Sample size for collecting germplasms – a polyploid model with mixed mating system

R L SAPRA†, PREM NARAIN*, S V S CHAUHAN**, S K LAL and B B SINGH

Genetics Division, Indian Agricultural Research Institute, Pusa, New Delhi 110 012, India
**B-3/27A, Lawrence Road, New Delhi 110 035, India*
***Department of Botany, School of Life Sciences, BR Ambedkar University, Agra 282 002, India*

†*Corresponding author (Fax, 91-11-25766420; Email, saprarl@rediffmail.com)*

The present paper discusses a general expression for determining the minimum sample size (plants) for a given number of seeds or vice versa for capturing multiple allelic diversity. The model considers sampling from a large 2 k-ploid population under a broad range of mating systems. Numerous expressions/results developed for germplasm collection/regeneration for diploid populations by earlier workers can be directly deduced from our general expression by assigning appropriate values of the corresponding parameters. A seed factor which influences the plant sample size has also been isolated to aid the collectors in selecting the appropriate combination of number of plants and seeds per plant. When genotypic multiplicity of seeds is taken into consideration, a sample size of even less than 172 plants can conserve diversity of 20 alleles from 50,000 polymorphic loci with a very large probability of conservation (0·9999) in most of the cases.

[Sapra R L, Narain P, Chauhan S V S, Lal S K and Singh B B 2003 Sample size for collecting germplasms – a polyploid model with mixed mating system; *J. Biosci*. **28** 155–161]

1. Introduction

Several probability models have been developed over the past three decades which can estimate minimum or optimal sample size (plants as well as seeds) for the purpose of collection, regeneration and conservation of plant germplasm. Minimum sample size required for conserving allelic, genotypic and phenotypic diversity has been discussed in a number of papers (Oka 1969; Allard 1970; Bennett 1970; Marshal and Brown 1975; Oka 1975; Qualset 1975; Bogyo *et al* 1980; Gregorius 1980; Hawkes 1980; Chapman 1984; Yonezawa 1985; Namkoong 1988; Yonezawa and Ichihashi 1989; Crossa 1989; Brown and Briggs 1991; Crossa *et al* 1993; Brown and Marshall 1995; Lawrence *et al* 1995a,b; Sapra *et al* 1998). Discussions have focused from diallelic with mono locus models to multiallelic and multilocus models. Yonezawa and Ichihashi (1989) were the first to study mathematically the contribution of seeds per plant to the outcome of the sampling by taking appropriate combination of the number of plants and number of seeds per plant. They considered a diploid model and worked out sample sizes by taking 2 or 3 alleles at each of the 2 or 4 independent loci. However, they could not come out with a general expression which can clearly give required sample sizes for capturing multi allelic and multi loci variability. The purpose of the present paper, is therefore, to develop a general expression by considering the probability of all the alleles contained in a target population and to provide extra flexibility and savings to the collector in terms of seeds as well as plants depending upon the species.

2. Diploid model

2.1 *Diallelic locus*

In a particular species, the plant sample size primarily depend upon the number of alleles, their frequencies, number of loci, ploidy level and the mating behaviour. In

Keywords. Conservation; diploid; exploration; germplasm; inbreeding; polyploid; seeds

considering the problem of sample size, it is convenient to begin by considering the simplest case of conserving a pair of alleles at a single locus in a sample drawn at random from a population of plants. The analysis of the problem is then extended to the situation where there are more than two alleles per locus at several of the independent loci. First we will deal with sampling of plants without considering the contribution of seeds to the outcome of sampling. Let us consider a diploid plant population as the target population with two alleles A_1 and A_2 having frequencies p_1 and p_2 at a locus and reproducing seeds by constant proportion of selfing (s) and random mating (1-s). These two alleles have been assumed to be selectively neutral and the population is at genetic equilibrium. Suppose we wish to conserve a gene at a single locus. The genetical variation at this locus will have been completely conserved when a sample drawn from the population contains at least one copy of each of the alleles. We assume the species in question is diploid one, and the frequencies of the three genotypes (A_1 A_1, A_1A_2, A_2 A_2) at this locus be G_1, G_2, and G_3. The probability P, that a randomly drawn sample of size m plants contains at least one copy of each allele then is,

P = 1 – probability of missing A_1 in m number of plants

– probability of missing A_2 in m number of plants

$= 1 - (G_3)^m - (G_1)^m$,

where G_1, G_2 and G_3 are the frequencies of A_1A_1, A_1A_2 and A_2A_2 in the target population and are related to the allelic frequencies as:

$G_1 = p_1^2 (1\text{-}F_1) + p_1F_1$;

$G_2 = 2\, p_1p_2\, (1\text{-}F_1)$;

$G_3 = p_2^2 (1\text{-}F_1) + p_2F_1$,

and, $F_1 = s/(2\text{-}s)$ is the inbreeding coefficient measuring the deviation from random mating. Let the allele A_2 be considered as rare in nature with a frequency p_0. The term $(G_3)^m$ then becomes negligible, can be ignored, and the probability expression now reduces to,

$$P \approx 1 - (G_1)^m.$$

If now there are l independent loci, collecting both the alleles at all the loci will yield the probability expression as,

$$P \approx (1 - G_1^m)^l.$$

The number of plants (m) to be sampled can be easily determined by evaluating the above said expression. Dealing with more than 2 alleles becomes complicated, involving the use of principle of inclusion and exclusion of probabilities. We shall now extend our mathematical treatment to the contribution of seeds per plant to the outcome of the sampling. Yonezawa and Ichihashi (1989) formulated an expression for determining number of plants and seeds/plant to be sampled for including at least one copy of both the alleles with a given probability of conservation (P) as:

$$P = 1 - (G_1 g_{11}^n + G_2 g_{21}^n)^m - (G_2 g_{23}^n + G_3 g_{33}^n)^m, \quad (1)$$

where n is the number of seeds sampled from a single plant; g_{11}, g_{21}, g_{23} and g_{33} are the frequencies of genotypes (seeds) produced on single plants, as shown in table 1 which is reproduced from Yonezawa and Ichihashi (1989) for better understanding.

Let us assume that A_2 is rare in nature, has a frequency ($p_2 = p_0$) of the order of 0·05 or less, and the explorer samples large number of seeds from each of the plant; the contribution of all the terms except $G_1g_{11}^n$, in the

Table 1. Genotypic array in target population and seeds on single plants.

Plants in population		Seeds on single plants				
Genotype	Frequency	Genotype	Frequency in selfed seeds	Frequency in out crossed seeds	Frequency in the total	Designation
A_1A_1	G_1	A_1A_1	1	p_1	$s + (1\text{-}s)p_1$	g_{11}
		A_1A_2	0	p_2	$(1\text{-}s)p_2$	g_{12}
		A_2A_2	0	0	0	g_{13}
A_1A_2	G_2	A_1A_1	1/4	$p_1/2$	$s/4 + (1\text{-}s)p_1/2$	g_{21}
		A_1A_2	1/2	1/2	$s/2 + (1\text{-}s)/2$	g_{22}
		A_2A_2	1/4	$p_2/2$	$s/4 + (1\text{-}s)p_2/2$	g_{23}
A_2A_2	G_3	A_1A_1	0	0	0	g_{31}
		A_1A_2	0	p_1	$(1\text{-}s)p_1$	g_{32}
		A_2A_2	1	p_2	$s + (1\text{-}s)p_2$	g_{33}

expression (1) now is almost negligible. Therefore with elimination of these terms, the expression (1) now reduces to:

$$P \approx 1\text{-}\ (G_1\, g_{11}{}^n)^m. \qquad (1.1)$$

Taking logarithm on both sides of (1.1) and simplifying, we arrive at the following expression for minimum plant sample size as:

$$m > \log\,(1\text{-}P)/(\log\, G_1 + n\, \log\, g_{11}). \qquad (1.2)$$

After assigning the values of G_1 and g_{11}, the final expression we arrive at is:

$$m > \log\,(1\text{-}P)/\Big\{\log\,[\,(1\text{-}p_0)^2\,(1\text{-}F_1) + (1\text{-}p_0)F_1\,] + n\ \log\,[s + (1\text{-}s)(1\text{-}p_0)]\Big\}. \qquad (2)$$

From expression (2), we can deduce two important results:

$$\text{(i) } m > \log\,(1\text{-}P)/\log\,(1\text{-}p_0);\ \text{when } s = 1 \qquad (2.1)$$

$$\text{(ii) } m > \log\,(1\text{-}P)/\Big\{(2 + n)\ \log\,(1\text{-}p_0)\Big\};\ \text{when } s = 0 \qquad (2.2)$$

2.2 *Multiallelic multilocus*

Expression (2) estimates the minimum number of plants to be sampled for a given number of seeds for including both the alleles of a locus only. The equations for handling more than 2 alleles at a locus become quite complicated as the number of alleles increase. Crossa *et al* (1993), and Hernandez and Crossa (1993) developed these equations and computer algorithm for determining optimum plant sample size for regeneration of germplasm. However they didn't consider the selfing parameter, as well as the effect of number of seeds/plant on the sample size (plants) in their model. They developed a simplified equation for (a) number of alleles per locus at all the l independent loci subject to the condition that (a-1) alleles occur with an identical low frequency of (p_0), and that the ath allele occurs with a frequency of [1-(a-1)p_0]. This kind of assumption is more logical for two reasons: (i) the sample size is mainly affected by the frequencies of rare alleles; and (ii) obtaining the required sample size for many alleles at different frequencies is very impractical. The algebraic treatment is quite cumbersome (in our case) when the number of alleles are large. Nevertheless, we have deduced expression (3) using the results of Crossa *et al* (1993) for multiallelic and multilocus diploid population, and our results (1.1). However, in the § 4, we shall provide an empirical verification of our expression (3).

Crossa *et al* (1993) used the following probability expression for 'a' allelic classes,

$$P \approx 1\text{-}(a\text{-}1)(1\text{-}p_0)^m. \qquad (2.3)$$

Comparing the above said Crossa *et al* (1993) results with (1.1), it is clear that the term (a-1) controls the multialleic situation while the term $(1\text{-}p_0)^m$ takes care of the rare alleles. Using the term (a-1) in (1.1) we arrive at,

$$P \approx 1 - (a\text{-}1)(G_1\, g_{11}{}^n)^m. \qquad (2.4)$$

If there are l independent loci, (2.4) reduces to

$$P \approx [\ 1 - (a\text{-}1)\ (G_1\, g_{11}{}^n)\ m\ ]^l. \qquad (3)$$

Expression (3) for minimum sample size can be further written with an inequality sign (>), after incorporating the values of G_1 and g_{11}, as:

$$m > \Big\{\log\,[\,(1\text{-}(P)^{1/l}\,] - \log\,(a\text{-}1)\Big\}/\Big\{\log\,[\,(1\text{-}p_0)^2\ (1\text{-}F_1)_{+}(1\text{-}p_0)F_1)] + n\ \log\,[s + (1\text{-}s)(1\text{-}p_0)\,]\Big\} \qquad (4)$$

As expected, when $a = 2$ and $l = 1$ expression (4) reduces to (2).

3. Auto-polyploid model

3.1 *Diallelic locus*

Let us consider a 2 k-auto-polyploid population as the target population with two alleles, A_1 and A_2, at a single locus, having frequencies p_1 and p_2, respectively, reproducing seeds by constant proportions of selfing (s) and random mating (1-s), with no double reduction or selection under chromosomal segregation. Such a population at equilibrium can be denoted as:

$$Z \equiv \left(\frac{A_1^{2k}}{p_1^{2k}(1-F_k) + p_1F_k}, \frac{A_1^{2k-1}A_2}{{}_{2k}C_1 p_1^{2k-1} p_{2(1-F_k)}}, \frac{A_1^{2k-i}A_2^i}{{}_{2k}C_i\, p_1^{2k-i}\, p_2^i(1-F_k)}, \ldots \frac{A_2^{2k}}{p_2^{2k}(1-F_k) + p_2F_k} \right).$$

The sum of the allelic frequencies and the sum of the genotypic frequencies as given above is one. F_k is the theoretical populational inbreeding coefficient at equilibrium for a 2 k-ploid organism, and is related to the proportion of selfing (s) by the formula (McConnell and Fyfe 1975):

$$F_k = s/\{2k\text{–}(2k\text{–}1)s\}.$$

When $k = 1$, the population becomes a diploid and can be represented as:

$$Z \equiv \left(\frac{A_1^2}{p_1^2(1-F_1) + p_1F_1} \frac{A_1A_2}{2\,p_1p_2(1-F_1)} \frac{A_2^2}{p_2^2(1-F_1) + p_2F_1} \right).$$

Let us again assume allele A_2 to be rare, having a frequency (p_0), and the collector samples large number of seeds from each of the selected plant of the target population. We can now formulate probability expression for auto-polyploids similar to expression (1) which we desired for diploids for sampling seeds for retaining at least a copy of both the alleles (A_1, A_2) with a certain degree of probability. For formulating such expressions, tables can be easily generated for auto-tetraploids, auto-hexaploids etc. in the same way as for diploids represented in table 1. On dropping the terms other than $G_1 g_{11}^n$ and comparing with the diploid expression (2), we can write an expression for auto-polyploids by merely incorporating the ploidy parameter, (k), as follows:

$$m > \log (1\text{-}P) / \{\log [\ (1\text{-}p_0)^{2k} (1\text{-}F_k)] + n \log [s + (1\text{-}s)(1\text{-}p_0)^k\ ]\}. \quad (5)$$

3.2 *Multiallelic multilocus*

Writing expressions for multiallelic and multilocus situation is straightforward and leads to the following expression by incorporating the ploidy parameter. Using (3), (4) and (5), and after simplification, we get the final expression (6),

$$m > A/(B + C + D); \quad (6)$$

where $A = \log [\ (1\text{-}(P)^{1/l}\ ] - \log (a\text{-}1)$,

$B = \log (1\text{-}p_0)$,

$C = \log [\ (1\text{-}p_0)^{2k-1} (1\text{-}F_k) + F_k]$,

$D = n \log [s + (1\text{-}s)(1\text{-}p_0)^k]$.

4. Empirical verification

For empirical verification of expression (6), we constructed three tables similar to table 1, by writing down the genotypic frequencies for plants as well as genotypic frequencies of seeds borne on individual plant for diploid species with 3 and 4 alleles, and for auto-tetraploid species with 3 alleles. Three large probability expressions, as represented by expression (1) were formulated. For verification of our expression, we considered ten levels of loci (1, 2, 10 20, 100, 1000, 5000, 10000 20000, 50000), two levels of probability of conservation (95%, 99·99%), two levels of rare allele frequency (0·05, 0·01), eleven levels of selfing (0, 0·1, 0·2, 0·3, 0·4, 0·5, 0·6, 0·7, 0·8, 0·9, 1·0) and 14 seeds number (1, 2, 3, 4, 5, 10, 15 20, 25, 30, 35, 40, 45, 50) for calculation of sample sizes. The sample sizes obtained from these probability expressions were compared with those obtained from our simplified expression (6). It was observed that the exact values (calculated from the said probability expressions) were in close agreement with those calculated by using expression (6) for $n > 5$. However, for $n \leq 5$, the developed expression underestimated the plant sample sizes for a given number of seeds. Thus, expression (6) can be safely utilized with accuracy over varied conditions for determination of minimum sample sizes for large populations.

5. Discussion

The above derived expression (6) is the most generalized one for determining minimum number of plants required for collecting the multiallleic and multilocus diversity from large populations, keeping in view the genotypic multiplicity of seed embryos borne on a single plant. By assigning certain values to the parameters, this interesting expression yields several important and useful results derived by some of the earlier workers, as given below:

(i) When $k = 1$ it gives results for diploid populations.
(ii) When $k = 1$ and $s = 1$ it reduces to $m > A/B$. This is the same equation as developed by Crossa *et al* (1993) for determining the optimal sample size for capturing all the alleles from a homogeneous diploid population. Interestingly, the equation $m > A/B$ does not contain the parameter for number of seeds (n) indicating that the plant sample size is not affected by the number of seeds sampled from each plant if the population is an inbred one i.e. ($s = 1$).
(iii) When $k = 1$, $s = 1$ and $a = 2$, the expression (6) reduces to the one as suggested by Chapman (1984).
(iv) When $k = 1$, $a = 2$, $l = 1$ and $s = 1$, our expression gives the same results as given by the gametic model of Marshall and Brown (1975) for selfed diploid population; and when $k = 1$, $a = 2$, $l = 1$, $s = 0$ and $n = 0$, it gives results for apomictic diploid population.
(v) When $k = 1$ and $n > 5$, the results obtained agree with those of Yonezawa and Ichihashi (1989). However, when the number of seeds sampled per plant is taken as less than 5, the model may underestimate the plant sample size.
(vi) When the genotypic multiplicity of seed embryos borne on single plant are not taken into consideration as proposed by Yonezawa and Ichihashi (1989) i.e. $n = 0$, the expression (6) reduces to $m > A/(B + C)$, which is the same as developed by Sapra *et al* (1998) for determining the minimum plant sample size (vegetative) to conserve all the alleles at all the independent loci in a 2 k-ploid population with varying degrees of selfing. The component 'C' identified here accounts for reduction in plant sample size owing to deviation from selfing or diploidy or both (Sapra *et al* 1998). The results obtained by Law-

rence *et al* (1995a) can be easily obtained from the expression $m > A/(B + C)$.

(vii) The component '*D*' in expression (6) is a 'seed factor' and determines the efficiency in terms of number of plants to be sampled. How many seeds per plant should be sampled, or whether we can achieve greater efficiency by collecting more seeds from a smaller number of plants, depends upon the value of this seed factor. The value of *D* lies between $nk[\log(1\text{-}p_0)]$ and 0. The lower limit is attained under exclusive random mating in out crossing population ($s = 0$) and the upper limit is attained when the population is either highly inbred one ($s = 1$) or n is taken as zero. Thus, the seed factor, '*D*' provides the explorer and additional flexibility for selecting an appropriate combination of plants and seeds/plant.

(viii) The expression (6) gives following three additional useful results:

$$m_{1r}/m_{kr} \approx k, \tag{6.1}$$

$$m_{1s}/m_{ks} \approx 1, \tag{6.2}$$

$$m_{ks}/m_{kr} \approx k(2 + n), \tag{6.3}$$

where m_{1r} and m_{kr} are the number of plants sampled from a diploid population and polyploid population respectively, under exclusive random mating for a given set of parameters; and m_{1s} and m_{ks} are the number of plants sampled from diploid population and polyploid population for a given set of parameters under exclusive selfing.

The result (6.1) indicates that auto-polyploidy is almost k times conservative of genetic variability in comparison to diploid population under random mating. The result (6.2) is interesting, and indicates that under exclusive selfing the plant sample size for a given set of parameters gets fixed irrespective of their ploidy level. The result (6.3) indicates that the number of plants to be sampled from a random mating population for a given set of parameters is almost inversely proportional to the product of ploidy level and number of seeds, when the number of seeds sampled are very large.

In our situation we have assumed (a-1) alleles occur with an identical low frequency at all the loci. Though this situation will be hardly met in the actual field conditions, yet, it will set a upper limit of the sample size for all practical purposes. We calculated theoretical plant sample sizes for 10, 50, 100 and 500 seeds per plant under varying degrees of selfing for conserving alleles as large as 20 (one abundant and 19 rare alleles with frequency 0·01) from 50,000 polymorphic loci (table 2). Our calculations show that for a fully cross-fertilizing species, a plant sample size of 23 is sufficient for conserving diversity when 100 seeds are collected from each plant whereas only 5 plants are to be sampled when 500 seeds are collected from each plant. Plant sample size increases slowly with the increase in selfing rate up to 90% of selfing. However, beyond 90%, the sample size increases drastically with the increase in selfing rate and attains a maximum value of 2286 when the species set the seeds fully through self-pollination ($s = 1$). At this stage the plant sample size becomes fixed and can not be reduced by increasing the number of seeds as all the seeds are in homozygous and homogeneous conditions.

Table 2. Plant sample size under varying degrees of selfing for capturing 20 alleles (one abundant and 19 rare alleles occurring at a frequency of 1% each) at each of the 50000 loci with a probability of conservation of 0·9999 when 10, 50, 100 and 500 seeds are sampled from each plant.

	Number of plants			
Selfing (s)	10 seeds	50 seeds	100 seeds	500 seeds
0·00	191	44	23	5
0·10	209	49	25	6
0·20	232	55	28	6
0·30	260	63	32	7
0·40	296	73	38	8
0·50	344	86	45	10
0·60	412	107	56	12
0·70	514	140	73	16
0·80	688	203	108	23
0·90	1051	372	206	45
0·91	1110	406	226	50
0·92	1177	446	251	56
0·93	1252	496	283	64
0·94	1337	558	323	74
0·95	1436	639	377	89
0·96	1551	746	453	109
0·97	1686	897	566	144
0·98	1847	1124	755	209
0·99	2043	1507	1135	382
1·00	2286	2286	2286	2286

Table 3 gives the seed sample sizes under varying degrees of selfing when 23 plants are sampled (23 is the plant sample size required under exclusive random mating conditions in table 2). The table shows that up to 80% selfing, less than 500 seeds will be sufficient for capturing the said variability and beyond 90% the seed sample size may be very large ranging from a thousand to ten-thousand per plant. Lawrence *et al* (1995a) suggested that a sample of about 172 plants, drawn at random from a population of target species, is of sufficient size to conserve at a very high probability, all or very nearly all of the polymorphic genes that are segregated in the population, provided their frequency is not less than 0·05, irrespective of whether the individuals of the species set all of their seed by self- or by cross–fertilization or a mixture of both. Table 4 shows the values of various seed sample sizes to be collected for some 172 plants as

Table 3. Seed sample size under varying degrees of selfing for capturing 20 alleles (one abundant and 19 rare alleles occurring at a frequency of 1% each) at each of the 50000 loci with a probability of conservation of 0·9999 when 23 plants are sampled.

Selfing (*s*)	Seeds (*n*)	Selfing (*s*)	Seeds (*n*)
0·00	98	0·91	1098
0·10	110	0·92	1235
0·20	125	0·93	1410
0·30	145	0·94	1650
0·40	164	0·95	1980
0·50	198	0·96	2470
0·60	246	0·97	3295
0·70	328	0·98	4945
0·80	495	0·99	9890
0·90	987	1·00	

Table 4. Seed sample size corresponding to a plant sample size of some 172 plants (proposed by Lawrence *et al* 1995a) under varying degrees of selfing for capturing 20 alleles (one abundant and 19 rare alleles occurring at a frequency of 1% each) at each of the 50000 loci with a probability of conservation of 0·9999.

Selfing (*s*)	Seeds (*n*)	Plants (*m*)
0·00	11–12	164–176
0·10	12–13	168–180
0·80	63	173
0·90	122	172
0·99	1240	172

suggested by Lawrence *et al* (1995a). Compared to Lawrence *et al* (1995a) model, our model provides the explorer an extra flexibility and higher certainty for collecting all the genes with a frequency as low as 0·01.

6. Conclusions

We have developed a single simplified expression which can give ample flexibility to the practical explorer for deciding minimum sample size (a suitable combination of number of plants and seeds/plant) for capturing allelic diversity during the field collection. A seed factor that is an indicative of efficiency in terms of number of plants has also been isolated in the expression. Our task has been to find the minimum plant sample size for a given number of seeds or vice versa to give a very high probability of conserving all of the genetical variation of the species. This model provides the explorer an option of flexibility while collecting for conservation purposes suited to the varying field conditions, and limiting resources. We believe our recommendations could lead to considerable saving of resources during collection and storage of material in gene banks.

References

Allard R W 1970 Population structure and sampling methods; in *Genetic resources in plants – their exploration and conservation* (eds) O H Frankel and E Bennett (Oxford, Edinburgh: Blackwell) pp 97–107

Bennett E 1970 Adaptation in wild and cultivated plant populations; in *Genetic resources in plants – their exploration and conservation* (eds) O H Frankel and E Bennett (Oxford, Edinburgh: Blackwell) pp 115–129

Bogyo T P, Poreceddu E and Perrino P 1980 Analysis of sampling strategies for collecting genetic material; *Econ. Bot.* **34** 160–174

Brown A H D and Marshall D R 1995 A basic sampling strategy: theory and practice; in *Collecting plant genetic diversity* (eds) L Guarino, V Ramanath Rao and R Reid (Wallingford, Oxon: CAB International) pp 75–91

Brown A H D and Briggs J D 1991 Sampling strategies for genetic variation in *ex situ* collections and endangered plant species; in *Genetics and conservation of rare plants* (eds) D A Falk and K E Holsinger (New York: Oxford University Press) pp 99–119

Chapman C G D 1984 On the size of genebank and the gene tic variation it contains; in *Crop genetic resources: conservation and evaluation* (eds) J H W Hoden and J T Williams (London: Allen and Unwin) pp 102–108

Crossa J 1989 Methodologies for estimating the sample size required for conservation of outbreeding crops; *Theor. Appl. Genet.* **77** 153–161

Crossa J, Harnandez C M, Bretting P, Eberhart S A and Taba S 1993 Statistical genetic considerations for maintaining germplasm collections; *Theor. Appl. Genet.* **86** 673–678

Gregorious H R 1980 The probability of loosing an allele when diploid genotypes are sampled; *Biometrics* **36** 643–652

Hawkes J G 1980 *Crop genetic resources and field collection manual* (IBPGR and EUCARPIA: University of Birmingham) pp 1–37

Hernandez C M and Crossa J 1993 A program for estimating the optimum sample size for germplasm conservation: *J. Hered.* **84** 85–86

Lawrence M J, Marshall D F and Davies P 1995a Genetics of genetic conservation I; Sample size when collecting germplasm: *Euphytica* **84** 89–99

Lawrence M J, Marshall D F and Davies P 1995b Genetics of genetic conservation II; Sample size when collecting seed of cross pollinating species and the information that can be obtained from the evaluation of material held in gene banks; *Euphytica* **84** 101–107

Marshall D R and Brown A H D 1975 Optimum sampling strategies in genetic conservation; in *Crop genetic resources today and tomorrow* (eds) O H Frankel and J G Hawkes (Cambridge: Cambridge University Press) pp 53–80

McConnell Gillian and Fyfe J L 1975 Mixed selfing and random mating with polysomic inheritance; *Heredity* **34** 271–272

Namkoong G 1988 Sampling for germplasm collections; *Hortic. Sci.* **23** 79–81

Oka H I 1969 A note on the design of germplasm presentation work in grain crops; *SABRAO Newslet.* **1** 127–134

Oka H I 1975 Consideration on the population size necessary for conservation of crop germplasms; in *JIBP synthesis* (ed.) T Matsuo (Tokyo: Tokyo University Press) Vol. 5, pp 57–63

Qualset C O 1975 Sampling germplasm in a center of diversity: an example of disease resistance; in *Crop genetic resources today and tomorrow* (eds) O H Frankel and J G Hawkes (Cambridge: Cambridge University Press) pp 81–96

Sapra R L, Prem Narain and S V S Chauhan 1998 A general model for sample size determination for collecting germplasm; *J. Biosci*. **23** 647–652

Yonezawa K 1985 A definition of the optimal allocation of effort in conservation of plant genetic resources – with application to sample size determination for field collection; *Euphytica* **34** 345–354

Yonezawa K and H Ichihashi 1989 Sample size for collecting germplasm from natural plant population in view of genotypic multiplicity of seed embryos borne on a single plant; *Euphytica* **41** 91–97

MS received 17 September 2002; accepted 17 December 2002

Corresponding editor: S MAHADEVAN