Heredity **87** (2001) 537–543 *Received 10 May 2000, accepted 25 May 2001*

Analysis of segregation data from selfed progeny of allopolyploids

M. S. RIDOUT*, J. A. BELL & D. W. SIMPSON
HRI-East Malling, West Malling, Kent ME19 6BJ, U.K.

This paper discusses the inference of parental genotype based on segregation data from selfed progeny of allopolyploids when there is incomplete information about genotypes and when alleles are codominant or null. The distinct alleles that are present in a genotype are assumed to be known, but not the frequency with which they occur. These assumptions may be appropriate when genotypes are deduced from DNA or protein banding patterns on electrophoretic gels. A computer program, SELF, is described that can generate all possible parental genotypes and rank them on the basis of their agreement with the progeny data. The program caters for tetraploids, hexaploids and octoploids. The methods are illustrated using data from a study of the inheritance of isoenzymes in selfed progeny of octoploid strawberry cultivars.

Keywords: banding pattern, *Fragaria* × *ananassa*, isoenzyme, polyploid, selfing.

Introduction

This paper discusses the inference of parental genotype based on segregation data from selfed progeny of allopolyploids when there is incomplete information about genotypes and when alleles are codominant or null. It is assumed that the distinct alleles that are present in a genotype can be identified, but that the frequency with which each allele occurs is not known. This situation arises in the interpretation of electrophoretic gels, when the presence of a DNA or protein band on the gel indicates the presence of the corresponding codominant allele in the genotype. Although the relative staining intensities of different bands may provide some information about the frequencies with which the different alleles are present, this information is often very difficult to quantify reliably. It is important therefore to consider how data can be analysed when the allele frequencies are unknown. Specifically, what can be inferred about the parental genotype from knowledge of the banding patterns that occur in selfed progeny? Knowledge of parental gen otypes would be important for linkage studies, for example.

The method of analysis described in this paper was developed as part of a study of the inheritance of several isoenzymes in selfed progenies of various strawberry cultivars. The cultivated strawberry, *Fragaria* × *ananassa* Duch., is an octoploid ($2n = 56$). There is good cytological and genetic evidence, reviewed by Galletta & Maas (1990) and Bringhurst (1990), that *F.* × *ananassa* behaves as an allopolyploid, showing disomic inheritance. Thus, the eight sets of chromosomes comprise four pairs and inheritance proceeds by ordinary diploid segregation within each pair. Based on this assumption, we use a computer program to generate all possible parental genotypes and to rank them on the basis of their compatibility with the progeny data.

The paper is organized as follows. Firstly, statistical methods are described. Secondly, these methods are illustrate using data on segregation of the enzyme system phosphoglucomutase (*PGM*) in selfed progeny of two strawberry cultivars, 'Cambridge Favourite' and 'Elvira'. For 'Cambridge Favourite', only a few parental genotypes are possible, and these are easily identified by inspection. For 'Elvira', there are many more possibilities and the use of a computer program is essential for efficiency and accuracy. Finally, we discuss extensions of the methodology to allow for null alleles and for the possibility that the locus of interest may not be present on all four of the chromosome pairs.

A Fortran77 computer program, SELF, was developed for the calculations and caters for tetraploids, hexaploids and octoploids. An executable version of the program, which runs under MS-DOS on an IBM-compatible personal computer, is available by E-mail from the first author.

*Correspondence and present address. Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent, CT2 7NF, UK. E-mail: m.s.ridout@ukc.ac.uk

Statistical methods

We write the octoploid genotype of an individual as a list of four pairs of alleles. The general form is thus

$$(ab, cd, ef, gh),$$

although, in most instances, not all of the eight alleles will be distinct. All alleles are assumed to be codominant.

When an allopolyploid is selfed, the four pairs of alleles segregate independently of one another, in normal Mendelian fashion, giving rise to a maximum of $2^4 = 16$ distinct gametes. As each pair of alleles can give rise to three distinct offspring genotypes, there are, in total, $3^4 = 81$ possible offspring genotypes, although this number is reduced by a factor of three whenever the two alleles of a pair are identical. Moreover, because the four pairs are assumed to segregate independently of one another, the probabilities of the different genotypes are obtained by multiplying together the probabilities for each pair. For example, the probability of obtaining the offspring genotype (*ab*, *cc*, *ef*, *hh*) is

$$\frac{1}{2} \times \frac{1}{4} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{64}.$$

For this individual, the banding pattern *abcefh* would be observed and, as this pattern can arise only from this offspring genotype, the probability of observing this banding pattern is also 1/64.

However, when the alleles are not all distinct, a particular banding pattern can often arise from several different offspring genotypes and then the overall probability of obtaining the banding pattern is the sum of the individual genotype probabilities. For example, when the parent (*ab*, *cd*, *ef*, *af*) is selfed, the banding pattern *acef* arises from any of the following offspring genotypes:

Offspring genotype	Probability
(aa, cc, ee, af)	1/128
(aa, cc, ee, ff)	1/256
(aa, cc, ef, aa)	1/128
(aa, cc, ef, af)	1/64
(aa, cc, ef, ff)	1/128

Summing these probabilities gives the overall probability of obtaining the banding pattern *acef* as 11/256. Thus, for a given parental genotype, it is straightforward to identify the distinct offspring genotypes and to calculate the probabilities with which they occur.

In practice, however, the parental genotype is unknown. We assume that the distinct parental alleles are known, but not the frequency with which they occur or the way in which they are paired. The objective is to infer this information from the selfed progeny. To this end, we consider all possible parental genotypes and, for each of these, evaluate the different banding patterns that can occur in the progeny, and their expected frequencies. These expected frequencies may then be compared with the observed frequencies. When the agreement is poor, it is unlikely that the postulated parental genotype is correct. Different parental genotypes can therefore be ranked on the basis of the degree of agreement between observed and expected frequencies.

Generation of possible parental genotypes

Suppose that there are m distinct parental alleles ($1 \leq m \leq 8$) and that the ith allele occurs r_i times, where

$$r_1 + r_2 + \ldots + r_m = 8.$$

In mathematical terminology (r_1, r_2,..., r_m) is a *partition* of eight. Table 1 gives the possible partitions for different values of m. The SELF program considers each partition in turn.

Suppose now that the partition (r_1, r_2,..., r_m) contains p distinct values, say (s_1, s_2,..., s_p) and that the number of alleles that occur s_i times in the genotype is u_i. This implies that

$$s_1u_1 + s_2u_2 + \ldots + s_pu_p = 8.$$

For example, if $m = 5$, with alleles *abcde*, and the partition is (1,1,2,2,2), then $P = 2$ and we have $s_1 = 1$, $s_2 = 2$, $u_1 = 2$, $u_2 = 3$. There are 10 different combinations of the alleles *abcde* in which two of the alleles occur once and the other three occur twice, namely

abccddee abbcddee abbccdee abbccdde aabcddee
aabccdee aabccdde aabbcdee aabbcdde aabbccde.

The number of combinations, C, of m alleles occurring with the specified frequencies (r_1, r_2,..., r_m) is given by the multinomial coefficient

$$C = \frac{m!}{u_1!u_2! \ldots u_p!},$$

which is the number of different ways of choosing the u_1 alleles that occur s_1 times, the u_2 alleles that occur s_2 times, etc. Each such combination gives rise to several possible genotypes, V say, depending on the way in

Table 1 The partitions of 8 that contain exactly m elements, where m is the number of distinct parental alleles. For each partition, the table shows the number of distinct values in the partition (p), the number of possible combinations of alleles that can arise (C), the number of ways in which each of these can be arranged in four pairs (V) and hence the number of possible genotypes ($G = C \times V$). See text for more details

m	Partition	p	C	V	G
1	(8)	1	1	1	1
2	(7,1)	2	2	1	2
	(6,2)	2	2	2	4
	(5,3)	2	2	2	4
	(4,4)	1	1	3	3
3	(6,1,1)	2	3	2	6
	(5,2,1)	3	6	3	18
	(4,3,1)	3	6	4	24
	(4,2,2)	2	3	6	18
	(3,3,2)	2	3	6	18
4	(5,1,1,1)	2	4	4	16
	(4,2,1,1)	3	12	7	84
	(3,3,1,1)	2	6	8	48
	(3,2,2,1)	3	12	11	132
	(2,2,2,2)	1	1	17	17
5	(4,1,1,1,1)	2	5	10	50
	(3,2,1,1,1)	3	20	16	320
	(2,2,2,1,1)	2	10	23	230
6	(3,1,1,1,1,1)	2	6	25	150
	(2,2,1,1,1,1)	2	15	36	540
7	(2,1,1,1,1,1,1)	2	7	60	420
8	(1,1,1,1,1,1,1,1)	1	1	105	105

which the alleles are paired. When all eight alleles are distinct, $V = 7 \times 5 \times 3 = 105$. We are not aware of any simple general formula for V, but numerical values are shown in Table 1. The number of parental genotypes that can arise from a given partition is then $G = C \times V$, and the total number of parental genotypes that can arise from a given set of m alleles is obtained by summing the G-values for all partitions that contain m elements. The results are shown in the second column of Table 2. The number of possible genotypes increases as m increases, to a maximum of 690 when $m = 6$, and decreases thereafter.

Ranking parental genotypes

For each parental genotype that is generated, the SELF program enumerates the possible offspring banding patterns, and the probabilities with which they occur. In most instances, many of the parental genotypes may be ruled out immediately, because they are unable to generate all of the banding patterns that occur in the data. The remaining genotypes are ranked on the basis

Table 2 The number of possible octoploid parental genotypes, given the number of distinct parental alleles, m, depending on whether or not null alleles are allowed

	Number of possible parental genotypes	
m	Without null alleles	With null alleles
1	1	14
2	13	97
3	84	381
4	297	897
5	600	1290
6	690	1110
7	420	525
8	105	105

of their agreement with the observed data. The measure of agreement used is the standard chi-squared statistic, χ^2, calculated from the observed and expected frequencies (n_i and e_i) as

$$\chi^2 = \sum \frac{(n_i \quad e_i)^2}{e_i},$$

where the summation is over all banding patterns that can occur, given the parental genotype. This may include banding patterns that do not occur in the data ($n_i = 0, \ e_i \neq 0$). The degrees of freedom for the chi-squared statistic are one less than the number of possible banding patterns.

When parental genotypes give nonzero probabilities for banding patterns that do not occur in the data, the degrees of freedom for χ^2 are increased, often without greatly increasing χ^2 itself. Thus, based on the chi-squared test, these genotypes may compare favourably with parental genotypes that do not give rise to unobserved banding patterns. However, χ^2 is only one possible measure of agreement between observed and expected banding patterns. We can also calculate the probability that none of the unobserved banding patterns occurs as

$$(1 \quad p_0)^N,$$

where p_0 is the sum of the probabilities of the banding patterns that do not occur in the data, and N is the progeny size. If this probability is small, the proposed parental genotype is unlikely to be correct.

Finally, we calculate one further measure of agreement, the posterior probability that a given parental genotype is the true parental genotype, assuming that all possible genotypes are equally likely *a priori*. This type of calculation, which uses the ideas of Bayesian statistics, has been advocated recently by Olson (1997),

in a similar context. The posterior probability for the j^{th} possible parental genotype is calculated as

$$\frac{P_j}{\sum P_j},$$

where P_j denotes the multinomial probability of the observed offspring segregation given the jth parental genotype. Two specific genotypes, i and j, may be compared by means of the Bayes factor, which, because all genotypes have the same prior probability, is simply $B_{ij} = P_i/P_j$. Thus, the Bayes factor is the ratio of the likelihoods for the two genotypes. Based on the suggestions of Kass & Raftery (1995), in an extensive review of Bayes factors, a Bayes factor between 3 and 20 provides 'positive' evidence in favour of the genotype with higher posterior probability, whilst a value in excess of 20 provides 'strong' evidence.

Data

We illustrate the approach using data on *PGM* segregation in selfed progeny of two strawberry cultivars, 'Cambridge Favourite', which has alleles *a*, *c* and *h*, and 'Elvira', which has alleles *a*, *c*, *e*, *f* and *h*. The possibility that null alleles might be present is also considered. The experiment methods are described by Bell & Simpson (1994).

Results

PGM segregation in selfed progeny of 'Cambridge Favourite'

Four different banding patterns were observed in 57 offspring (Table 3). We note first that there is a single individual in which only the allele *c* was detected; this implies that each pair of alleles must include at least one *c* allele, and the genotype is therefore of the form

$$(c\ ,c\ ,c\ ,c\ ).$$

In addition, the alleles *a* and *h* occur in some progeny, and the genotype therefore has the form

$$(ac, ch, c\ ,c\ ).$$

As further alleles do not occur in the progeny, there are six possible parental genotypes in all, namely

$$(ac, ch, ac, ac)\ (ac, ch, ac, cc)\ (ac, ch, ac, ch)$$

$$(ac, ch, cc, cc)\ (ac, ch, cc, ch)\ (ac, ch, ch, ch).$$

Although none of these parental genotypes can be ruled out entirely, some are far more plausible than others, based on the observed frequency of different offspring banding patterns. Table 3 shows the expected frequencies of different banding patterns for the six possible parental genotypes. Only the first genotype gives a satisfactory fit when judged by χ^2.

For three of the parental genotypes, χ^2 has four degrees of freedom rather than three, because the banding pattern *ah* can also arise. This banding pattern did not occur amongst the progeny that were recorded, and therefore has an observed frequency of zero. This is quite compatible with the expected frequency, which is only $57/256 = 0.22$ for each of the parental genotypes from which it can arise. Indeed, the probability that a banding pattern with probability 1/256 will not be observed in a sample of size 57 is $(255/256)^{57} = 0.800$. However, for these parental genotypes, the expected frequencies of other banding patterns agree poorly with the observed values.

Table 3 Frequency of different banding patterns for *PGM* in 57 selfed progeny of the strawberry cultivar 'Cambridge Favourite'. Expected frequencies and summary statistics are shown for each of the six possible parental genotypes

Banding pattern	Observed frequency	Expected frequency for different parental genotypes					
		(*ac*, *cc*, *cc*, *ch*)	(*ac*, *ac*, *cc*, *ch*)	(*ac*, *ac*, *ch*, *ch*)	(*ac*, *cc*, *ch*, *ch*)	(*ac*, *ac*, *ac*, *ch*)	(*ac*, *ch*, *ch*, *ch*)
ach	34	32.1	40.1	49.9	40.1	41.9	41.9
ac	14	10.7	13.4	3.3	2.7	14.0	0.7
ch	8	10.7	2.7	3.3	13.4	0.7	14.0
c	1	3.6	0.9	0.2	0.9	0.2	0.2
ah	0	0	0	0.2	0	0.2	0.2
χ^2		3.66	11.59	48.52	51.11	84.89	273.10
d.f.		3	3	4	3	4	4
P		0.30	< 0.01	< 0.001	< 0.001	< 0.001	< 0.001
Posterior probability		0.854	0.146	< 0.00001	< 0.00001	< 0.00001	< 0.00001

The Bayesian analysis also favours the parental genotype (*ac*, *cc*, *cc*, *ch*), assigning it a posterior probability of 0.85. The Bayes factor for comparing this model with its closest rival (*ac*, *ac*, *cc*, *ch*) is 5.85, indicating 'positive' evidence in favour of (*ac*, *cc*, *cc*, *ch*).

This is a simple example, primarily because one individual had only the single allele *c*, and consequently there were only six possible parental genotypes. It is possible that this individual was misrecorded and also carried one of the alleles *a* or *h*. If the individual is excluded from the analysis the number of possible parental genotypes increases to 25 and the parental genotypes (*ah*, *cc*, *cc*, *cc*) and (*ac*, *ah*, *cc*, *cc*) give satisfactory fits to the data as well as (*ac*, *cc*, *cc*, *ch*), when assessed by the chi-squared test. The Bayesian analysis assigns a posterior probability of only 0.05 to (*ac*, *cc*, *cc*, *ch*), compared to posterior probabilities of 0.52 and 0.40 for (*ah*, *cc*, *cc*, *cc*) and (*ac*, *ah*, *cc*, *cc*), respectively. Therefore, it provides 'positive' evidence that the parental genotype is *not* (*ac*, *cc*, *cc*, *ch*).

PGM segregation in selfed progeny of 'Elvira'

Six different banding patterns were observed in 57 offspring, as shown in Table 4. The SELF program identified 122 possible parental genotypes. The $m = 5$ distinct parental alleles can be arranged into 600 distinct parental genotypes (Table 2). It follows that 478 ($= 600 - 122$) genotypes have been eliminated because they were unable to generate the six observed banding patterns in Table 4.

The smallest χ^2 arose from the parental genotypes (*ae*, *cc*, *ee*, *fh*) and (*ac*, *cc*, *ee*, *fh*). It is impossible to discriminate between these parental genotypes using selfed progeny, because they give identical expected segregations. The expected frequencies are shown in Table 4; χ^2 is 10.27 on 5 d.f. ($P = 0.068$).

Table 4 also shows the expected frequencies arising from the parental genotypes (*ac*, *ce*, *ee*, *fh*) and (*ae*, *cc*, *ce*, *fh*); χ^2 is 12.61 on 8 d.f. ($P = 0.126$). The degrees of freedom are increased by three because these parental genotypes can give rise to three banding patterns, *aef, aeh* and *aefh*, which do not occur in the observed data. Based on the chi-squared *P*-values, these latter genotypes apparently provide a better fit to the observed frequencies, even though χ^2 is larger. Although two of the expected frequencies for these genotypes are slightly less than one, the large sample *P*-value appears to be accurate; using a Monte Carlo testing procedure (Hope, 1968) involving a simulation of 10 000 data sets of size 57 based on this parental genotype, 1273 sets gave χ^2 greater than the observed value of 12.61, implying an almost identical *P*-value of 0.127. Thus the observed banding patterns do not differ significantly from those expected from these genotypes when judged by the chi-squared test. However, the total probability of obtaining one of the three banding patterns *aef, aeh* or *aefh* which do not occur in the observed data is 16/256. The probability that none of these banding patterns will occur when there are 57 offspring is therefore $(240/256)^{57} = 0.025$. Assessed in this way, there is therefore a significant departure from the model.

This additional test is particularly important for analysing the 'Elvira' data. Of the 122 possible parental genotypes, only the genotypes (*ae*, *cc*, *ee*, *fh*) and (*ac*, *cc*,

Table 4 Frequency of different banding patterns for *PGM* in 57 selfed progeny of the strawberry cultivar 'Elvira'. Expected frequencies and summary statistics are shown for four of the 122 possible parental genotypes

Banding pattern	Observed frequency	Expected frequency for given parental genotype	
		(ae, cc, ee, fh) or *(ac, cc, ee, fh)*	*(ac, ce, ee, fh)* or *(ae, cc, ce, fh)*
acefh	18	21.4	19.6
acef	8	10.7	9.8
aceh	7	10.7	9.8
aefh	0	0	1.8
cefh	14	7.1	7.1
aef	0	0	0.9
aeh	0	0	0.9
cef	5	3.6	3.6
ceh	5	3.6	3.6
χ^2		10.27	12.61
d.f.		5	8
P		0.07	0.13
Posterior probability		0.458	0.026

ee, *fh*) do not give rise to banding patterns that did not occur in the observed data. Moreover, for the other 120 genotypes, the significance level associated with these unexpected banding patterns is always 0.025 or less. Most of these parental genotypes would also be ruled out on the basis of the chi-squared test, but a few would not. The most extreme examples are the parental genotypes (*ae*, *ce*, *ce*, *fh*) and (*ac*, *ce*, *ce*, *fh*) for which the chi-squared statistic is 15.38 on 14 d.f. ($P = 0.35$). However, there are nine banding patterns that do not occur in the observed data, and the test based on these gives a *P*-value of 0.00969. These discrepancies in *P*-values are not surprising; the chi-squared test is a general-purpose goodness-of-fit test that will not be as good as a specially designed test at detecting specific departures from the assumed model.

The Bayesian analysis favours the genotypes (*ae cc*, *ee*, *fh*) and (*ac*, *cc*, *ee*, *fh*) over (*ac*, *ce*, *ee*, *fh*) and (*ae*, *cc*, *ce*, *fh*), the Bayes factor being 17.7.

To summarize, the most likely *PGM* genotype for 'Elvira' appears to be (*ae*, *cc*, *ee*, *fh*) or (*ac*, *cc*, *ee*, *fh*), although the agreement between observed and expected frequencies of different banding patterns is not quite as good as one might hope (Table 4). Much of the discrepancy arises because of the large number of offspring that had the banding pattern *cefh*. Other possible parental genotypes appear unlikely, because of the absence of particular banding patterns in the observed data.

Extensions

Null alleles

In many species, the interpretation of electrophoretic data on isozyme segregation has relied on the introduction of *null* alleles, which are inherited in the normal way but which do not generate a visible band on the gel. The SELF program has an option to allow for the possibility that some of the parental alleles may be null. If null alleles are allowed, the number of possible parental genotypes that must be considered increases, as shown in the rightmost column of Table 2. In fact, it is easy to see that, for $m < 8$, the number of possible parental genotypes when null alleles are allowed is the sum of the number of possible parental genotypes that can arise when there are m or $m + 1$ distinct alleles and null alleles are *not* allowed.

For 'Cambridge Favourite', the minimum value of the chi-squared statistic was unchanged by the introduction of null alleles. Without null alleles, the mostly likely parental genotype for *PGM* was *(ac*, *cc*, *cc*, *ch)*. If null alleles are introduced in any combination of the following ways:

(i) replace *ac* by *an*
(ii) replace *ch* by *hn*
(iii) replace one of the *cc*'s by *cn* or *nn*,

then the probabilities of the different banding patterns that can arise in the selfed progeny are unaffected. There are thus 12 ($= 2 \times 2 \times 3$) genotypes, 11 of them involving null alleles, which give the minimum value of χ^2 (3.66).

The minimum value of χ^2 for *PGM* segregation in 'Elvira' was also unchanged by the introduction of null alleles. The minimum value arose from the genotype (*an*, *cc*, *ee*, *fh*) as well as (*ae*, *cc*, *ee*, *fh*) and (*ac*, *cc*, *ee*, *fh*).

Locus not represented on all chromosome pairs

The locus of interest may not be present on all four chromosome pairs. The SELF program allows the number of chromosome pairs on which the locus is present to be specified as 2, 3 or 4. This allows the analysis to be extended to hexaploids (e.g. plum) and tetraploids (e.g. sour cherry).

The most likely *PGM* genotype for 'Cambridge Favourite' was *(ac*, *cc*, *cc*, *ch)*, assuming that null alleles are not involved. Clearly one of the *cc* pairs can be excluded without altering the possible segregation patterns. Based on this cultivar, it is therefore possible that the *PGM* locus is located on only three of the four chromosome pairs. It is even possible that the genotype is simply *(ac*, *ch)*, though this is unlikely because the banding pattern *ah* did not occur and the probability of this is only 0.025.

For 'Elvira', there are five distinct *PGM* alleles; so the only possibility that needs to be considered is that the locus occurs on only three out of the four chromosome pairs. However, although there are 15 possible parental genotypes, all of these give rise to banding patterns whose absence from the observed data is completely implausible ($P < 0.0001$). The conclusion from 'Elvira', and from unpublished data on other cultivars, is that the *PGM* locus is present on all four pairs of chromosomes. Arulsekar *et al.* (1981) proposed that the loci for phosphoglucoisomerase (*PGI*) and leucine aminopeptidase (*LAP*) were also present on all four pairs of chromosomes.

Discussion

Although both examples have involved octoploid strawberry cultivars, the fact that the locus can be specified to occur on only two or three of the chromosome pairs means that the SELF program can also be used to analyse data from allotetraploids or allohexaploids.

Ignoring the possibility that null alleles may be present, the *PGM* genotype for 'Cambridge Favourite'

appears to be well established from the data, providing that the reading of the gel with only a single band can be considered reliable. The 'Elvira' data illustrate the fact that, even without null alleles, certain genotypes may be indistinguishable on the basis of selfed progeny, because they give identical probabilities for the different banding patterns. Indeed, even crossing to a completely homozygous 'tester' would not distinguish between these pairs of genotypes when only the presence or absence of alleles can be observed in the progeny. Although one particular pair of genotypes appears to provide the best fit to the *PGM* segregation in 'Elvira', the fit is not especially good ($P = 0.07$). It would be useful to have additional data to determine whether this is simply a chance effect, or whether a different model is needed.

Even when parental genotypes do not give identical segregation patterns, they may give patterns that are difficult to distinguish without very large samples. For example, for the strawberry cultivar 'Bogota', the parental genotypes that gave the two smallest χ^2 values imply banding pattern probabilities that differ by $4/256 = 0.016$ at most. One of the genotypes gives rise to a banding pattern that does not occur in the data, but the probability of this banding pattern is only 0.016. The nonoccurrence of this banding pattern would not be considered unusual ($P < 0.05$) unless the sample size exceeded 185.

The Bayesian analysis proposed by Olson (1997) appears to be quite successful in analysing selfed progeny. The assignment of prior probabilities in Bayesian analysis is often difficult but here the assumption that all possible parental genotypes have the same prior probability seems uncontroversial. In the examples given, the conclusions from the Bayesian analyses agree with the conclusions based on the chi-squared tests when the chi-squared tests are supplemented by tests related to the absence of particular offspring genotypes in the observed data. These latter tests are necessitated by the difficulties associated with degrees of freedom for χ^2.

The Bayesian analysis appears simpler to interpret, because it provides only a single value. Nonetheless, it is important to calculate an absolute measure of goodness-of-fit, such as χ^2, as there is no guarantee that the parental genotype with the highest posterior probability will necessarily provide a good fit. For example, we have noted earlier that by assuming that the *PGM* locus is located on only three of the four chromosome pairs gives rise to a substantial lack of fit. Nonetheless, the two (indistinguishable) parental genotypes (*ac*, *ee*, *fh*) and (*ae*, *cc*, *fh*) have a combined posterior probability exceeding 0.99. The difficulty here is that the Bayesian analysis assumes that the correct model is included in the set of possible models. In this example, no model in which the locus is present on only three chromosome pairs is satisfactory. In principle, it would be possible to consider models involving both three and four chromosome pairs simultaneously, but it is then not clear how prior probabilities should be assigned. A referee has also pointed out that the Bayesian analysis is not immune from the problems of variable sample space that cause difficulties in deciding on appropriate degrees of freedom for χ^2; in the Bayesian context one model may have higher posterior probability than another partly because it has a smaller sample space.

Although this paper has concentrated on selfing, we do not wish to imply that this is necessarily the best approach to determining parental genotypes in allopolyploids; indeed, the development of efficient crossing strategies is a difficult, but interesting, problem. One formulation of the problem is as follows. Given a set of c cultivars for which the distinct alleles are known, but not the allele frequencies, devise a set of n crosses which, when taken together, will maximize the information about parental genotypes. It would be useful to develop a computer program to generate suitable crossing schemes. An important first step in developing such a program would be to generalize the SELF program to allow crosses between parents with different genotypes.

Acknowledgements

This work was done as part of project HH0113SSF, funded by the Ministry of Agriculture, Fisheries and Food. We thank C Lefebvre for assistance with the experiments and with initial data analysis, J P McSorley for discussions of the combinatorial aspects of pairing, and K R Tobutt and J R Lynn for helpful comments on the manuscript.

References

ARULSEKAR, S., BRINGHURST, R. S. AND VOTH, V. 1981. Inheritance of PGI and LAP isozymes in octoploid cultivated strawberries. *J. Am. Soc. Hort. Sci.*, **106**, 679–683.

BELL, J. A. AND SIMPSON, D. W. 1994. The use of isoenzyme polymorphisms as an aid for cultivar identification in strawberry. *Euphytica*, **77**, 113–117.

BRINGHURST, R. S. 1990. Cytogenetics and evolution in American *Fragaria*. *Hortsci.*, **25**, 879–881.

GALLETTA, G. J. AND MAAS, J. L. 1990. Strawberry genetics. *Hortscience*, **25**, 871–879.

HOPE, A. C. A. 1968. A simple Monte Carlo test procedure. *J. R. Statist. Soc. B*, **30**, 582–598.

KASS, R. E. AND RAFTERY, A. E. 1995. Bayes factors. *J. Am. Stat. Ass.*, **90**, 773–795.

OLSON, M. S. 1997. Bayesian procedures for discriminating among hypotheses with discrete distributions: inheritance in the tetraploid *Astilbe biternata*. *Genetics*, **147**, 1933–1942.