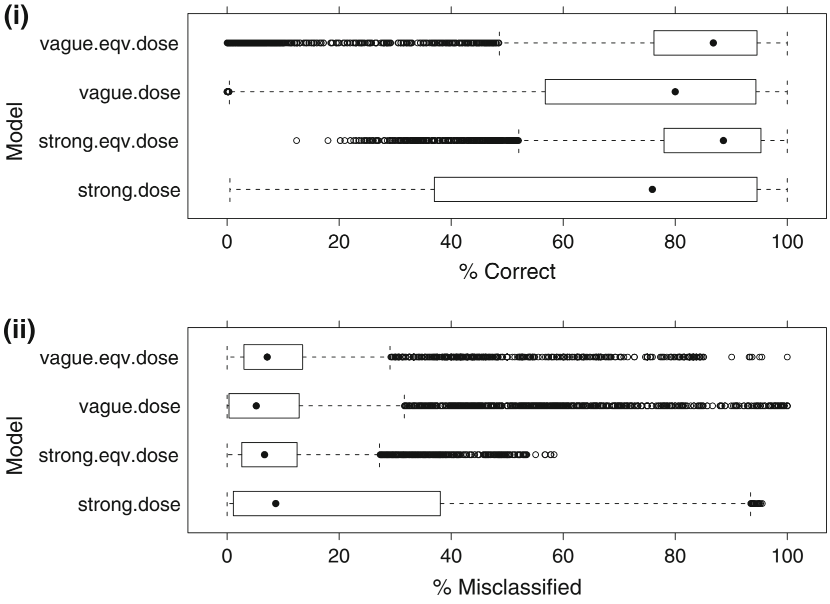

Fig. 7 Box plots of the percentage of markers with **i** dosage correctly allocated and **ii** misclassified by all mixture models which had vague or strong priors, equal variances (*eqv*) or different variances. Note that the box plots include results for all other simulation factors and for a range of posterior probability cutoffs

Choosing the best mixture model approach

In general, as judged by the Geweke (1992) statistic, convergence was achieved with a burn in of 5,000 and MCMC sample of 10,000 although some model/data set combinations with equal variances required longer chains and in these cases further diagnostics were also employed (Best et al. 1995).

From the boxplots in Fig. 7, on balance, the mixture model with equal variances incorporating strong prior knowledge about the theoretical segregation ratios gave more correct dosage allocations with lower misclassification rates. Boxplots include results for all ploidy levels, numbers of markers and individuals as well as where allocation was made by setting the posterior probability cutoff between 0.5 and 0.99. These factors did not influence the overall conclusion and hence choice of best model/prior combination.

For simulated data with a ploidy of 8 and 10, both a 3 or 4 component mixture model was fitted. The former model was considered in order to avoid potential problems with fitting a component consisting of very few four dose markers. However, convergence problems were not evident. This was probably due to strong prior information and also less parameters being required when a common variance parameter was assumed for each component. In general, models with more components performed slightly better as outlined in Appendix 3. Such models may indicate that fewer components are warranted if there is only one or two markers in the highest dosage component. For medium or severe overdispersion at high ploidy levels results proved to be more variable and so, in terms of allocating marker dosage, some data sets did worse.

Finally, to choose the threshold for the posterior probability in cutting off the percentage of correctly allocated markers and misclassification rates were examined. There was no effect of either the number of progeny or number of markers on classification or misclassification rates. There is no clear optimal threshold value although 0.8 seemed a reasonable rate when balancing correct classification and misclassification rates (see Appendix 3).

In summary, it would appear that incorporating strong prior knowledge and equal variances on the logit scale for all components produces better allocations of marker dosage. However, if there is medium to severe overdispersion or if the ploidy level is over 6 then allocation will be less accurate.

Comparison of current and proposed methods

The best approaches of each type of standard and mixture model methods were compared via Tukey mean difference plots. In "Comparison of current methods", the χ^2 test at the 1% level, which is very similar to the binomial 99% CI, proved to be the most effective standard test while in "Choosing the best mixture model approach" the mixture model incorporating strong prior information and where all components had equal variance on the logit scale proved to correctly identify marker dosage than other models.

Overall, from Fig. 8, it can be seen that markers were correctly classified by the mixture model method more often than the best standard method, namely a χ^2 test at the 1% level. The difference between the methods was more pronounced as overdispersion increased and to a lesser extent as ploidy increased. Also, more markers were correctly classified as with increasing numbers of progeny. No trend was seen with increasing marker numbers (not shown). Note that each point represents one simulated data set and a difference greater than 0 indicates the mixture model correctly allocates more markers.

Figure 9 indicates that both mixture models and the a χ^2 test at the 1% level have similar misclassification rates, although for all but the case when there is no overdispersion, more data sets had a lower misclassification rate when the mixture model was employed. However, misclassification rates increased with increasing overdispersion and ploidy levels.

Analysis of sugarcane AFLP data set

The standard methods and the mixture models for dose allocation were applied to the 566 markers inherited from KQ99-1410 from the sugarcane backcross study in "Case study: a sugarcane AFLP data set".

To assign marker dosage to each marker, χ^2 tests, binomial confidence intervals and mixture models were fitted via MCMC as outlined in sections "The χ^2 test", "The binomial confidence interval method", "A mixture model for assessing dosage", respectively. Mixture models were fitted under four scenarios. Priors were chosen as either vague or informative priors as outlined in Appendix 1. Variances (and hence precisions) of the normal mixture components on the logit scale were set to be either free to vary $(\tau_1, \tau_2 \text{or} \tau_3)$ or to be all the same $(\tau = \tau_1 = \tau_2 = \tau_3)$. The numbers of markers allocated as simplex, duplex or triplex are shown in Table 2. No quadruplex markers were allocated by any method.

The observed segregation proportions for 566 AFLP markers inherited from KQ99-1410, fitted mixture model and theoretical distributions on the logit scale are shown in Fig. 10. The mixture model incorporated prior information and set component variances to be equal since this was shown to correctly allocate more markers with lower misclassification rates than other models in the presence of strong overdispersion which is evident here.

Two features are apparent from Fig. 10. First, the observed segregation proportion distribution is moved to the left compared to the theoretical expectations. The expected values are (0.50, 0.79, 0.93) for simplex, duplex and triplex markers, respectively but the estimates are (0.03, 0.09, 0.07) lower (see Table 3). It should be noted that differences are magnified on the logit scale. Secondly, as expected the best estimated distribution is the simplex component. This distribution has variance wider than that of the theoretical distribution. Thirdly, even though strong priors were employed, these appeared to not unduly influence the model fit since the fitted distributions were shifted to the left since these fit the data rather than reflect the priors.

MCMC convergence was achieved after a burn in of 5,000 iterations and with a sample of 10,000 as judged by

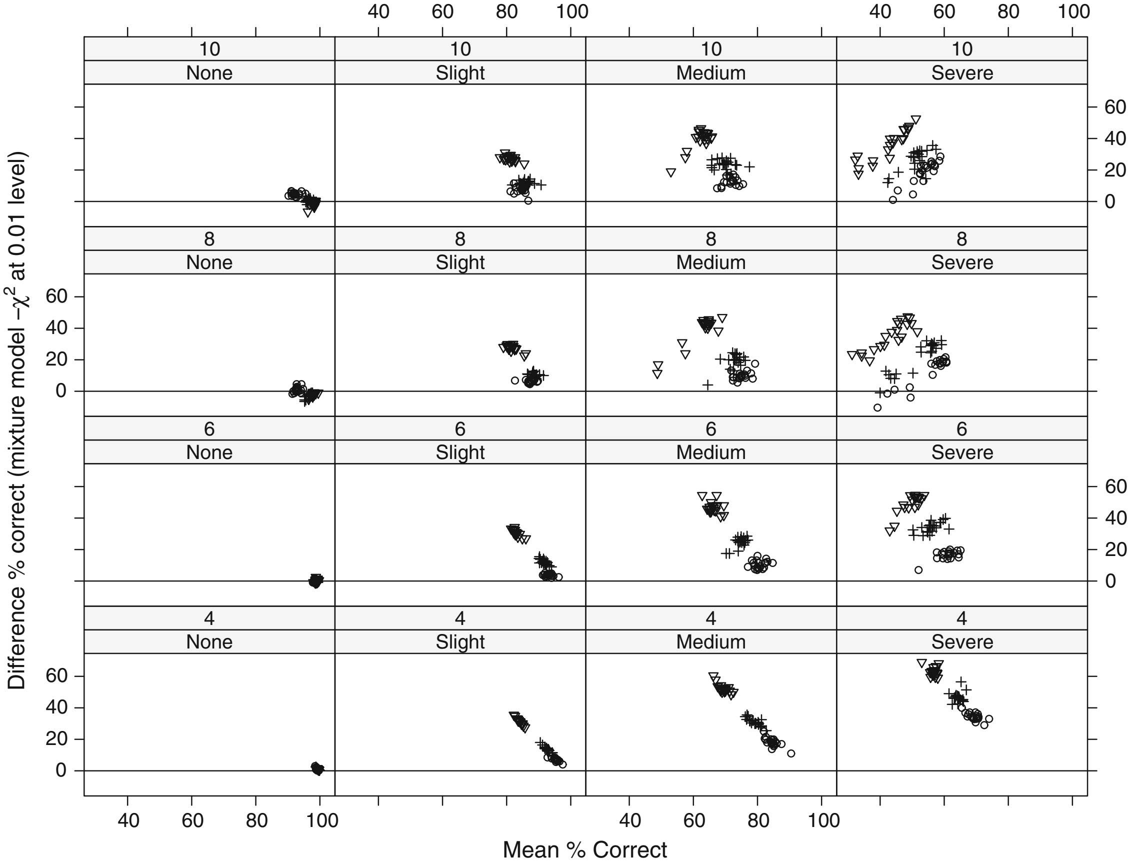

Fig. 8 Tukey mean difference plots comparing the percentage of markers correctly classified for a $\chi^2_{0.01}$ test with a mixture model employing strong prior information and equal variances. Data sets where both methods are equivalent fall on the *horizontal line*. The majority of points are above the reference line. Each such point represents one simulated data set and indicates the mixture method correctly allocates more marker dosages. The plots are conditional on the ploidy level (4, 6, 8, 10) and the level of overdispersion (*None* to *Severe*). The no. of progeny (100, 200, 500) are shown as a *inverted triangle*, *plus sign* and a *open circle*, respectively

the Geweke (1992) statistic and the diagnostics of Raftery and Lewis (1992) and Heidelberger and Welch (1983). Parameter estimates of models incorporating prior information corresponding to the mean of the posterior distributions are shown in Table 3. The deviance information criteria are shown in Table 2. The model, which proved to be best in "Choosing the best mixture model approach" had component means closer to the theoretical ones.

Discussion

Assigning marker dosage is the first analytical step in constructing molecular marker maps or QTL linkage analysis in autopolyploids. Incorrectly assigning dosage can result in marker map distortion. Allocating marker dosage to fewer markers than the data warrants may also significantly decrease the efficiency of the statistical analysis. Improvements to methods for assigning dosage should have the opposite effect by increasing efficiency.

Standard methods for assigning marker dosage appear to have limitations in that these approaches do not allow for multiple testing or overdispersion. Extra-binomial variation or overdispersion may be induced by correlation or measurement error. Since linked markers are correlated, overdispersion might be expected as a matter of course. Overdispersion may also be expected in sugarcane due to aneuploidy and irregular ploidy. For instance, since sugarcane is not a regular octoploid, different double-dose markers may come from homology groups with different ploidies and so have similar but different expected segregation ratios. Chromosome loss may also result in a change of segregation ratio from the expected value. Of course, if a marker is linked to a lethal gene say then the resulting segregation distortion may yield a lower than expected segregation ratio, whereas the opposite may occur if the marker is linked to a favourable gene. Another potential cause of overdispersion could be that of preferential pairing. In terms of allocating marker dosage, it is clear that preferential pairing will not alter segregation proportions

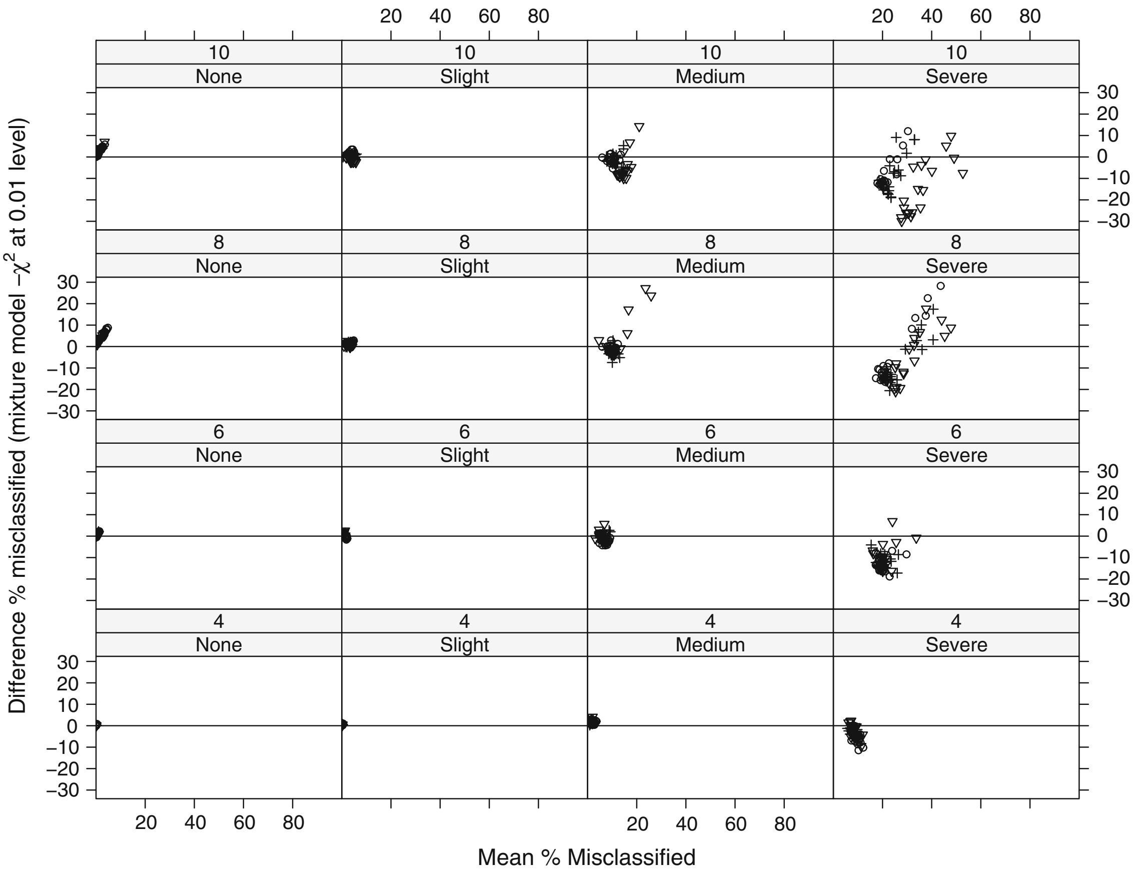

Fig. 9 Tukey mean difference plots comparing the misclassification rates for a $\chi^2_{0.01}$ test with a mixture model employing strong prior information and equal variances. Data sets where both methods are equivalent fall on the *horizontal line*. Points above the line indicate the mixture method is worse in that more markers are misclassified. More points fell below the line for increasing overdispersion although results were more variable as ploidy increased

substantially for octoploids or higher ploidy levels and so the methods presented here should correctly identify dosage.

The χ^2 test at the 5% level is commonly used to allocate dosage to a panel of markers. Alternatively, employing a binomial confidence interval should result in most markers being allocated a similar dosage. If there are no departures from theoretical assumptions then 5% of markers, which are clearly in the tails of the specific dose distributions, will remain unclassified and so a sparser marker map may be produced than if all markers were used. Employing a smaller α level should result in more markers being correctly classified except there may be some ambiguity at higher doses due to overlapping distributions. Nevertheless, simulation studies in "Comparison of current methods" showed that more markers will be correctly classified for $\alpha = 0.01$ under a range of conditions such as increasing ploidy or overdispersion. Also, in general, misclassification rates were not increased by reducing the α level.

From the simulation studies in section "Simulation study", it appears that current methods for assigning dosage work well when there is no overdispersion. Fewer markers are classified and misclassification rates increase with increasing overdispersion or ploidy. The number of progeny or markers did not affect the rates of correct allocation or misclassification for the range of conditions studied here.

The Bayesian approach presented above augments current methods for assigning marker dosage in autopolyploids by estimating marker dosage for all markers simultaneously. This method directly estimates the posterior probabilities of each dosage for every marker, incorporating prior genetic knowledge, allowing for extra-binomial variation and finally for providing a framework for assessing departures from genetic theory by comparing observed and expected parameter values. The drawbacks of the method centre around the fact that it is more complicated than current tests and so more computing power and consideration of further issues like assessment of convergence and model choice are required.

From the simulation studies in "Simulation study", it would appear that incorporating strong prior knowledge

Table 2 Number of markers assigned a dosage by various methods for 566 AFLP markers inherited from KQ99.1410 in a sugarcane trial at Ayr, Queensland

Standard methods		Marker dosage				
Method	α	1	2	3	Not	% Allocated
χ^2	0.05	301	45	12	208	63
χ^2	0.01	363	56	16	131	77[a]
Binomial	0.05	308	44	13	201	64
Binomial	0.01	363	56	17	130	77
Mixture model		**Marker dosage**				
Prior	τ_k	1	2	3	Not	% Allocated
Vague	$\neq$	425	67	42	32	94
Vague	$=$	389	53	17	107	81
Strong	$\neq$	425	56	42	43	92
Strong	$=$	430	71	32	33	94[a]

χ^2 tests, binomial CIs with type I error α and mixture models with a posterior probability threshold of 0.8 were used to allocate dosage. The mixture models employed priors that were either vague or informative based on $\tau_k(1 \pm 20\%)$ with precision $\tau_k = 1/s_k^2$.

[a] Best methods in simulation study

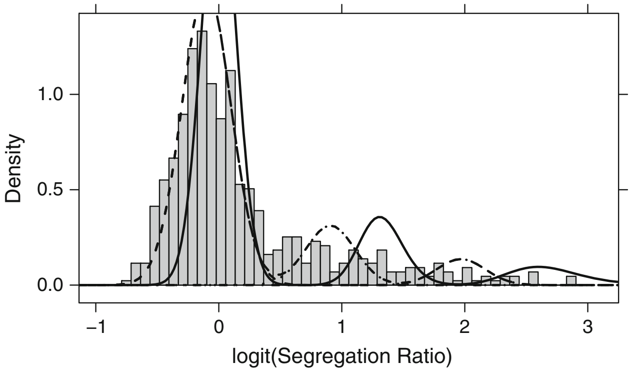

Fig. 10 *Histogram* of observed segregation ratios for 566 AFLP markers inherited from KQ99-1410 in a sugarcane trial at Ayr, Queensland. The *solid line* is the expected theoretical distribution while the *dashed line* is a mixture model fitted on the logit scale. The mixture model employed strong priors on the means μ_k and precisions τ_k. The largest component, consisting of simplex markers theoretically centred around 0, exhibits a larger variance than the theory would suggest. All components exhibit a shift to the left which may be due to aneuploidy

Table 3 Parameter estimates and standard errors in brackets on the logit scale for mixture models employed to assign marker dosage for 566 AFLP markers inherited from KQ99.1410 in a sugarcane trial at Ayr, Queensland

Parameter	$\sigma = \sigma_k$	$\sigma \neq \sigma_k$	Expected
$\hat{\pi}_1$	0.78 (0.02)	0.76 (0.02)	
$\hat{\pi}_2$	0.16 (0.02)	0.14 (0.02)	
$\hat{\pi}_3$	0.07 (0.01)	0.10 (0.02)	
$\hat{\mu}_1$	−0.10 (0.01)	−0.11 (0.01)	0.0
$\hat{\mu}_2$	0.91 (0.04)	0.78 (0.06)	1.3
$\hat{\mu}_3$	1.97 (0.07)	1.74 (0.10)	2.6
$\hat{\sigma}_1$	0.20 (0.01)	0.17 (0.01)	
$\hat{\sigma}_2$		0.19 (0.02)	
$\hat{\sigma}_3$		0.36 (0.05)	

Priors were informative based on τ (1 ± 20%) with unequal or equal variances. Component means μ_k were less than theory suggests

and equal variances on the logit scale for all components produces better allocations of marker dosage. However, as with standard methods, if there is medium to severe overdispersion or if the ploidy level is over 6 then allocation will be less accurate. This is to be expected since for higher ploidy levels, markers with dosage over 3 will not be distinguishable on the basis of segregation ratios since the theoretical distributions are so similar and so overlap. In the presence of overdispersion, this situation will be even worse since potentially the underlying distributions will be even wider.

Perhaps surprisingly, the number of markers (from 200 to 1,500) yielded very similar results in the simulation study and so no direct comparisons based on marker numbers have been presented. This is to be expected for conventional methods where tests are carried out separately but it is interesting that the number of markers had no bearing on the results employing the Bayesian mixture model.

In terms of correctly identifying marker dosage, both the χ^2 and mixture model methods produce similar results where no overdispersion is present. However, misclassification rates tend to be slightly higher with the mixture model as ploidy increases. On the other hand if any overdispersion is present, mixture models are to be preferred since, more markers in general are classified and correctly assigned dosage. Also, in the presence of overdispersion, fewer markers are misclassified than with the χ^2 test.

While there is no omnibus statistical test for overdispersion in mixture models, a simple histogram with superimposed theoretical mixture density should prove adequate for identifying overdispersion. The theoretical mixture density may be approximated by a mixture of binomial distributions with probability set to the appropriate theoretical values and proportions set to those observed by applying the χ^2 test. Additionally, graphical methods enable quick identification of outliers. Additionally, the Bayesian mixture method provides a more objective mechanism to investigate departures from the expected distributions by comparing the posterior distributions of the mean segregation proportions for each dosage component to their expected values in Eq. 1.

Extra-binomial variation or overdispersion is evident in the case study outlined in "Case study: a sugarcane AFLP data set". It also appears that, on average, the sugarcane markers appear to have smaller segregation ratios than expected. Given that the parent KQ99-1410 resulted from a cross of a *S. officinarum* IJ76-514 and a commercial sugarcane Q165, it is perhaps not surprising since there may be a number of unknown processes occurring like $2n + n$ transmission, chromosome loss or aneuploidy. Cytological studies rather than segregation ratio studies may be required to investigate this hypothesis further.

Finally, it should be noted that fitting finite mixtures is not as straight forward as more common statistical techniques like linear models and so care must be taken. Since analytic solutions are not available, MCMC is employed for computation. The EM algorithm Dempster et al. (1977) could also be used directly but this would prove less straightforward if prior information were incorporated which is the case here. Care must also be taken to ensure convergence is achieved.

The mixture model approach developed in this paper, along with standard tests have been implemented as an R (R Development Core Team 2007) package with MCMC carried out by calling JAGS (Plummer 2005). The R packages polySegratio and polySegratioMM for standard and mixture model approaches, respectively are available from the corresponding author on request and also at http://www.r-project.org/

In summary, Bayesian mixture models add a new method to assign marker dosage in autopolyploids. These methods possess the advantages that they produce estimates of the posterior probability of dosage for each marker, can objectively allocate more markers than current methods under departures from genetic theory and also provide a tool to highlight, and to a limited extent, quantify departures from the theory.

Acknowledgments We thank Jingchuan Li for assistance with AFLP marker data production and Ross Darnell and Kerrie Mengersen for useful statistical discussions. This work was partially funded by the Cooperative Research Centre for Sugarcane Industry Innovation through Biotechnology.

Appendix 1: Informative prior specification for τ_k

Conjugate prior distributions are employed for the means μ_k and precisions $\tau_k, k = 1, \ldots, K$. A method to determine the hyperparameters as $\mathrm{logit}(P_k)$ and T_k for the prior distribution of the mean μ_k are outlined in "Priors". Prior distributions for the means μ_k in autooctoploids are provided in Table 4.

Table 4 Expected segregation ratios for autooctoploids on the logit scale and hyperparameters set for strong priors assuming theoretical segregation ratios

K	Segregation ratio			
	p	logit (p)	s = SD (logit (p))	$\tau = 1/s^2$
1	1/2	0.000	0.17	33.36
2	11/14	1.299	0.41	5.85
3	13/14	2.565	1.00	1.00
4	69/70	4.234	1.15	0.76

The priors on the means are $\mu_k\ N\left(\mathrm{logit}\left(P_k, T_k^{-1}\right), k = 1, \ldots, 4\right)$ and on the precisions are $\tau \sim \mathrm{Gamma}(A_k, B_k)$ where $\tau_k = 1/\sigma_k^2$. Hyperparameters are set by by constructing a 95% CI and logit transforming this interval to obtain approximate precisions on the logit scale. Means $\mu_k \pm 0.05$ on the untransformed scale while standard deviations were set to be ±20% on the logit scale

Employing a similar approach for the mean of $\tau_k \sim \mathrm{Gamma}(A_k, B_k)$, we can calculate the logit transformed 2.5 and 97.5 percentiles of the theoretical binomial distribution with parameters in Eq. 1 to obtain the expected value $\hat{\tau}_\mathrm{k}$ of τ_k. Denoting the percentiles as $q_{0.025}$ and $q_{0.975}$ on the untransformed scale then for a 95% confidence region on the logit scale

$$\tilde{s}_k \approx [\mathrm{logit}(q_{0.0975}) - \mathrm{logit}(q_{0.0925})]/4 \tag{12}$$

and so expected value $\tilde{\tau}_k = 1/\tilde{s}_k^2$,, where $\tilde{s}_k$ is defined in Eq. 12 and so may be obtained directly from the percentiles $q_{0.025}$ and $q_{0.075}$ of the binomial distribution with size equal to the number of individuals N_k and probability P_k.

The conjugate prior distribution for the precision τ_k is a Gamma(A_k,B_k) which has mean A_k/B_k and variance A_k/B_k^2. Ideally, the mean of the prior distribution would be specified as $\hat{\tau}_\mathrm{k}$ but the variance may be specified by several methods. Those considered here involve setting an interval around either $\tilde{\tau}_\mathrm{k}$ or $\tilde{s}_k$.

First, if the prior distribution is specified to have a 95% confidence region around τ_k as $\tau_k(1 \pm x)$, then the interval has length $2x\tau_k$ which is 4 SD(τ_k). A_k and B_k are obtained by equating the mean A_k/B_k and variance A_k/B_k^2 to their observed values $\hat{\tau}_\mathrm{k}$ and $2x\hat{\tau}$, respectively.

In a similar fashion, if the 95% confidence region around s_k is set to $s_k(1 \pm x)$ then the SD(τ_k) is a quarter of the interval on the on the τ_k scale. This is simply

$$\mathrm{SD}(\tau_k) = \frac{1}{4}\left(\frac{1}{s_k^2(1-x)^2} - \frac{1}{s_k^2(1+x)^2}\right) \times \frac{x}{s_k^2(1+x)^2(1-x)^2}$$

and once the observed and expected means and variances are equated then $A_k = C.\hat{\tau}_k^4$and $B_k = C\hat{\tau}_k^3$ where $C = x^2/(1+x)(1-x)$. This produces a narrower prior distribution than specifying limits around τ_k.

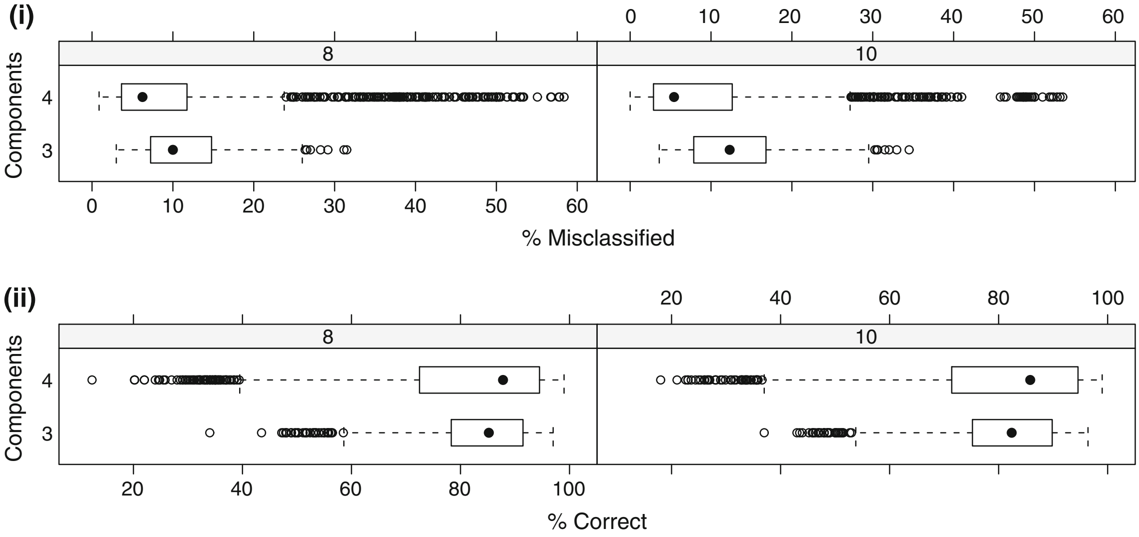

Fig. 11 Box plots of the **i** percentage of markers with dosage correctly allocated and **ii** misclassified by models with three or four components where equal variances on the logit scale were assumed and strong prior information was incorporated. Three component models may avoid computational problems but could result in more markers being misclassified

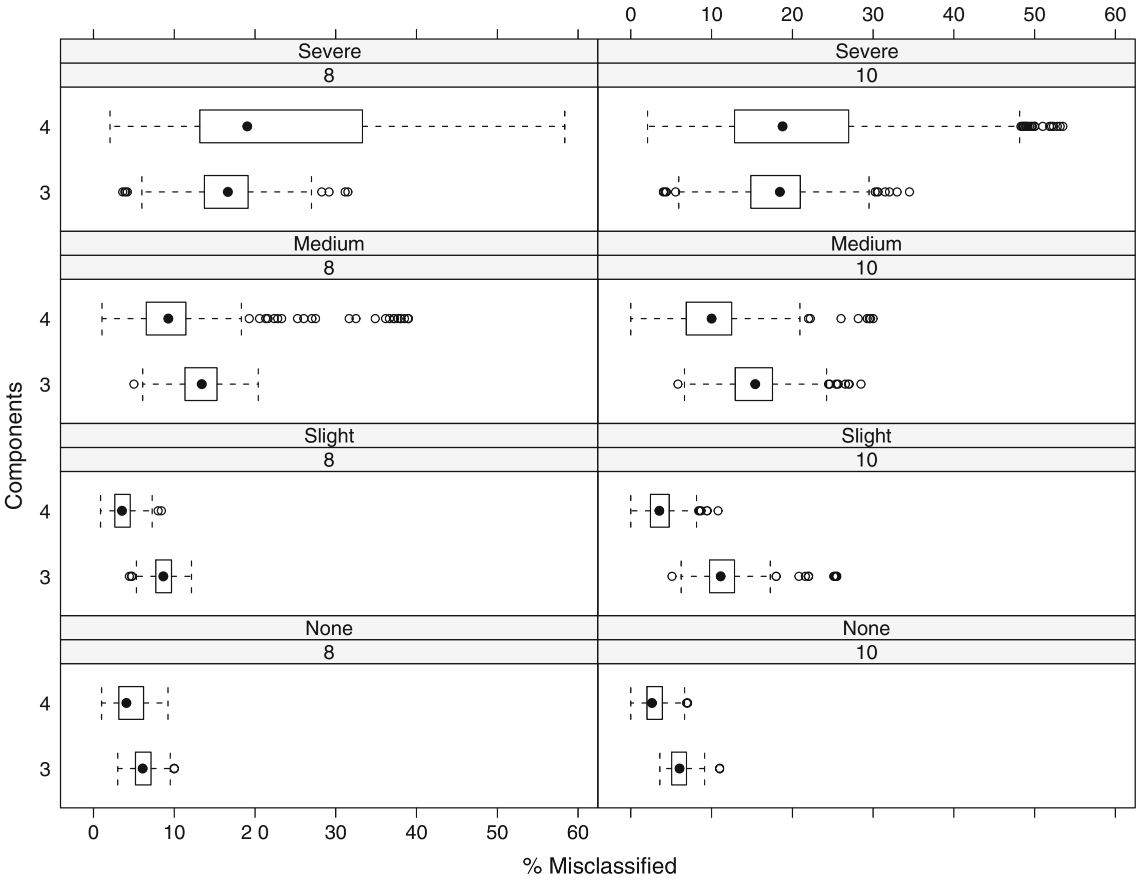

Fig. 12 Box plots of the percentage of misclassified markers for non to severely overdispersed data for models with three or four components. The model employed was that of equal variances on the logit scale with strong prior information incorporated. The range of results increases with increasing overdispersion which could result in worse classification for some data sets

Appendix 2: Generating markers with overdispersion

For simulation studies, markers with specified dosage may be generated from a Binomial (n,p) distribution where p is the appropriate segregation proportion in Eq. 1. When overdispersion is present, a beta-binomial distribution may be used with $p \sim$ beta (α,β), where α and β are the first and second shape parameter, respectively (Skellams 1948). If the theoretical segregation ratio P_{jk} in Eq. 1 is equated to the expected value $E(p) = \alpha/(\alpha + \beta)$ then simply setting the first shape parameter α fixes the value of $\beta = \alpha(1- P_{jk})/P_{jk}$. Note that larger values of α correspond to smaller values of Var $(p) = ab(a+b)^2(a+b+1)$ which results in less overdispersion.

Appendix 3: Comparison of mixture model options

From Fig. 11 the model with more components performs slightly better in that, on average, the median percentage of correctly allocated markers was higher with more components and the misclassification rate was lower. In general, while results are better when more components are employed at higher ploidy levels, the range may appear to indicate that worse results may actually be obtained for particular data sets. Further investigation reveals that this only occurs for medium to severe overdispersion (see Fig. 12).

While the percentage of correctly allocated markers decreases with increasing threshold (see Fig. 13), the trend becomes more pronounced with increasing overdispersion and ploidy levels. On the other hand, misclassification rates increase with smaller thresholds and increasing ploidy or overdispersion levels. While there is no clear optimal threshold value, it would seem that that a value of around 0.8 is a reasonable compromise and corresponds in some ways to the value of 0.2 which is commonly used in false discovery rate studies and commonly used as a reasonable power when designing experimental studies (Fig. 14).

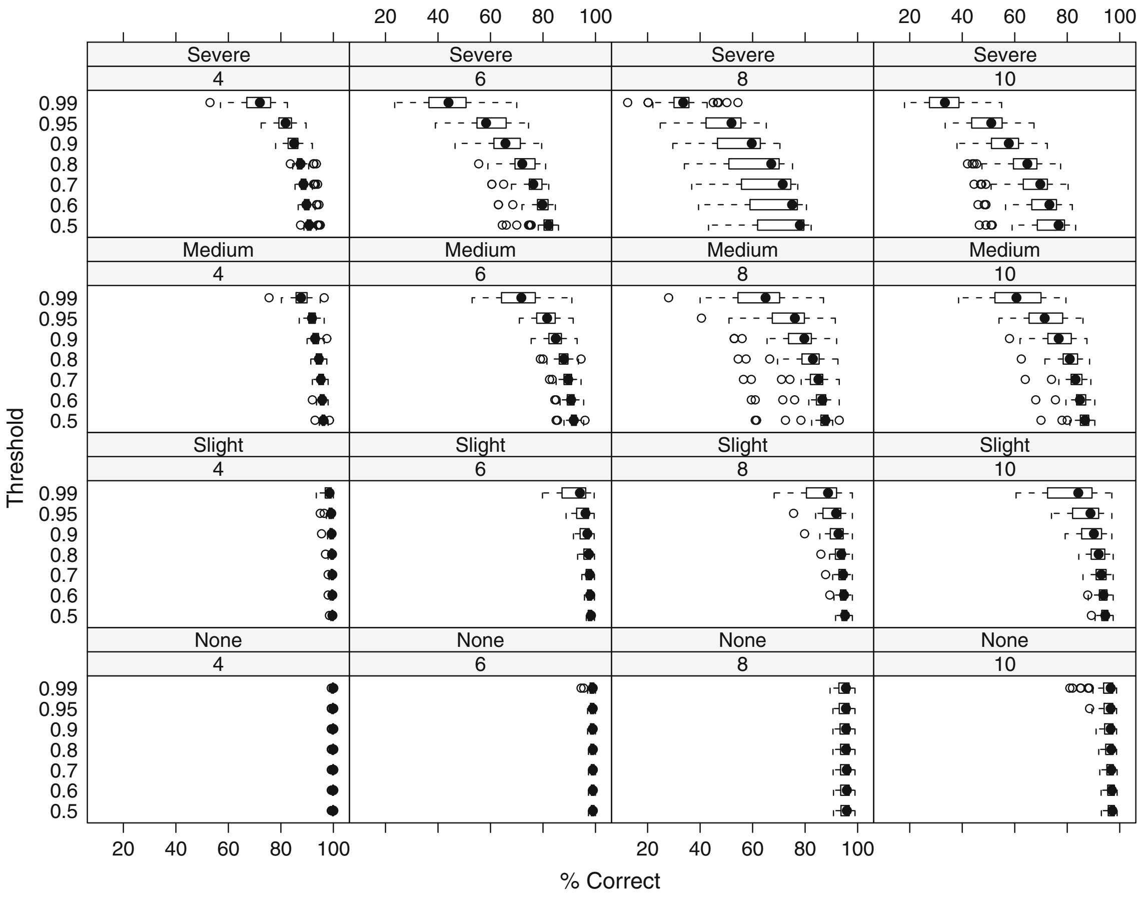

Fig. 13 Box plots of the percentage of markers with dosage correctly allocated by mixture models for a range of thresholds by four levels of overdispersion (*None, Slight, Medium, Severe*) and ploidy (4, 6, 8, 10). Dosage is allocated when the posterior probability exceeds the threshold. The models fitted were chosen to be those with the maximum number of components, equal variances on the logit scale were assumed and strong prior information incorporated. The percentage of correctly allocated markers tails off for thresholds larger than 0.8

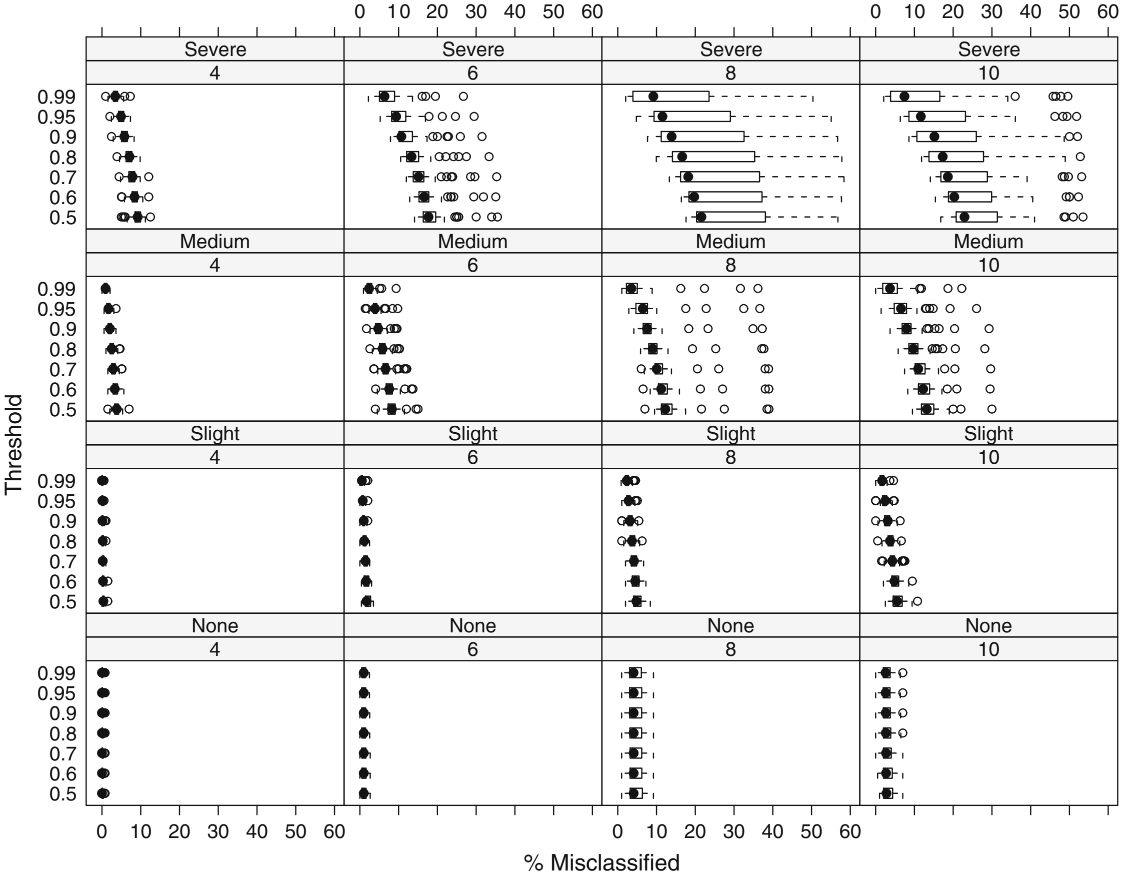

Fig. 14 Box plots of the percentage of misclassified markers for a range of thresholds with increasing overdispersion (*None, Slight, Medium, Severe*) varying ploidy levels (4, 6, 8, 10). Dosage is allocated when the posterior probability exceeds the threshold. When there is little or no overdispersion, very few markers are misclassified. However, for moderate to severe overdispersion the misclassification rate decreases as the threshold increases but is greater for higher ploidy levels

References

Aitken K, Jackson P, McIntyre C (2005) A combination of aflp and ssr markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. Theor Appl Genet 110:789–801

Aitken KS, Jackson PA, McIntyre CL (2007) Construction of a genetic linkage map for *Saccharum officinarum* incorporating both simplex and duplex markers to increase genome coverage. Genome 50:742–756

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Control AC-19:713–716

Al-Janabi SM, Honeycutt RJ, McClelland M, Sobral BW (1993) A genetic linkage map of *Saccharum spontaneum L.* 'SES 208'. Genetics 134:1249–1260

Altman DG, Bland JM (1983) Measurement in medicine: the analysis of method comparison studies. Statistician 32:307–317

Besag J, Green P, Higdon PJ, Mengersen K (1995) Bayesian computation and stochastic systems. Stat Sci 10:3–66 (with discussion)

Best N, Cowles MK, Vines K (1995) CODA convergence diagnosis and output software for Gibbs sampling output Version 0.30. MRC Biostat Unit, Cambridge

Bland J, Altman D (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 346:1085–1087

Burner DM (1997) Chromosome transmission and meiotic behaviour in various sugarcane crosses. J Am Soc Sugar Cane Tech 17:38–50

Celeux G, Forbes F, Robert C, Titterington D (2006) Deviance information criteria for missing data models. Bayesian Anal 4:651–674

da Silva JAG (1993) A methodology for genome mapping of autopolyploids and its application to sugarcane *Saccharum spontaneum*. PhD thesis, Cornell University, Ithaca, NY

da Silva JAG, Sorrells ME, Burnquist WL, Tanksley ST (1993) RFLP linkage map and genome analysis of *Saccharum spontaneum*. Genome 36:782–791

da Silva J, Honeycutt RJ, Burnquist W, Al-Janabi SM, Sorrells ME, Tanksley SD, Sorbral BWS (1995) *Saccharum spontaneum* L. SES 208 genetic linkage map combining RFLP-and PCR-based markers. Mol Breed 1:165–179

De Winton D, Haldane JBS (1931) Linkage in the tetraploid. J Genet 24:121–144

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

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman Hall, London

Geweke J (1992) Evaluating the accuracy of sampling based approaches to calculating posterior moments. In: Bernado JM, Berger JO, David AP, Smith AFM (eds) Bayesian statistics 4. Oxford University Press, Oxford, pp 169–194

Gilks W, Richardson S, Spiegelhalter D (1996) Markov chain Monte Carlo in practice. Chapman Hall, London

Grivet L, Arruda P (2002) Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr Opin Plant Biol 5:122–127

Hackett CA (2001) A comment on Xie and Xu: 'mapping quantitative trait loci in tetraploid species'. Genet Res 78:187–189

Hackett CA, Luo ZW (2003) TetraploidMap: construction of a linkage map in autotetraploid species. J Hered 94:358–359

Haldane JBS (1930) Theoretical genetics of autopolyploids. J Genet 22:359–372

Haley C, Knott S (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Heidelberger P, Welch P (1983) Simulation run length control in the presence of an initial transient. Oper Res 31:1109–1144

Janoo N, Grivet L, David J, D'Hont A, Glaszmann JC (2004) Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. Heredity 93:460–467

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

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

Luo ZW, Zhang RM, Kearsey MJ (2004) Theoretical basis for genetic linkage analysis in autotetraploid species. PNAS 101:7040–7045

Mao CX (2007) Estimating population sizes for capture–recapture sampling with binomial mixtures. Comput Stat Data Anal 51:5211–5219

Mather K (1951) The measurement of linkage in heredity. Methuen, London

Mengersen KL, Robert CP (1996) Testing for mixtures: a Bayesian entropic approach. In: Bernando JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian statistics 5. Oxford University Press, Oxford, pp 225–276

Meyer R, Milbourne D, Hackett C, Bradshaw J, McNichol J, Waugh R (1998) Linkage analysis in tetraploid potato and association of markers with quantitative resistance to late blight (*Phytophthora infestans*). Mol Gen Genet 259:150–160

Ming R, Liu S, Lin Y, Braga D, da Silva J, van Deynze, Wenslaff A, Wu K, Moore P, Burnquist W, Sorrells M, Irvine J, Paterson A (1998) Alignment of Sorghum and Saccharum chromosomes: comparative organization of closely-related diploid and polyploid genomes. Genetics 150:1663–1882

Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics, 3rd edn. McGraw–Hill, New York

Plummer M (2005) JAGS Version 0.90 manual. International Agency for Research on Cancer, Lyon

Qu L, Hancock JF (2001) Detecting and mapping repulsion-phase linkage in polyploids with polysomic inheritance. Theor Appl Genet 103:136–143

Qu L, Hancock J (2002) Pitfalls of genetic analysis using a doubled-haploid backcrossed to its parent. Theor Appl Genet 105:392–396

Raftery AL, Lewis S (1992) How many iterations in the Gibbs sampler? In: Bernado JM, Berger JO, David AP, Smith AFM (eds) Bayesian statistics 4. Oxford University Press, Oxford, pp 763–774

Ripol MI, Churchill A, da Silva JA, Sorrells M (1999) Statistical aspects of genetic mapping in autopolyploids. Gene 235:31–41

Robert C (1996) Mixtures of distributions: inference and estimation. In: Gilks W, Richardson S, Spiegelhalter D (eds) Markov chain Monte Carlo in practice. Chapman Hall, London

Rufo MJ, Pérez CJ, Martìn J (2007) Bayesian analysis of finite mixtures of multinomial and negative-multinomial distributions. Comput Stat Data Anal 51:5452–5466

Skellam JG (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. J R Stat Soc Ser B 10:257–261

Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov Monte Carlo Methods. J R Stat Soc Series B 55:3–23

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. PNAS 97:7051–7057

Spiegelhalter D, Thomas A, Best N, Gilks W (1995) BUGS. Bayesian inference Using Gibbs Sampling, Version 0.50. MRC Biostatistics Unit, Cambridge

Spiegelhalter DJ, Best N, Carlin B, van der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B 64:583–639

Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York

Stephens M (2000) Bayesian analysis of mixtures with an unknown number of components—an alternative to reversible jump methods. Ann Stat 28:40–74

Sybenga J (1994) Preferential pairing estimates from multivalent frequencies in tetraploids. Genome 37:1045–1055

Sybenga J (1995) Meiotic pairing in autohexaploid *Lathyrus*: a mathematical model. Heredity 75:343–350

Sybenga J (1996) Chromosome pairing affinity and quadrivalent formation in polyploids: do segmental allopolyploids exist? Genome 39:1176–1184

Tanner MA (1993) Tools for statistical inference, 2nd edn. Springer, New York

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation: with discussion. J Am Stat Assoc 82:528–550

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0

Tweedie RL, Mengersen K (1996) Rates of convergence of the Hastings and Metropolis algorithms. Ann Stat 24:101–121

Ukoskit K, Thompson PG (1997) Autopolyploidy versus allopolyploidy and low-density randomly amplified polymorphic DNA linkage maps of sweetpotato. J Am Soc Hortic Sci 122:822–828

Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet 83:294–300

Wu R, Gallo-Meagher M, Littell RC, Zeng ZB (2001a) A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. Genetics 159:869–882

Wu SS, Wu R, Ma CX, Zeng ZB, Yang MC, Casella G (2001b) A multivalent pairing model of linkage analysis in autotetraploids. Genetics 159:1339–1350

Wu R, Ma CX, Casella G (2002) A bivalent model for linkage analysis in outcrossing tetraploids. Theor Popul Biol 62:129–151

Xie CG, Xu SH (2000) Mapping quantitative trait loci in tetraploid populations. Genet Res 76:105–115