Theor Appl Genet (2009) 119:483–496
DOI 10.1007/s00122-009-1054-x

ORIGINAL PAPER

Estimation of copy number in polyploid plants: the good, the bad, and the ugly

Andrew W. George

Received: 21 December 2008 / Accepted: 24 April 2009 / Published online: 18 May 2009

Abstract Genetic studies in polyploid plants rely heavily on the collection of data from dominant marker loci. A dominant marker locus is a locus for which only the presence or absence of an observable (dominant) allele is recorded. Before these marker loci can be used for genetic exploration, the number of copies of a dominant allele carried by a parent (copy number) must be determined for each marker locus. Copy number in polyploids is estimated using a hypothesis testing procedure. The performance of this estimation procedure has never been evaluated. In this paper, I quantify whether the highly sought after single-copy markers can be accurately identified, if the performance of the estimation procedure improves with increasing sample size, and whether the estimation procedure is capable of accurately estimating the copy number of high copy markers. I found that the probability of incorrectly estimating copy number is quite low and that more data can actually reduce the accuracy of the estimation procedure when the testing assumptions are violated. Fortunately, when a significant result is obtained, it is almost always correct. The challenge often is in obtaining a significant result.

Introduction

Polyploids are living things that have more than twice the haploid number of chromosomes in the nuclei of their cells. In animals and humans, polyploidy is rare and often deleterious. However, in plants, polyploids are often superior to their diploid cousins (Leitch and Bennett 1997; Wendel 2000). Consequently, over 70% of all flowering plants are polyploids. More importantly, most of the crop plants are highly polyploid. Crops such as banana and apple are triploid; maize, potato, and tobacco are tetraploid; bread wheat is hexaploid; and strawberry and sugarcane are octoploid. For centuries, farmers have domesticated their crops through selection, but have also unknowingly been breeding higher-ploidy plants.

Much interest surrounds better understanding the genetic machinery controlling important economic traits in polyploid crops. With increased understanding comes increased opportunity to develop new breeding strategies that combine traditional breeding practices with innovative genetic technologies which greatly improve selection. However, the genetic machinery underlying polyploids can be extremely complex. In diploids each chromosome pairs with its homologous (sister) partner. In polyploids, multiple homologous and homoeologous (partially homologous) chromosomes may be competing for synapsis (Sybenga 1992, 1996) and it is a plant's often poorly known homology that determines the meiotic behaviour of its chromosomes. Furthermore, it can be difficult to collect genetic data on polyploid systems from codominant marker loci such as SNPs and microsatellites. A codominant marker locus may have a large number of possible heterozygous genotypes segregating in a population, and these genotypes cannot always be readily identified.

Genetic studies in polyploids avoid these difficulties by making use of dominant marker loci. A dominant marker locus is a locus for which only the presence or absence of an observable (dominant) allele is recorded. However, only those marker loci for which a dominant allele is

Communicated by B. Friebe.

A. W. George (✉)
Mathematical and Information Sciences,
CSIRO, Brisbane, QLD 4067, Australia
e-mail: andrew.george@csiro.au; geo047@csiro.au

segregating in a full-sib family from one parent but not the other is of use (Wu et al. 1992). Consequently, an important analysis step in polyploid studies is to estimate the copy number (also referred to as a marker dosage) of a dominant allele. All further analyses of the genetic data are based on these marker copy number estimates.

The copy number of a dominant allele is most commonly estimated in polyploids by using the unorthodox approach of hypothesis testing. Here, a series of hypothesis tests are performed. In each test, the null hypothesis is the copy number equalling a hypothesised value, and the alternative is the copy number not equalling the hypothesised value. The significance of the alternative hypothesis is measured using the chi-squared test (Dasilva et al. 1995; Grivet et al. 1996; Aitken et al. 2005). The "best" estimate of marker copy number is the value associated with the null hypothesis that is accepted. If multiple null hypotheses are accepted, an inconclusive result is obtained and the copy number of the dominant allele cannot be estimated.

In this paper, I examine the performance of the hypothesis testing procedure for the estimation of copy number in polyploid systems. This paper is motivated by the knowledge that this testing procedure has never been evaluated. Furthermore, the hypotheses for estimating the copy number of a dominant allele are constructed under the simplifying assumptions of known ploidy level and random pairing of chromosomes. However, several important polyploid crop plants have variable polyploidy across homology groups and/or exhibit non-random (preferential) chromosome pairing (Rhoades 1952; Doyle 1963; Pfosser et al. 1995; Missaoui et al. 2005). Hence, I pay special attention to the performance of the estimation procedure, when the underlying assumptions are violated. Specifically, I quantify whether the sought after single-copy alleles can be accurately identified given a wrongly specified ploidy level and/or the existence of preferential pairing; whether an increase in sample size guards against misspecification of a dominant allele's copy number; and whether the test is capable of accurately estimating high copy numbers.

Methods

In this section, I begin by describing the hypothesis testing procedure for estimating the copy number of a dominant allele in polyploid full-sibling families. I focus on those families for which only a single parent carries a dominant allele. Next, I present a new way of parameterizing the non-uniform (preferential) pairing of chromosomes and then use this parameterisation to construct a probability distribution of the segregation ratio for a dominant allele. That is, I derive the probability distribution of the proportion of progeny carrying a dominant allele conditional on ploidy level and amount of preferential pairing. I then use this probability distribution to calculate the probability of (a) correctly estimating copy number, (b) misclassifying the copy number of a dominant allele, and (c) the estimation procedure yielding an inconclusive result and hence failing to estimate copy number. I examine the performance of the test with respect to known ploidy level and no preferential pairing (the good); known ploidy level but an unknown level of preferential pairing (the bad); and unknown ploidy level and an unknown level of preferential pairing (the ugly).

Estimating marker copy number using hypothesis testing

To estimate the copy number of a dominant allele, a series of hypotheses are constructed and their significance determined using the chi-squared test. Let ω be the true unknown copy number of a dominant allele and $\hat{\omega}$ an estimate of the copy number. Here, $\hat{\omega} \in \{1, 2, \ldots, \ell/2\}$ where l is the ploidy level. The null hypotheses to be tested are $H_0 : \omega = 1$, $H_0 : \omega = 2, \ldots, H_0 : \omega = \ell/2 - 1$, and $H_0 : \omega = \ell/2$. The associated alternative hypotheses are $H_1 : \omega > 1$, $H_1 : \omega \neq 2, \ldots, H_1 : \omega \neq \ell/2 - 1$, and $H_1 : \omega < l/2$. A one-tailed test of significance is performed if $\hat{\omega} = 1$ or $\hat{\omega} = l/2$. Alternatively, a two tailed test of significance is performed if $1 < \omega < l/2$. An estimate of ω is obtained if one of the null hypotheses is accepted; otherwise the marker's copy number is unknown, as it cannot be estimated from the observed data.

Pearson's chi-squared test is used to test if there is significant evidence to accept or reject the null hypothesis. The test statistic can be written as

$$T_l(\omega_0) = \frac{(x - ns_{\omega_0})^2}{ns_{\omega_0}} \quad (1)$$

where x is the number of progeny observed to carry a dominant allele, s_{ω_0} is the expected segregation ratio for the dominant allele with a parent carrying ω_0 hypothesised copies of a dominant allele, and n is the total number of progeny. The test statistic $T_l(\omega_0)$ closely follows a chi-squared distribution with one degree of freedom. The test is performed assuming the ploidy level is known and no preferential pairing. Under these assumptions, measuring the significance of the null hypothesis is equivalent to assessing the strength of evidence that the observed number of progeny carrying a dominant allele follows a binomial distribution with probability of success s_{ω_0}. Ripol et al. (1999) give the expected segregation ratios, assuming an absence of preferential pairing, for various polyploids.

Preferential pairing

The pairing affinity of chromosomes during meiosis is measured in tetraploids by the preferential pairing factor p (Wu et al. 2002). Suppose a tetraploid has two pairs of homologous chromosomes, 1_A 2_A and 3_B 4_B where chromosomes with the same subscripts are homologous. There are three unique chromosome pairing configurations (Ψ) that may occur during meiosis; 1_A pairs with 2_A and 3_B pairs with 4_B (Ψ_1), 1_A pairs with 3_B and 2_A pairs with 4_B (Ψ_2), and 1_A pairs with 4_B and 2_A pairs with 3_B (Ψ_3). Here, the two bivalents in Ψ_1 are between homologous chromosomes and the two bivalents in Ψ_2 and Ψ_3 are between homoeologous chromosomes. If there is no preferential pairing, each of these configurations occurs during meiosis with equal frequency. However, if the founding genomes A and B are genetically diverse, there will be a strong pairing affinity between homologous chromosomes and Ψ_1 will predominantly occur during meiosis. The probabilities of the pairing configurations are $13 + p$ for Ψ_1, $13 - 12p$ for Ψ_2 and $13 - 12p$ for Ψ_3 where p is a deviation from random chromosomal pairing, the deviations sum to zero, and the configuration probabilities sum to one. As p approaches 2/3, pairing occurs predominantly between homologous chromosomes (allopolyploidy) and as p approaches 0, pairing occurs randomly between all chromosomes (autopolyploidy).

A limitation of this parameterisation is that it does not extend to higher ploidy systems. For example, suppose we have a hexaploid with homologous chromosomes 1_A 2_A, 3_B 4_B, and 5_C 6_C. Then, during meiosis, there are 15 unique chromosome pairing configurations for which eight configurations contain only homoeologous bivalents, six contain two homoeologous bivalents and one homologous bivalent, and one configuration contains only homologous bivalents. Analogous to the parameterisation used in tetraploids, we could assign probability $115 + p$ to the pairing configuration containing only homologous bivalents and $115 - 114p$ to the remaining configurations. However, equal weight is given to configurations containing a homologous bivalent and configurations containing only homoeologous bivalents. An additional preferential pairing parameter, q say, could be introduced, but it is unclear how this can be done while still ensuring the deviations sum to zero.

Instead, I account for preferential pairing via the probability of a configuration containing m homologous bivalents (p_m). To demonstrate its use, let us return to our hexaploid example. Here, p_0 is the probability of a pairing configuration occurring with no homologous bivalents, p_1 is the probability of a configuration containing a single homologous bivalent, and p_3 is the probability of a configuration containing three homologous bivalents. Since a pairing configuration containing two homologous bivalents and a homoeologous bivalent is not possible in a hexaploid system with three pairs of homologous chromosomes, p_2 is not considered here. The probabilities have a natural ordering: $p_3 \geq p_1 \geq p_0$. If the probabilities are approximately equal, there is little preferential pairing and the chromosomes are pairing uniformly. If p_3 is approximately one, then pairing is occurring only between homologous chromosomes and there is near complete preferential pairing.

Distribution of segregation ratios for a dominant allele

Suppose data are collected from a full-sib family with n progeny. The marker phenotype is a dichotomous trait for which only the presence or absence of a dominant allele is observed. Dominant alleles are assumed to segregate from only one parent. Then it can be shown ("Appendix 1") that for a dominant marker locus with ω copies of a dominant allele carried by a parent of ploidy level l, the probability distribution of the segregation ratio s is:

$$\Pr_{l,\omega}(s|\mathbf{p}) = \sum_{j=1}^{m} w_j \mathrm{Bin}\left(ns\,;n\,,\pi_j\right) \qquad (2)$$

where m is the number of mixture components, w_1, w_2,…, w_m are mixing weights satisfying the conditions $w_j \geq 0$ and $\sum w_j = 1$, Bin($ns;\ n,\ \pi_j$) is the binomial probability of observing ns progeny exhibiting a dominant allele in a family with n progeny, and π_j is the probability of a progeny exhibiting a dominant allele given the preferential pairing probabilities $\mathbf{p} = (p_0, p_1, \ldots, p_{\ell/2-2}, p_{\ell/2})$.

When an absence of preferential pairing is assumed, Eq. 2 collapses into a single binomial probability distribution with π being the standard expected segregation ratio (see Ripol et al. 1999). Equation 2 also collapses into a single binomial probability with $\pi = 0.5$ when $\omega = 1$. That is, the probability distribution of the segregation ratio for a single-copy allele is independent of the amount of preferential pairing and ploidy level. For multiple-copy alleles, the probability distribution of s follows a mixture distribution (Table 1) where the number of component mixtures is determined by the copy number. For an example of how the probability distribution of the segregation ratio for a multiple-copy allele is constructed, see "Appendix 2".

Measuring the performance of the copy number estimation procedure

The copy number of a dominant allele in polyploids is estimated by performing a series of one- and two-tailed Pearson's chi-squared tests. These tests may result in the

Table 1 Probability distribution of the segregation ratio for a dominant marker locus given the copy number ω and vector of preferential pairing probabilities **p**

Copy number	Mixture distribution			
	Mixture component (j)	Mixing Weights (w_j)	Binomial probabilities (nc $\times$ π_j)	
Tetraploid ($l = 4$)				
2	1	0.33	$a + 4b$	$a = 2p_0$, $b = p_0 + p_2$
	2	0.07	$2a + 3b$	nc $= 2a + 4b$
Hexaploid ($l = 6$)				
2	1	0.2	$8a + 8b$	$a = 4p_0 + 2p_1$
	2	0.8	$10a + 6b$	$b = 2p_0 + 3p_1 + p_3$
3	1	0.4	$12a + 7b$	nc $= 12a + 8b$
	2	0.6	$11a + 8b$	
Octoploid ($l = 8$)				
2	1	0.14	$3a + 36b + 16c$	$a = 24p_0$
	2	0.86	$5a + 38b + 12c$	$b = 14p_0 + 8p_1 + 2p_2$
3	1	0.43	$5a + 44b + 16c$	$c = 9p_0 + 8p_1 + 6p_2 + p_4$
	2	0.57	$6a + 45b + 14c$	nc $= 6a + 48b + 16c$
4	1	0.09	$5a + 48b + 16c$	
	2	0.23	$6a + 48b + 15c$	
	3	0.69	$6a + 47b + 16c$	
Decaploid ($l = 10$)				
2	1	0.11	$36a + 128b + 32c$	$a = 96p_0 + 24p_1$
	2	0.89	$48a + 124b + 24c$	$b = 64p_0 + 42p_1 + 12p_2 + 2p_3$
3	1	0.33	$51a + 148b + 32c$	$c = 44p_0 + 45p_1 + 20p_2 + 10p_3 + p_5$
	2	0.67	$57a + 146b + 28c$	nc $= 60a + 160b + 32c$
4	1	0.05	$54a + 160b + 32c$	
	2	0.38	$60a + 156b + 30c$	
	3	0.57	$58a + 156b + 32c$	
5	1	0.13	$60a + 160b + 31c$	
	2	0.24	$59a + 160b + 32c$	
	3	0.64	$60a + 159b + 32c$	

In constructing these theoretical probability distributions, I have assumed that only one of the parents carries ω copies of the dominant allele. nc is the normalizing constant

true copy number of an allele being estimated, an incorrect estimate of the underlying copy number, or an inconclusive result and unknown copy number for the dominant allele. An obvious measure of performance is the probability with which each of these events is expected to occur. Specifically, I calculate the outcome probability of a hypothesis under the true ω and l, conditional on a specific set of preferential pairing probabilities. However, this measure is dependent upon the vector of preferential pairing probabilities, and hence is multidimensional and not easily reported for polyploids with high ploidy levels. Instead, I reduce the dimensionality of this measure by averaging over $p_0, p_1, \ldots, p_{\ell/2-2}$. The expected outcome probability is then the probability of an event, under the true copy number and ploidy level, conditional on $p_{\ell/2}$.

Now, I discuss the calculation of the outcome probability. First, I obtain the critical values for the hypothesis tests at a 5% significance level. Then, using an algebraic rearrangement of Eq. 1 where x is made the subject, I calculate the corresponding critical segregation ratio values. These critical segregation ratios are the boundaries for the acceptance/rejection regions of the hypothesis tests. Second, Eq. 2 is used to derive the probability distribution of the segregation ratio, under a specified ω and l, given **p**. Third, this probability distribution is used to calculate the outcome probability of a copy number estimate by computing the area of the distribution bounded by the associated acceptance region.

As an example, suppose marker data are observed on a nuclear family with 100 progeny, and we assume that the

progeny are hexaploids. Let us assume that a single parent carries two copies of a dominant allele and the amount of preferential pairing is $\mathbf{p} = (0.05, 0.15, 0.8)$. The upper critical segregation ratio associated with testing the significance of $H_0{:}\omega = 1$ and $H_0{:}\omega = 2$ is 0.6 and 0.88, respectively. The lower critical segregation ratio associated with testing the significance of $H_0{:}\omega = 2$ and $H_0{:}\omega = 3$ is 0.72 and 0.91, respectively. The probability distribution of the associated segregation ratio for this marker is $\Pr_{\ell=6,\omega=2}(s|\mathbf{p}) = 0.2\text{Bin}(100s; 100, 0.88) + 0.8\text{Bin}(100s; 100, 0.78)$ where the mixture weights and binomial probabilities are calculated using Table 1. From this distribution, the probability of incorrectly classifying the dominant allele as a single-copy allele is zero; the probability of correctly estimating the allele's copy number is 0.8; the probability of misclassifying the allele as a triple copy allele is 0.05; and the probability of the estimation procedure yielding an inconclusive result is 0.15.

Results

In this section, I present the findings from evaluating the performance of the copy number estimation procedure. I compute the probability of correctly/incorrectly estimating the copy number of a dominant allele and of the estimation procedure yielding an inconclusive result. The performance of the estimation procedure is measured under three different scenarios where for each scenario, the family size, ploidy level, and amount of preferential pairing are varied. In the first scenario, I assume an absence of preferential pairing of the chromosomes during meiosis (the good). There is only one set of values for **p** because under no preferential pairing all preferential pairing probabilities are equal. Hence, since the outcome probability is univariate, I do not need to compute its expected value. In the second scenario, I assume that the ploidy level is correctly known, but the amount of preferential pairing is unknown (the bad). Here, I compute expected outcome probabilities. I then plot these probabilities against $p_{l/2}$ and investigate the impact of preferential pairing on the estimation procedure. In the third scenario, the ploidy level and amount of preferential pairing are unknown (the ugly). Although I could compute expected outcome probabilities for different underlying ploidy levels, I have focused on octoploids and investigated the effect of wrongly assuming a hexaploid or decaploid. This situation together with preferential pairing has been observed in sugarcane.

Known ploidy and no preferential pairing (The good)

For single- and double-copy alleles, the probability of correctly estimating their copy number is approximately 95% (Table 2). For higher copy alleles, sample sizes of 250 or even 500 are required to achieve the same level of accuracy. Reassuringly, the probability of misclassifying an allele's copy number remains extremely low across ploidy levels for single-, double-, and triple-copy alleles. The probability of misclassifying an allele's copy number begins to increase only for quadruple- and quintuple-copy alleles. Also, as would be expected, estimating the copy number of an allele for which a parent is carrying five copies is difficult, because the probability of the estimation procedure yielding an inconclusive result is very high. See Table 2 for results.

Known ploidy level and unknown preferential pairing (the bad)

The expected outcome probabilities of correctly estimating copy number are shown in Fig. 1. Each plot contains three curves corresponding to the expected outcome probabilities for a family with 100, 250, and 500 progeny. Results for a single-copy allele are not shown. The probability of correctly estimating a single-copy allele remains at 95%, since it is independent of ploidy level and amount of preferential pairing.

From Fig. 1, preferential pairing clearly affects the accuracy of the estimation procedure. In tetraploids and hexaploids, the probability curves decrease substantially with increasing amounts of preferential pairing. In higher polyploids such as octoploids and decaploids, at least for double-copy alleles, the probability curves are more stable. Also, note the behaviour of the probability curves with respect to family size. For a double-copy allele, an increase in family size actually leads to less accurate results. However, the opposite is true when estimating the copy number of quadruple- and quintuple-copy alleles.

Fortunately, although in some cases the ability of the chi-squared test to correctly estimate an allele's copy number may be poor, this does not imply that the probability of misclassifying an allele's copy number is high. It is only for extreme levels of preferential pairing that the probability of misclassifying a dominant allele's copy number begins to increase and then only in hexaploids and octoploids. See Fig. 2 for details. The challenge in using hypothesis testing for copy number estimation is that the probability of an inconclusive result can be high. This probability increases sharply with increasing amounts of preferential pairing (Fig. 3). It is also apparent that collecting data on large families does not always reduce the probability of an inconclusive result.

Unknown ploidy levels and unknown preferential pairing (the ugly)

Here, I reveal the performance of the estimation procedure for the most challenging scenario, unknown ploidy level

Table 2 Outcome probabilities ($\times 10^2$) associated with the copy number estimation procedure

True copy number	n[a]	Estimated copy number																	
		1				2				3			4		5	Unknown			
		Ploidy Level																	
		4	6	8	10	4	6	8	10	6	8	10	8	10	10	4	6	8	10
1	100	96	96	96	96	0	0	0	0	0	0	0	0	0	0	4	4	4	4
	250	96	96	96	96	0	0	0	0	0	0	0	0	0	0	4	4	4	4
	500	95	95	95	95	0	0	0	0	0	0	0	0	0	0	6	6	6	6
2	100	0	0	0	0	94	94	95	95	0	0	1	0	0	0	6	6	6	4
	250	0	0	0	0	95	95	96	95	0	0	0	0	0	0	5	5	5	5
	500	0	0	0	0	95	95	95	95	0	0	0	0	0	0	5	5	5	5
3	100		0	0	0		0	1	2	93	91	81	2	1	0		7	6	6
	250		0	0	0		0	0	0	95	95	95	0	0	0		5	5	5
	500		0	0	0		0	0	0	95	95	95	0	0	0		5	5	5
4	100			0	0			0	0		6	3	60	26	1			34	70
	250			0	0			0	0		0	1	91	90	1			9	8
	500			0	0			0	0		0	0	95	95	1			5	4
5	100				0				0					0	5				95
	250				0				0					8	37				55
	500				0				0					2	82				16

The ploidy level is known and chromosomes pair uniformly. The hypothesis tests are performed at the 5% significance level

[a] The number of progeny

and unknown levels of preferential pairing. For ease of explanation, I focus on octoploids and examine the impact on copy number estimation when I assume wrongly a hexaploid or decaploid. In Fig. 4, the expected outcome probabilities of correctly estimating an allele's copy number are shown. For double- and triple-copy alleles, each plot contains three probability curves corresponding to the expected outcome probabilities for hexaploids, octoploids, and decaploids. Since I would test for a quadruple-copy allele only if I believe the ploidy level to be greater than a hexaploid, plots for a quadruple-copy allele have two probability curves corresponding to the expected outcome probabilities for octoploids and decaploids.

From Fig. 4, we see that for a double-copy allele, misclassifying the ploidy level does not adversely affect the performance of the estimation procedure. In fact, incorrectly assuming a decaploid can increase the probability of correctly identifying an allele's copy number. However, the adverse affect of misclassifying a polyploid's ploidy level is evident when estimating triple- and quadruple-copy alleles.

Fortunately, even though the underlying assumptions are being grossly violated, the probability of misclassifying an allele's copy number is still reasonably low (Fig. 5). It is only for triple-copy alleles that the probability of misclassifying their copy number can be as high as 20%. Even then, this occurs only when there is a high level of preferential pairing, and one wrongly assumes the octoploid to be a hexaploid.

As we saw in the previous section, the weakness in using a series of hypothesis tests to infer a dominant allele's copy number is that the estimation procedure can yield an inconclusive result. As the observed data moves further away from the underlying test assumptions, the probability of an inconclusive result increases. This phenomenon is clearly evident in Fig. 6. For double-copy alleles, misclassifying the ploidy level does not substantially alter the probability curves. However, this is not true for triple- and quadruple-copy alleles for which misclassifying the ploidy level can greatly increase the chances of the test yielding an inconclusive result.

Discussion

In this paper, I reveal the true behaviour of the estimation procedure commonly used to infer a dominant allele's copy number in polyploids. By deriving the theoretical distribution of the segregation ratio for a dominant allele, I was able to quantify the performance of the estimation procedure when the assumptions of known ploidy level and no preferential pairing are relaxed. I found that:

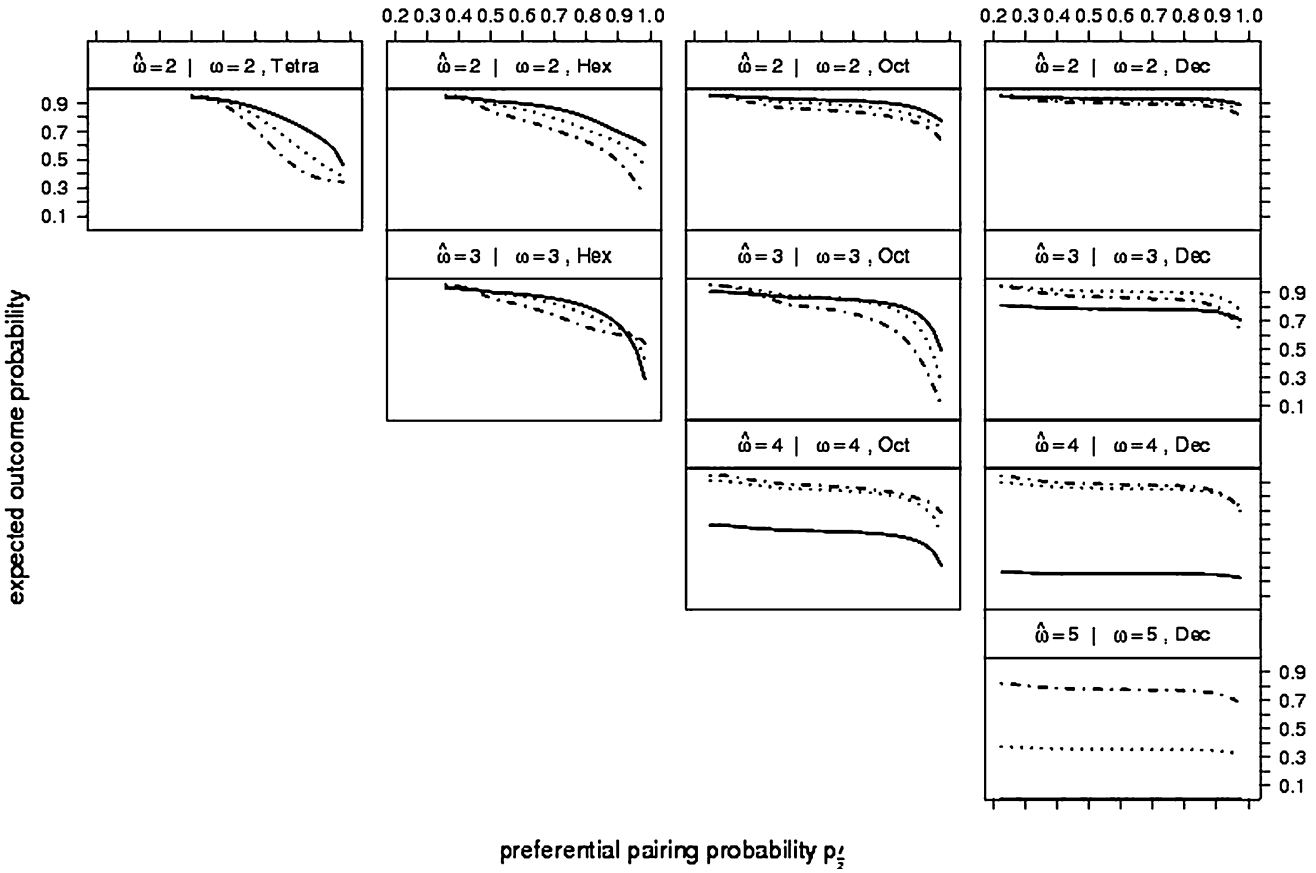

Fig. 1 Expected outcome probability curve of correctly estimating the copy number of a multiple copy allele conditioned on number of progeny and preferential pairing probability $p_{\ell/2}$. I have assumed that the ploidy level is known, but the amount of preferential pairing is unknown (the bad). The *solid*, *dotted*, and *dot-dashed lines* represent the expected probability curve for a family with 100, 250, and 500 progeny, respectively. In the *bottom right plot*, the *solid curve* is on top of the horizontal axis

1. Single-copy alleles are accurately estimated even if the ploidy level and amount of preferential pairing is uncertain.
2. Increased amounts of preferential pairing affect copy number estimation in tetraploids and hexaploids more than in polyploids with higher ploidy levels such as octoploids and decaploids.
3. Family size has greater impact on the performance of the estimation procedure for the higher ploidy levels and dominant alleles with higher copy numbers.
4. Increasing the number of progeny with marker information can lead to an increase in the probability of the estimation procedure failing to unambiguously determine the copy number.
5. When the assumptions of known ploidy level and no preferential pairing are violated, the test is much more likely to fail to unambiguously determine the copy number.

This work highlights the need for a more effective strategy for estimating copy number in polyploids. Ripol et al. 1999 derives a binomial test for copy number estimation, when the ploidy level is known and chromosomes randomly pair. For small sample sizes ($n < 30$), this test is slightly more accurate than using Pearson's chi-squared test. However, plant studies typically involve hundreds of plants. Consequently, the acceptance regions used in these studies associated with the chi-squared and binomial tests are almost identical. Currently, I am developing a Bayesian approach for copy number estimation based on the probability distributions derived in this paper. By using numerical integration techniques, I compute the posterior probability of ω given $\mathbf{Y}$ where $\mathbf{p}$ has been marginalised out of the joint posterior distribution. This marginalised posterior probability is an intuitive measure of how much evidence I have for a particular allele's copy number. Preliminary results suggest that although estimating the true amount of preferential pairing is difficult, including preferential pairing probabilities in the probability model gives results superior to the chi-squared test.

Overall, the strategy of forming a series of hypotheses and testing their significance using Pearson's chi-squared test works surprisingly well as an estimation procedure.

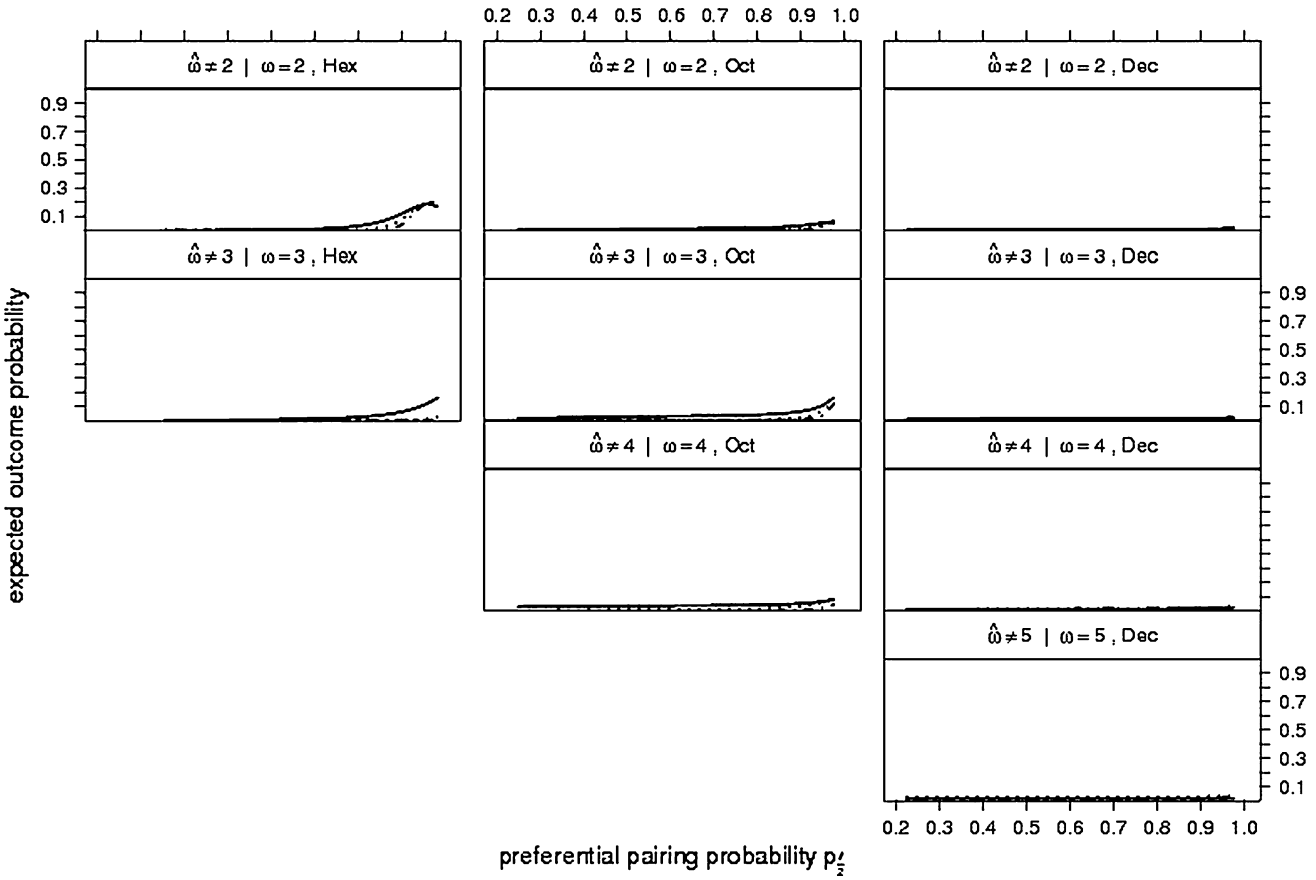

Fig. 2 Expected outcome probability curve of misclassifying the copy number of a multiple copy number allele conditioned on number of progeny and preferential pairing probability $p_{\ell/2}$. I have assumed that the ploidy level is known, but the amount of preferential pairing is unknown (the bad). The *solid*, *dotted*, and *dot-dashed lines* represent the expected probability curve for a family with 100, 250, and 500 progeny, respectively

The strengths of the approach are its simplicity and ability to accurately identify the highly sought after single-copy alleles. It is also reassuring that the probability of misclassifying an allele's copy number is quite low if the amount of preferential pairing is not high. However, the estimation procedure is much more likely to yield an inconclusive result for high copy alleles, especially if the sample size is small. The challenge in using the chi-square test for copy number estimation is not in accurately estimating an allele's copy number but in obtaining a significant result.

Appendix 1: Distribution of the segregation ratio for a dominant marker

Here, I derive the probability distribution for the segregation ratio (proportion of progeny carrying a dominant allele) in polyploids. I begin by constructing the joint probability of the observed marker data conditional on the amount of preferential pairing. I then simplify this joint distribution and use it to derive the probability distribution of the segregation ratio.

Suppose data on a dominant marker are collected from a full-sib family with n progeny. The marker phenotype is a dichotomous trait for which only the presence or absence of a dominant allele is observed. I denote a family's marker data by $\mathbf{Y} = (y_m, y_p, y_1, y_2, \ldots, y_n)$ where y_m and y_p are the marker phenotypes for the maternal and paternal parent, respectively, y_j is the marker phenotype for the jth progeny, and the marker phenotype is either 1 (presence) or 0 (absence). Dominant alleles are assumed to segregate from only one parent and I have, without loss of generality, chosen the paternal parent. The paternal parent carries ω copies of a dominant allele.

I begin my derivation of the probability distribution of the segregation ratio s by deriving the joint probability distribution of the observed data $\mathbf{Y}$. I can formulate the joint probability of $\mathbf{Y}$ in terms of the genetic information (i.e. the unobserved marker genotypes) that is passed from parent to progeny. This is due to data observed on a family being determined by a family's latent marker genotypes.

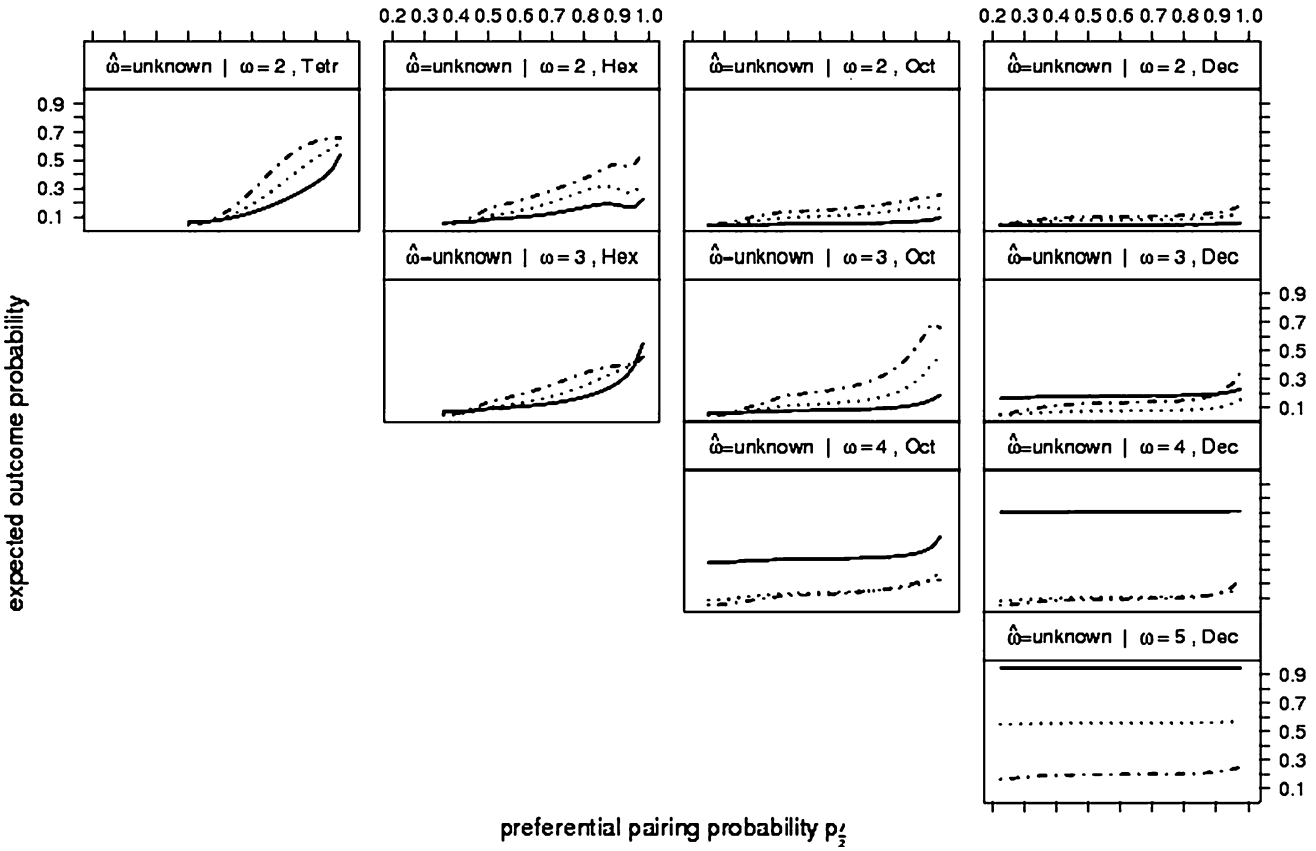

Fig. 3 Expected outcome probability curve of obtaining an inconclusive result given the number of progeny and preferential pairing probability $p_{\ell/2}$. I have assumed that the ploidy level is known, but the amount of preferential pairing is unknown (the bad). The *solid*, *dotted*, and *dot-dashed lines* represent the expected probability curve for a family with 100, 250, and 500 progeny, respectively. Note that the probability curve's kinked behaviour for high $p_{\ell/2}$ in some of the plots is due to the sharp increase in probability of misclassifying an allele's copy number

Note that an individual's latent marker genotype is also equivalent to the founder genes or identical by descent (ibd) genes that an individual inherits from its parents (George and Thompson 2003).

Suppose $\mathbf{G}$ is a vector of possible genotypes for a marker locus. Each element in $\mathbf{G}$ corresponds to an individual's marker genotype. There will be many possible $\mathbf{G}$ that will be consistent with the observed marker data. Expressing the joint probability of the observed data in terms of the possible latent marker genotypes, we have $\Pr_{\ell,\omega}(\mathbf{Y}|\mathbf{p}) = \sum_{\mathbf{G}} \Pr_{\ell,\omega}(\mathbf{Y},\mathbf{G}|\mathbf{p})$ where $\Pr_{\ell,\omega}(\mathbf{Y}|\mathbf{p})$ is the joint probability distribution of the observed data given preferential pairing probabilities $\mathbf{p}$ and assuming ploidy level l and copy number ω, and $\sum_{\mathbf{G}}$ is the sum over all possible genotypic configurations. For notational convenience, I will no longer subscript a probability distribution by its ploidy level l. All subsequent probabilities are constructed under an assumed ploidy level. Using a property of conditional probabilities, we can write

$$\Pr_{\omega}(\mathbf{Y}|\mathbf{p}) = \sum_{\mathbf{G}} \Pr(\mathbf{Y}|\mathbf{G}) \Pr_{\omega}(\mathbf{G}|\mathbf{p}) \tag{3}$$

where $\Pr(\mathbf{Y}|\mathbf{G})$ is the joint probability of the observed marker data conditional on the unobserved marker genotypes, and $\Pr_{\omega}(\mathbf{G}|\mathbf{p})$ is the joint probability of the latent marker genotypes conditional on the preferential pairing probabilities $\mathbf{p}$ and copy number ω.

Equation 3 is still not in a manageable form, since the joint probabilities on the right hand side are difficult to compute. In preparation for constructing a more tractable form of the joint probability of $\mathbf{Y}$, I note the following. First, as discussed previously, only those marker loci whose dominant alleles originate from a single parent (and I have arbitrarily chosen the paternal parent) need be considered for analysis. Consequently, the distribution of $\mathbf{Y}$ is independent of maternally inherited genetic information.

Second, a family member's marker phenotype is independent of the other phenotypic data given the family member's marker genotype. The joint conditional probability of $\mathbf{Y}$ given $\mathbf{G}$ can therefore be factored into the product of $n + 1$ marginal distributions

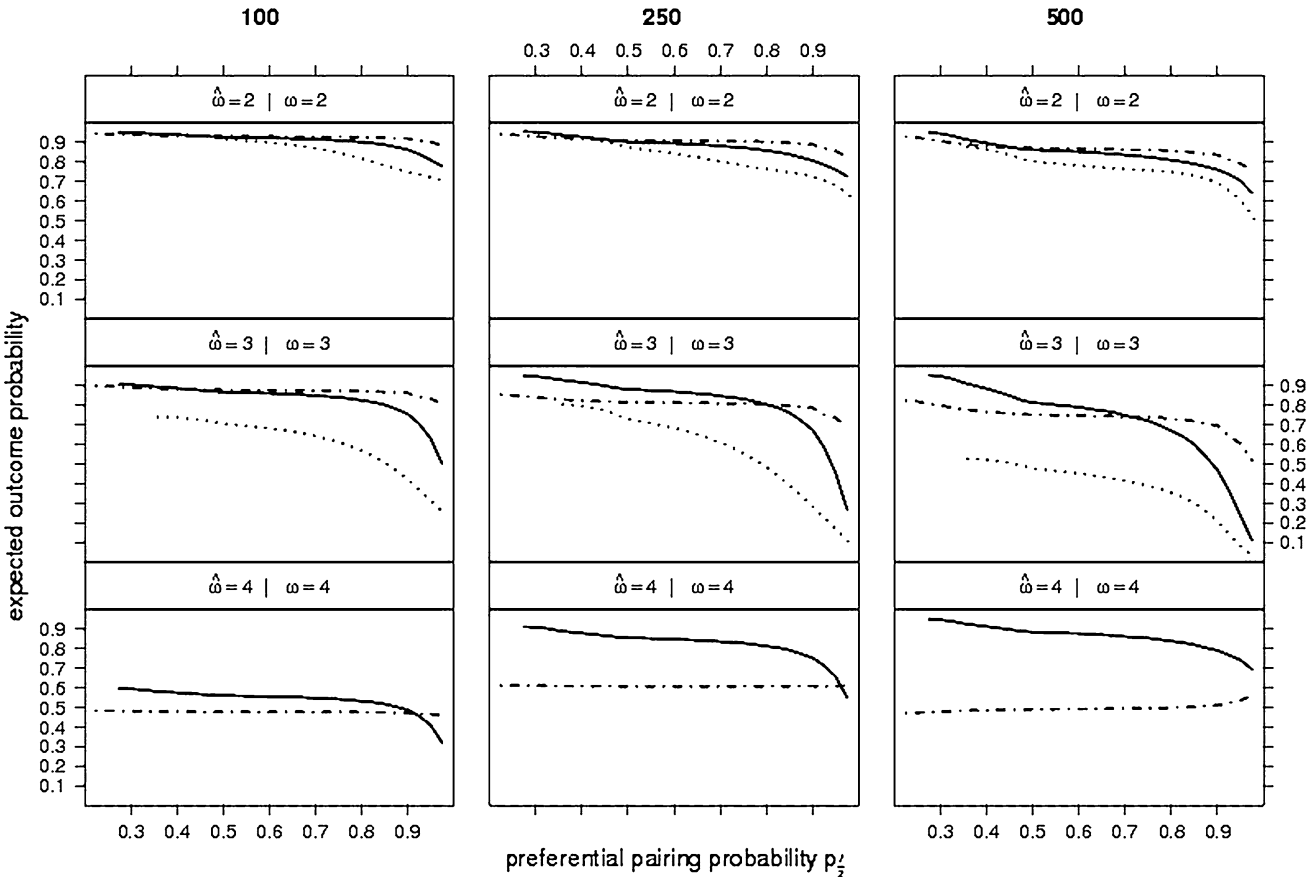

Fig. 4 Expected outcome probability curve of correctly estimating a dominant allele's copy number for a family with 100, 250, and 500 progeny. I have assumed an unknown ploidy level and an unknown amount of preferential pairing (the ugly). The *solid line* is the probability curve associated with correctly assuming an octoploid. The *semi-dashed line* is the probability curve associated with incorrectly assuming a decaploid when the polyploid is an octoploid. The *dashed line* is the probability curve associated with incorrectly assuming a hexaploid when the polyploid is an octoploid

$$\Pr(\mathbf{Y}|\mathbf{G}) = \Pr(y_p = 1|G_p)\prod_{j=1}^{n}\Pr(y_j|G_j) \tag{4}$$

where G_p is the paternal parent's latent marker genotype, and G_j is the paternally inherited marker genotype for the jth progeny.

Third, I assume that the marker data does not contain genotyping errors. Hence, the marginal probabilities in Eq. 4 are either zero or one such that:

$$\Pr(\mathbf{Y}|\mathbf{G}) = \begin{cases} 1 & \text{if } y = 1 \text{ and } G \in \{G^d\} \\ 1 & \text{if } y = 0 \text{ and } G \notin \{G^d\} \\ 0 & \text{otherwise} \end{cases}$$

where, for notational convenience, I have ignored the subscripts on $\mathbf{Y}$ and $\mathbf{G}$, and $\{G^d\}$ is the set of possible genotypes of a progeny that contains at least one dominant allele.

Fourth, a progeny's genotype is independent of its siblings' genotypes if the parental genotypes are known. The joint probability distribution of $\mathbf{G}$ in Eq. 3 can therefore be factorised as

$$\Pr_{\omega}(\mathbf{G}|\mathbf{p}) = \Pr_{\omega}(G_p|\mathbf{p})\prod_{j=1}^{n}\Pr_{\omega}(G_j|G_p,\mathbf{p}) \tag{5}$$

By substituting the factorised forms of Eqs. 4 and 5 into Eq. 3 and remembering that the data are observed without error, we can write

$$\Pr_{\omega}(\mathbf{Y}|\mathbf{p}) = \sum_{G_p^d}\Pr_{\omega}(G_p^d|\mathbf{p})\Big[\Pr_{\omega}(G_p = G_p^d|\mathbf{p})\Big]^{x} \times \Big[\Pr_{\omega}(G_p \neq G_p^d|\mathbf{p})\Big]^{(n-x)} \tag{6}$$

where $\sum_{G_p^d}$ is the summation over the set of paternal marker genotypes that carry at least one copy of a dominant allele, and $x = \sum_{j=1}^{n} y_j$ is the number of progeny that exhibits a dominant allele. The three terms on the right hand side of Eq. 6 are univariate probabilities and easy to calculate. The first term is the probability of the paternal parent having marker genotype G_p^d given copy number ω and preferential pairing probability $\mathbf{p}$. The second term is the probability that a progeny carries at least one copy of a dominant allele

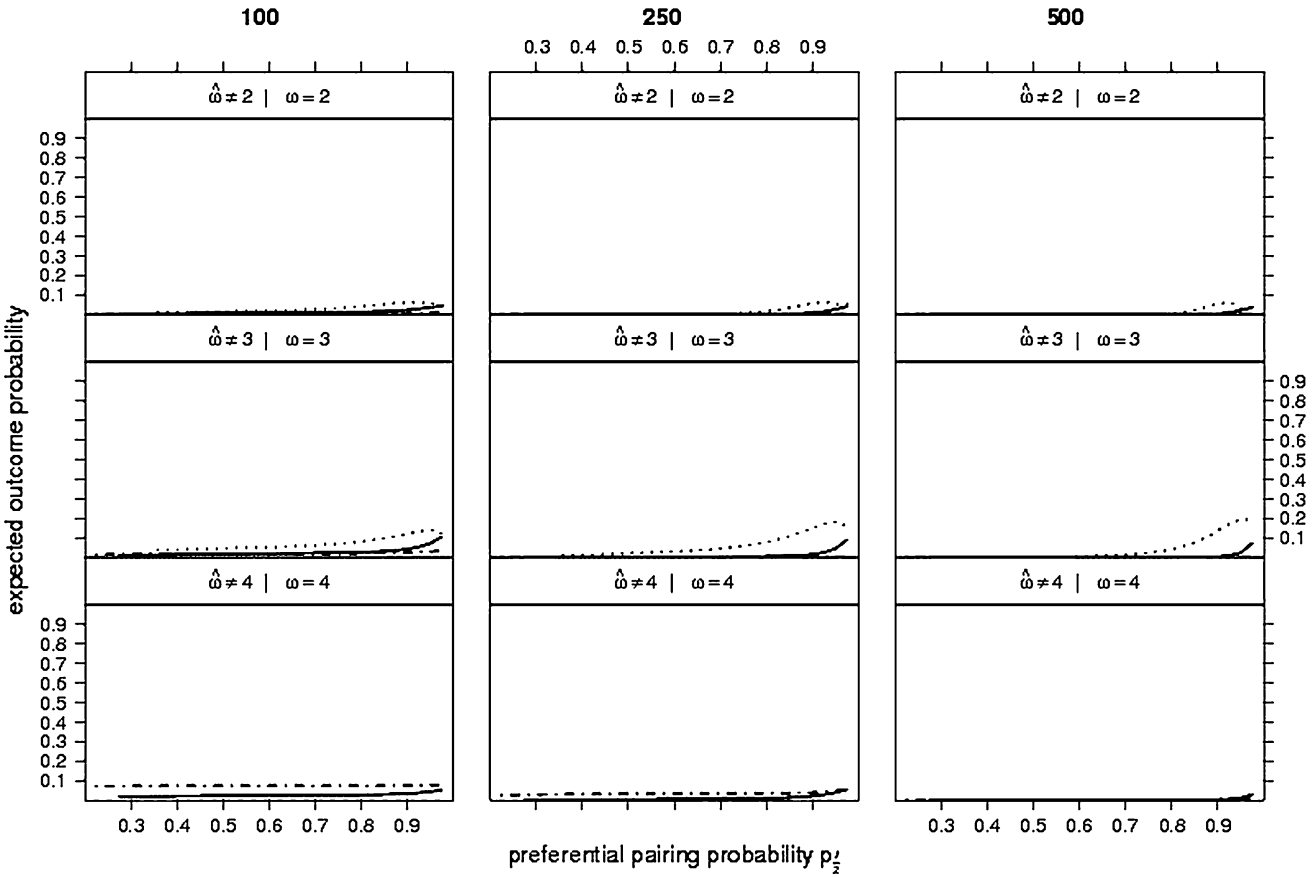

Fig. 5 Expected outcome probability curve of misclassifying a dominant allele's copy number for a family with 100, 250, and 500 progeny. I have assumed an unknown ploidy level and an unknown amount of preferential pairing (the ugly). The *solid line* is the probability curve associated with correctly assuming an octoploid. The *semi-dashed line* is the probability curve associated with incorrectly assuming a decaploid when the polyploid is an octoploid. The *dashed line* is the probability curve associated with incorrectly assuming a hexaploid when the polyploid is an octoploid

conditional on the paternal parent having at least one dominant allele, the copy number, and the preferential pairing probabilities. The third term is the probability that a progeny does not inherit any dominant alleles given that the paternal parent carries at least one copy of a dominant allele, the copy number, and the preferential pairing probabilities.

Before constructing the probability distribution of a segregation ratio, I note that segregation ratios are summary measures. For a marker locus with observed data $\mathbf{Y}$, the associated segregation ratio is x/n and there are many different data sets $\mathbf{Y}$ that can result in the same segregation ratio. In fact, for a family with n progeny, there are $\binom{n}{ns}$ combinations of ns progeny exhibiting a dominant allele.

The probability distribution of the segregation ratio for dominant marker in polyploids is then

$$\Pr_{\omega}(s|\mathbf{p}) = \binom{n}{ns}\Pr_{\omega}(\mathbf{Y}|\mathbf{p}) = \sum_{j=1}^{m} w_j \mathrm{Bin}(ns; n, \pi_j) \qquad (7)$$

where m is the number of mixture components, j is the index over the mixture components, $w_j \propto \Pr_{\omega}(G_p^d|\mathbf{p})$ is the jth mixture weight, $\mathrm{Bin}(\cdot)$ is the binomial probability distribution of observing ns progeny exhibiting a dominant allele in a family with n progeny, and $\pi_j = \Pr(G_p = G^d|G_p^d, \omega, \mathbf{p})$ is the jth probability of a progeny exhibiting a dominant allele given that the paternal parent carries a dominant allele, the copy number, and the preferential pairing probabilities. For an example of how to form Eq. 7 for a double-copy allele in hexaploids, see "Appendix 2".

Appendix 2: An example of constructing the distribution of the segregation ratio

In this example, the construction of the probability distribution for the segregation ratio of a double-copy allele in hexaploids is presented. I begin by assigning preferential pairing probabilities to the set of unique chromosome pairing configurations. Based on these preferential pairing probabilities, a probability is calculated for each possible parental gamete. I then construct the distribution of those

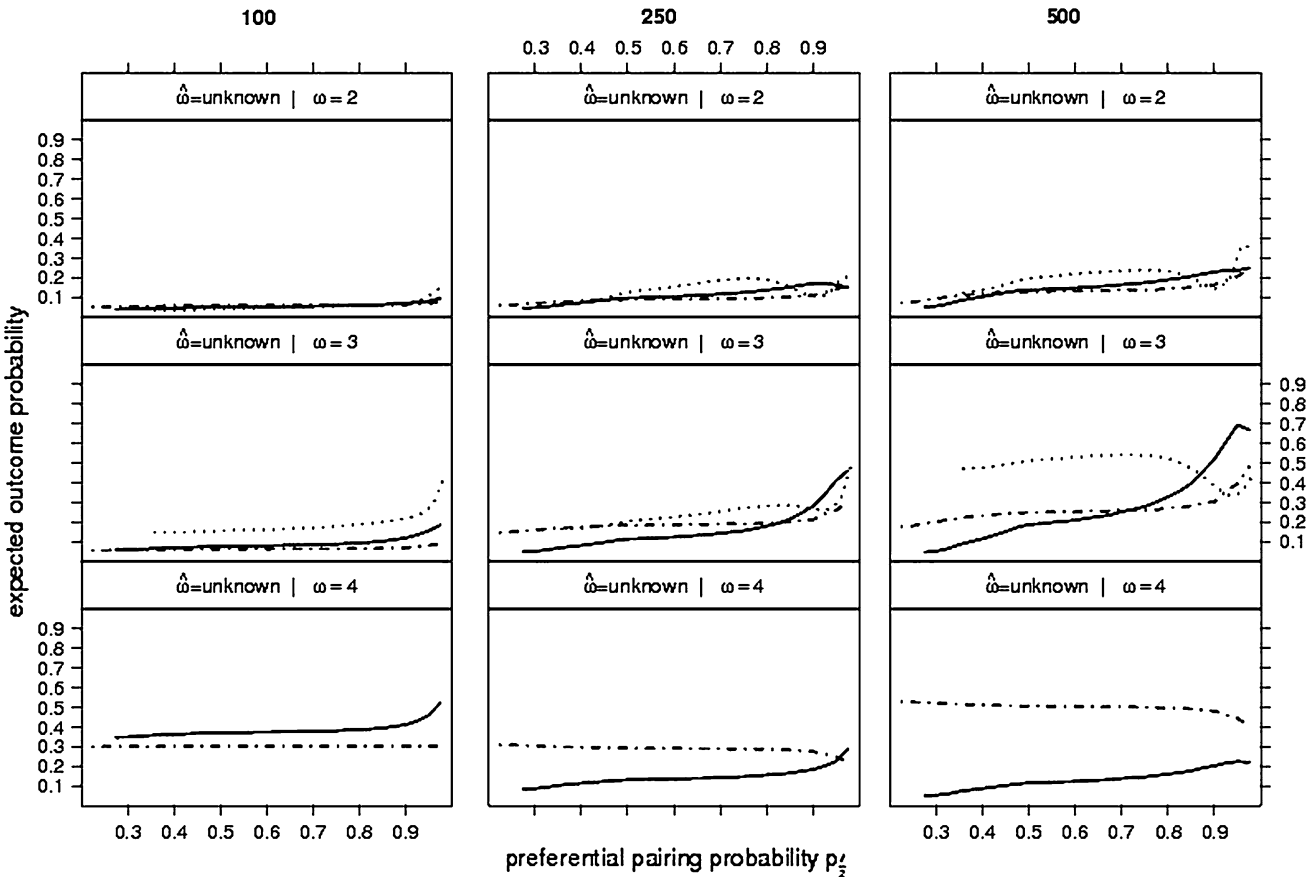

Fig. 6 Expected outcome probability curve of obtaining an inconclusive result for a family with 100, 250, and 500 progeny. I have assumed an unknown ploidy level and an unknown amount of preferential pairing (the ugly). The *solid line* is the probability curve associated with correctly assuming an octoploid. The *semi-dashed line* is the probability curve associated with incorrectly assuming a decaploid when the polyploid is an octoploid. The *dashed line* is the probability curve associated with incorrectly assuming a hexaploid when the polyploid is an octoploid. The probability curve's strange shape for high $p_{\ell/2}$ in some of the plots is due to an increase in the probability of misclassifying a dominant allele's copy number

gametes that result in a progeny exhibiting a dominant allele given the copy number. Finally, this distribution is used to calculate the mixture distribution parameters w_j and π_j. I assume that each chromosome has a single homologous partner and are labelled 1_A, 2_A, 3_B, 4_B, 5_C, and 6_C where chromosomes with the same subscript are homologous. I also assume that chromosomes pair during meiosis as bivalents and, without loss of generality, that dominant alleles originate from the paternal parent.

To begin, I construct the distribution of the unique chromosome pairing configurations (Table 3). There are 15 unique pairings and each configuration is assigned a probability that is based upon the number of associated homologous bivalents. I then construct a probability distribution for the possible gametes that may be inherited from a hexaploid father (Table 4). Each gamete may have originated from one of several pairing configurations. To assign a probability to a gamete, I take the sum of probabilities of the configurations from which the gamete may have originated. For example, from Table 3, the paternal gamete 1_A, 2_A, 3_B may have originated from pairing configurations 8, 9, 11, 12, 14, and 15. Therefore, the gametic probability is the sum of the associated configuration probabilities such that the probability of 1_A, 2_A, 3_B is $p_0 + 2p_1$.

There are several ways in which I could present the calculation of the mixture parameters w and π that appear in Eq. 2. I found a tabular form to be the easiest to demonstrate their calculation. In Table 4, each row corresponds to a different paternal gamete and each column corresponds to a pair of chromosomes carrying a dominant allele. A cell in the table contains a one (zero) if, given that the paternal chromosomes carrying the dominant allele, a progeny would (not) exhibit a dominant allele if they inherited the gamete. The binomial probabilities π_j are then easily calculated from the ratio of the sum of gamete probabilities with non-zero cell entries to the normalizing constant $12a + 8b$. The weights w_j are

Table 3 Unique chromosome pairing configurations associated with a hexaploid with chromosomes labelled 1_A, 2_A, 3_B, 4_B, 5_C, and 6_C where chromosomes with the same subscripts are homologous

Index	chromosome pairing configurations	Number of homologous bivalents	Preferential pairing probabilities	Index	chromosome pairing configurations	Number of homologous bivalents	Preferential pairing probabilities
1	12 34 56	3	p_3	9	14 26 35	0	p_0
2	12 35 46	1	p_1	10	15 23 46	0	p_0
3	12 36 45	1	p_1	11	15 24 36	0	p_0
4	13 24 56	1	p_1	12	15 26 34	1	p_1
5	13 25 46	0	p_0	13	16 23 45	0	p_0
6	13 26 45	0	p_0	14	16 24 35	0	p_0
7	14 23 56	1	p_1	15	16 25 34	1	p_1
8	14 25 36	0	p_0				

For notational convenience, the genome subscripts A, B, and C have been dropped. Homologous bivalents are underlined. All probabilities are normalized (i.e. $8p_0 + 6p_1 + p_3 = 1$)

Table 4 Tabular form for the calculation of the mixing weights w_j and binomial probabilities π_j in Eq. 2

Pat	Gam	Paternal chromosomes carrying dominant alleles														
Gam	Prob	12	13	14	15	16	23	24	25	26	34	35	36	45	46	56
123	a	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
124	a	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0
125	a	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1
126	a	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1
134	a	1	1	1	1	1	1	1	0	0	1	1	1	1	1	0
135	b	1	1	1	1	1	1	0	1	0	1	1	1	1	0	1
136	b	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1
145	b	1	1	1	1	1	0	1	1	0	1	1	0	1	1	1
146	b	1	1	1	1	1	0	1	0	1	1	0	1	1	1	1
156	a	1	1	1	1	1	0	0	1	1	0	1	1	1	1	1
234	a	1	1	1	0	0	1	1	1	1	1	1	1	1	1	0
235	b	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1
236	b	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1
245	b	1	0	1	1	0	1	1	1	1	1	1	0	1	1	1
246	b	1	0	1	0	1	1	1	1	1	1	0	1	1	1	1
256	a	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1
345	a	0	1	1	1	0	1	1	1	0	1	1	1	1	1	1
346	a	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1
356	a	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1
456	a	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1
π_j		π_1	π_2	π_2	π_2	π_2	π_2	π_2	π_2	π_2	π_1	π_2	π_2	π_2	π_2	π_1

Gamete probabilities (Gam Prob) are calculated from the sum of configuration probabilities associated with the pairing configurations the gamete may have originated from. Here, $a = 4p_0 + 2p_1$, $b = 2p_0 + 3p_1 + p_3$, $\pi_1 = \frac{8a+8b}{12a+8b}$, and $\pi_2 = \frac{10a+6b}{12a+8b}$ where p_0, p_1, and p_3 are preferential pairing probabilities

obtained from the proportion of columns that result in the same π_j. For example, there are three out of 15 columns in Table 3 that result in π_1 so $w_1 = {}^3/_{15}$ and there are 12 out of 15 columns that result in π_2 so $w_2 = {}^{12}/_{15}$. Hence, the probability distribution of s for a double copy-number allele in hexaploids is

$$\Pr_{\omega=2}(s|\mathbf{p}) = \frac{3}{15}\mathrm{Bin}(ns; n, \pi_1) + \frac{12}{15}\mathrm{Bin}(ns; n, \pi_2)$$

where $\pi_1 = \frac{8a+8b}{12a+12b}$, $\pi_2 = \frac{10a+6b}{12a+6b}$, and a and b are defined in Table 4.

References

Aitken K, Jackson P, McIntyre L (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

Dasilva J, Honeycutt RJ, Burnquist W, Aljanabi SM, Sorrells ME, Tanksley SD, Sobral BWS (1995) Saccharum-Spontaneum L SES-208 genetic-linkage map combining RFLP-based and PCR-based markers. Mol Breeding 1:165–179

Doyle GG (1963) Preferential pairing in structural heterozygotes of Zea mays. Genetics 48:1011–1027

George AW, Thompson EA (2003) Discovering disease genes: Multipoint linkage analysis via a new Markov chain Monte Carlo approach. Stat Sci 18:515–531

Grivet L, DHont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (Saccharum spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142:987–1000

Leitch IJ, Bennett MD (1997) Polyploidy in angiosperms. Trends Plant Sci 2:470–476

Missaoui AM, Paterson AH, Bouton JH (2005) Investigation of genomic organization in switchgrass (Panicum virgatum L.) using DNA markers. Theor Appl Genet 110:1372–1383

Pfosser M, Amon A, Lelley T, Heberlebors E (1995) Evaluation of sensitivity of flow-cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat-rye addition lines. Cytometry 21:387–393

Rhoades MM (1952) Preferential segregation in maize. In: Cowen JW (ed) Heterosis. Iowa State College Press, Ames, pp 66–80

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

Sybenga J (1992) Cytogenetics in plant breeding. Springer, Berlin

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

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol 42:225–249

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

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