
A general model for sample size determination for collecting germplasm

R L SAPRA[†], PREM NARAIN* and S V S CHAUHAN**

NBPGR, New Delhi 110 012, India

*50 Director's Office, Indian Agricultural Research Institute, Pusa, New Delhi 110 012, India

**Department of Botany, B R Ambedkar University, Agra, India

Corresponding author: C/o Dr Prem Narain, 50 Director's Office, Indian Agricultural Research Institute, Pusa, New Delhi 110 012, India (Fax, 91-11-5731495; Email, root@naasdel.ren.nic.in).

The paper develops a general model for determining the minimum sample size for collecting germplasm for genetic conservation with an overall objective of retaining at least one copy of each allele with preassigned probability. It considers sampling from a large heterogeneous $2k$ -ploid population under a broad range of mating systems leading to a general formula applicable to a fairly large number of populations. It is found that the sample size decreases as ploidy levels increase, but increases with the increase in inbreeding. Under exclusive selfing the sample size is the same, irrespective of the ploidy level, when other parameters are held constant. Minimum sample sizes obtained for diploids by this general formula agree with those already reported by earlier workers. The model confirms the conservative characteristics of genetic variability of polysomic inheritance under chromosomal segregation.

1. Introduction

Plant explorers and conservationists are faced with the problem of collecting genetic material (vegetative or seeds) from large populations with a view to conserving the germplasm with a certain degree of assurance. The number of plants needed to conserve the germplasm has been discussed in many papers (Allard 1970; Marshall and Brown 1975; Qualset 1975; Bogyo *et al* 1980; Gregorius 1980; Chapman 1984; Yonezawa 1985; Namkoong 1988; Crossa 1989; Yonezawa and Ichihashi 1989; Crossa *et al* 1993; Lawrence *et al* 1995a,b) using probability models mainly for diploid species.

Lawrence *et al* (1995a) suggested a sample of about 172 plants for conserving all or very nearly all of the polymorphic genes with high probability provided the 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 by a mixture of both. They generalized their model for multiple loci with 2 alleles

under extreme cases i.e., complete selfing or complete random mating. Crossa *et al* (1993) while discussing the optimal sample size for regeneration suggested sample sizes of 160–210 plants for capturing alleles at frequencies of 0.05 or higher in each of 150 loci, with 90–95% probability. When allele frequencies are unknown an equation for estimating an optimal sample size for capturing $(a-1)$ rare alleles having an identical frequency of (p_0) and the a th allele occurring at a frequency of $\{1 - (a-1)p_0\}$ at a number of independent loci was also developed by the same authors. However, they did not incorporate a parameter for association between genes within individuals (or inbreeding coefficient) in their expression.

The present study is concerned with the development of a comprehensive model giving the required minimum sample size when the sampling is done from a large diploid or autopolyploid ($2k$ -ploid) population with all degrees of inbreeding. We have also tried to isolate the general effects of polysomic inheritance and inbreeding, on sample size for collecting vegetative samples. Our

Keywords. Conservation; germplasm; polysomic; vegetative; inbreeding

trement extends models adopted by Lawrence *et al* (1995a) and Crossa *et al* (1993) for diploid species.

2. Polysomic model

2.1 Diallelic and single locus model

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

$$Z \equiv \left(\begin{array}{cc} A_1^{2k} & A_1^{2k-1} A_2 \\ p_1^{2k} (1 - F_k) + p_1 F_k & {}^{2k}C_1 p_1^{2k-1} p_2 (1 - F_k) \\ A_1^{2k-1} A_2 & A_2^{2k} \\ {}^{2k}C_i p_1^{2k-1} p_2 (1 - F_k) & p_2^{2k} (1 - F_k) + p_2 F_k \end{array} \right)$$

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 $2k$ -ploid organism and is related to the proportion of selfing(s) by the following formula (McConnell and Fyfe 1975):

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

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

$$Z \equiv \left(\begin{array}{ccc} A_{12} & A_1 A_2 & A_2^2 \\ p_1^2 (1 - F_1) + p_1 F_1 & 2p_1 p_2 (1 - F_1) & p_2^2 (1 + F_1) + p_2 F_1 \end{array} \right),$$

where $F_1 = s / (2 - s)$.

Our objective of conservation means that a randomly drawn sample from a $2k$ -ploid population should capture at least one copy of each allele. This can be achieved if the sample contains either one of the heterozygotes or one each of the homozygotes A_1^{2k} and A_2^{2k} . The probability of capturing at least one copy of each allele can be calculated simply by excluding the probability of selecting only one of the homozygotes in a sample of size n , as suggested by Lawrence *et al* (1995a) for diploid models. Thus the probability that a randomly drawn sample of size n contains at least one copy of each alleles at the said locus is

$$P[A_1, A_2] = 1 - \{p_1^{2k} (1 - F_k) + p_1 F_k\}^n - \{p_2^{2k} (1 - F_k) + p_2 F_k\}^n. \tag{1}$$

We can numerically evaluate equation (1) to obtain values of n for a given probability of conservation. For sake of simplicity, if we denote this probability by $(1 - \alpha)$, then

$$\alpha = \{p_1^{2k} (1 - F_k) + p_1 F_k\}^n + \{p_2^{2k} (1 - F_k) + p_2 F_k\}^n.$$

We can further simplify the above expression by assuming that the allele A_2 is rare in nature and occurs with a frequency (p_0) of the order of 0.05 or less. Then the term $\{p_0^{2k} (1 - F_k) + p_0 F_k\}^n$ is almost negligible and can be dropped. Taking the logarithm on both sides of the expression, we get

$$n > \log(\alpha) / \log(\beta), \tag{2}$$

where $\beta = (1 - p_0)^{2k} (1 - F_k) + (1 - p_0) F_k$

From expression (2), we can derive two important results for the extreme conditions of complete outcrossing and complete selfing, as follow:

$$s = 0; \quad n > \log(\alpha) / 2k \log(1 - p_0) \tag{3}$$

$$s = 1; \quad n > \log(\alpha) / \log(1 - p_0). \tag{4}$$

Expressions (3) and (4) indicate that the sample size under complete selfing is almost $2k$ times of that under random mating. As expected, with mixed mating systems, the sample size lies between these two limits and can be obtained numerically by evaluating either (1) or (2).

2. Multiallelic and multilocus model

Let us consider again a $2k$ -auto-polyploid population with a number of alleles at each of λ independent loci and producing a proportion of seed (s) by selfing and $(1 - s)$ by random mating. The number of genotypes for such a population become too large for computation when $k > 2$ and $a > 2$. Therefore we will deduce our results on the basis of methodology and results already published by Crossa *et al* (1993) without a complete algebraic explanation. However, we present an empirical verification of our results in an appendix.

Let us recall the results obtained by Crossa *et al* (1993) by assuming that $(a - 1)$ alleles have identical low frequencies of p_0 and the a th allele has a frequency of $\{1 - (a - 1)p_0\}$ for each of the λ independent loci. Thus, for

$$\text{Alleles (a), loci (1); } n > \log(\alpha) - \log(a - 1) / B \tag{5}$$

$$\text{Alleles(a), loci (\lambda); } n > A / B \tag{6}$$

where $A = \log \{1 - (1 - \alpha)^{1/\lambda}\} - \log(a - 1)$ and $B = \log(1 - p_0)$.

Comparing our result (4) with (5) and (6) obtained by Crossa *et al* (1993) we find that (4) is a special case of (5) and (6) when there are only 2 alleles at a single locus. All these results contain the term $(1 - p_0)$ which is nothing but the probability of capturing a rare allele in a sample of size one when the sampling is done either from an inbred population or a population containing only homozygous lines. However, when one samples from a population which deviates from selfing, a term other than $(1 - p_0)$ must be incorporated to account for this deviation from selfing as well as for diploidy in the expressions (5) and (6). When we replace the term $(1 - p_0)$ in expression (6) by β from expression (2) it can describe both multiallelic as well as multilocus populations. After replacement we get

$$n > A / \log(\beta). \quad (7)$$

Expression (7) evaluates sample sizes directly for a given set of parameters. This can be further rewritten to isolate the effects of deviation from selfing and diploidy present in our model on the sample sizes as follows:

$$n > A / (B + C), \quad (8)$$

where $C = \log\{1 - p_0^{2k-1}(1 - F_k) + F_k\}$.

Thus, in expression (8) we have minimal sample size as $n > A / (B + C)$ whereas, Crossa *et al* (1993) obtained $n > A / B$, which is possible when $C = 0$ or the population in consideration is completely inbred ($s = 1$). The term C involves the polyploidy parameter (k) and the corresponding inbreeding coefficient (F_k); hence, it accounts for a reduction in sample size owing to deviations from selfing and diploidy. For a given ploidy level, reduction in sample size continues until the state of complete random mating, where sample size reaches a minimum value of $A / 2kB$. Thus, for any given population, the minimum sample size lies between $A / 2kB$ and A / B . The upper bound (A / B) is attained under the condition of no random mating ($s = 1$) and is unaffected by the ploidy level. As we deviate from complete inbreeding, the role of ploidy in reducing minimal sample size increases until the minimum value ($A / 2kB$) is reached. This occurs under the condition of no selfing ($s = 0$). Thus, the sample sizes under this condition for diploid, tetraploid, hexaploid and octaploid populations are almost $n_u / 2$, $n_u / 4$, $n_u / 6$, and $n_u / 8$, respectively, where n_u is A / B . With increasing ploidy, the curves displayed in figure 1 almost become parallel to the X-axis except at very high rates of selfing where the minimum sample size rises steeply.

One may debate our justification for replacing the

term $(1 - p_0)$ by β in expression (2) to simplify the derivation of results without giving a mathematical proof. For a diallelic situation, no verification is necessary as the expression (2) itself contains the said term. However for the multiallelic situation, we have verified it empirically by considering a single locus with 3 and 4 alleles for diploids and single locus with 3 alleles for tetraploids (see appendix). We determined sample sizes exactly up to the first decimal place as described in the appendix for 60 cases with varying levels of probability of conservation, rare allele frequencies, and selfing rates. These exact values were then compared with those obtained from our expression (8). To our expectation, in 22 cases, our sample sizes matched exactly to the first decimal place, in 27 cases, the absolute difference was less than or equal to 0.5, in 7 cases, it was 0.6 to 0.9, and, only in 4 cases, we observe absolute differences ranging from 1.1 to 1.6. Of course actual sample size is a whole number. Rounding our calculated values to integers can at the most overestimate or underestimate the sample size by a single plant (tables 1, 2, 3). The close agreement of exact values with those calculated by expression (8) over varied conditions justifies our elimination of the term $\{p_0^{2k}(1 - F_k) + p_0 F_k\}^n$, thus, expression (8) can be safely applied for determining minimum sample sizes with a high degree of accuracy for large populations.

3. Conclusion

In the present paper we attempt to determine a theoretical minimum number of vegetative samples to capture all alleles from a population with a given probability of conservation. We developed a general model by considering a $2k$ -auto polyploid population under a broad range of mating systems. Our work is primarily based on the diploid models suggested by Yonezawa and Ichihashi (1989), Crossa *et al* (1993) and Lawrence *et al* (1995a). Theoretically speaking, the required minimum sample size under our model is $A / (B + C)$ which lies between the bounds $A / 2kB$ and A / B , attained under the extreme conditions of no selfing and no random mating. Crossa *et al* (1993) reported a similar conclusion, but for a diploid model. They indicated that if there are no associations between genes within individuals at any loci, then the required sample size is exactly half the sample size of that under perfect association. If the degree of association is unknown, then the required sample size is between $n / 2$ and n . Our general model yields the same results as given by the said authors when $k = 1$ and $s = 0$ or 1.

Minimum sample sizes for given probabilities of conservation, rare allele frequencies, and numbers of alleles and loci, under our set of assumptions have a minimum value of A / B for all inbred populations irrespective of ploidy level. Sample sizes are smallest

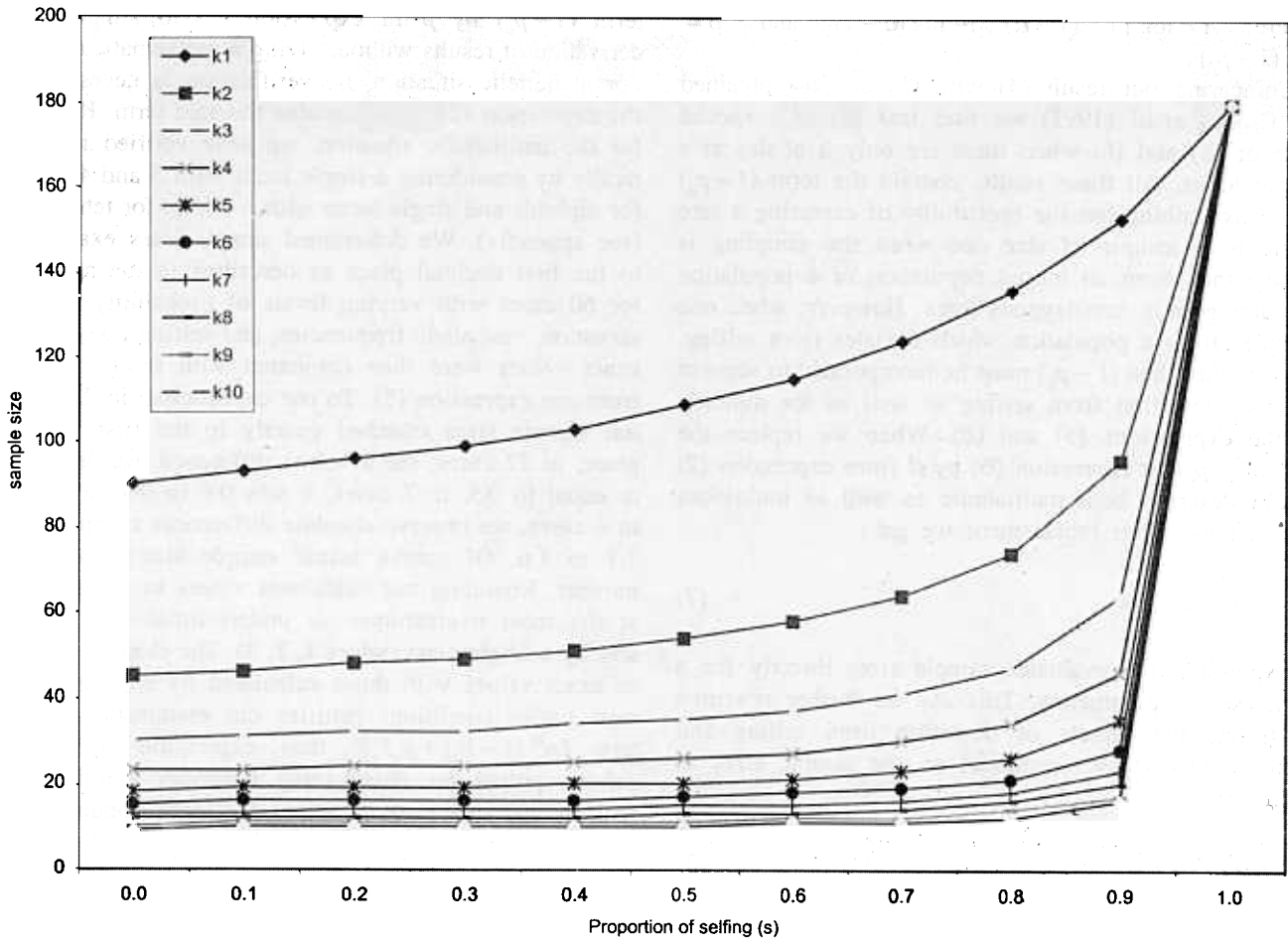


Figure 1. Relationship between sample size and proportion of selfing(s) at various ploidy levels for a population with 2 alleles, $p_0 = 0.05$ and probability of conservation, 99.99%.

Tables 1 and 2. Comparison of results of exact calculations (n_a) with those obtained by expression (8) n_f for various values of $a, s, k, 1 - \alpha$, and p_0 .

Table 1. Diploid with 3 alleles.

a	s	k	$1 - \alpha$	p_0	n_f	n_a	$n_a - n_f$
3	0.0		99.99	0.05	96.5	96.6	
3	0.2		99.99	0.05	102.4	102.4	
3	0.5		99.99	0.05	116.2	116.3	
3	0.8		99.99	0.05	145.4	145.5	
3	1.0		99.99	0.05	193.1	193.1	
3	0.0	1	95.00	0.05	36.0	36.0	
3	0.2	1	95.00	0.05	38.1	38.3	
3	0.5	1	95.00	0.05	43.3	43.5	
3	0.8		95.00	0.05	54.2	54.4	
3	1.0		95.00	0.05	71.9	72.2	
3	0.0		99.99	0.01	492.7	492.7	
3	0.2		99.99	0.01	521.8	521.9	
3	0.5		99.99	0.01	591.6	591.7	
3	0.8		99.99	0.01	739.7	739.7	
3	1.0		99.99	0.01	985.4	985.4	
3	0.0		95.00	0.01	183.5	184.2	
3	0.2		95.00	0.01	194.4	195.0	
3	0.5		95.00	0.01	220.4	221.1	
3	0.8		95.00	0.01	275.5	276.4	
3	1.0		95.00	0.01	367.0	368.3	

Table 2. Diploid with 4 alleles.

a	s	k	$1 - \alpha$	p_0	n_f	n_a	$n_a - n_f$
4	0.0	1	99.99	0.05	100.5	100.5	0.0
4	0.2		99.99	0.05	106.5	106.6	0.1
4	0.5		99.99	0.05	121.0	121.1	0.1
4	0.8		99.99	0.05	151.4	151.4	0.0
4	1.0		99.99	0.05	201.0	201.0	0.0
4	0.0		95.00	0.05	39.9	39.8	-0.1
4	0.2		95.00	0.05	42.3	42.2	-0.1
4	0.5		95.00	0.05	48.1	47.9	-0.2
4	0.8		95.00	0.05	60.1	60.0	-0.1
4	1.0		95.00	0.05	79.8	79.6	-0.2
4	0.0		99.99	0.01	512.9	512.9	0.0
4	0.2		99.99	0.01	543.2	543.2	0.0
4	0.5		99.99	0.01	615.9	615.9	0.0
4	0.8		99.99	0.01	769.9	770.0	0.1
4	1.0		99.99	0.01	1025.7	1025.8	0.1
4	0.0		95.00	0.01	203.7	202.9	-0.8
4	0.2		95.00	0.01	215.7	214.9	-0.8
4	0.5		95.00	0.01	244.6	243.7	-0.9
4	0.8		95.00	0.01	305.8	304.6	-1.2
4	1.0		95.00	0.01	407.4	405.8	-1.6

Table 3. Tetraploid with 3 alleles.

<i>a</i>	<i>s</i>	<i>k</i>	$-\alpha$	p_0	n_f	n_a	$n_a - n_f$
3	0.0	2	99.99	0.05	48.3	48.3	0.0
3	0.2	2	99.99	0.05	50.7	50.7	0.0
3	0.5	2	99.99	0.05	57.4	57.5	0.1
3	0.8	2	99.99	0.05	79.1	79.1	0.0
3	1.0	2	99.99	0.05	193.1	193.1	0.0
3	0.0	2	95.00	0.05	18.0	18.0	0.0
3	0.2	2	95.00	0.05	18.9	18.9	0.0
3	0.5	2	95.00	0.05	21.4	21.4	0.0
3	0.8	2	95.00	0.05	29.4	29.4	0.0
3	1.0	2	95.00	0.05	71.9	71.8	-0.1
3	0.0	2	99.99	0.01	246.3	246.4	0.1
3	0.2	2	99.99	0.01	257.9	257.9	0.0
3	0.5	2	99.99	0.01	290.4	290.5	0.1
3	0.8	2	99.99	0.01	395.9	396.0	0.1
3	1.0	2	99.99	0.01	985.4	985.4	0.0
3	0.0	2	95.00	0.01	91.8	91.5	-0.3
3	0.2	2	95.00	0.01	96.1	95.8	-0.3
3	0.5	2	95.00	0.01	108.2	107.9	-0.3
3	0.8	2	95.00	0.01	147.5	147.0	-0.5
3	1.0	2	95.00	0.01	367.0	365.9	-1.1

under random mating equilibrium. Sample sizes in this state reduce further with increasing ploidy levels. The behaviour of our model confirms the conservative characteristics of genetic variability related to polysomic inheritance as reported by Bray (1983) in the case of tetraploids.

Notably, our treatment only provides a model for the required minimum sample size for collecting the vegetative-materials. As mentioned by Lawrence *et al* (1995a) when seed is collected, sampling is done from the next generation, because the plants raised from this seed are the offspring of the plants from which the collections have been made. Potentially, questions of 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 have been investigated by Yonezawa and Ichihashi (1989) using probability models and Lawrence *et al* (1995b) using the analytical procedures of quantitative genetics for diploid species. This problem in relation to our model needs further investigation.

Appendix

As mentioned above, here we will outline procedures for determining the minimum sample size with given probability of conserving at least a copy of each allele present in the population for three special cases.

(i) Diploid with 3 alleles

Let us denote a diploid population with 3 alleles $A_1, A_2,$ and A_3 as

Genotype	A_1A_1	A_2A_2
	G_1	G_2
Frequency	$p_1^2(1 - F_1) + p_1F_1$	$p_2^2(1 - F_1) + p_2F_1$
Genotype	A_3A_3	A_1A_2
	G_3	G_4
Frequency	$p_3^2(1 - F_1) + p_3F_1$	$2p_1p_2(1 - F_1)$
Genotype	A_2A_3	A_3A_1
	G_5	G_6
Frequency	$2p_2p_3(1 - F_1)$	$2p_3p_1(1 - F_1)$

The expression for evaluating n can be formulated as

$$P[A_1, A_2, A_3] = (1 - \alpha) = 1 - \sum_{i=1}^3 P(A_i)^c + \sum_{1=i < j \leq 3} P(A_i A_j)^c,$$

where $P[A_1, A_2, A_3]$ is the probability of including all the alleles at least once in a sample of size n , $P(A_i)^c$ is the probability of missing A_i , and $P(A_i A_j)^c$ is the probability of missing both A_i and A_j . Putting the values of probabilities in terms of genotypic frequencies in the above expression, we get a simpler expression for evaluating n :

$$\alpha = C_1^n + C_2^n + C_3^n - G_1^n - G_2^n - G_3^n, \tag{a}$$

where

$$\begin{aligned} C_1 &= 1 - G_1 - G_4 - G_6, \\ C_2 &= 1 - G_2 - G_4 - G_5, \\ C_3 &= 1 - G_3 - G_5 - G_6. \end{aligned}$$

(ii) Diploid with 4 alleles

Similarly we can formulate the expression for 4 alleles as;

$$P[A_1, A_2, A_3, A_4] = (1 - \alpha) = 1 - \sum_{i=1}^4 P(A_i)^c$$

$$+ \sum_{1=i < j \leq 4} P(A_i A_j)^c - \sum_{1=i < j < k \leq 4} P(A_i A_j A_k)^c$$

where $P[A_1, A_2, A_3, A_4]$ is the probability of including all alleles (A_1, A_2, A_3, A_4) at least once in a sample of size n , $P(A_i)^c$ is the probability of missing A_i , $P(A_i A_j)^c$ is the probability of missing both A_i and A_j ; and $P(A_i A_j A_k)^c$ is the probability of missing 3 alleles (A_i, A_j, A_k) at a time. After putting the values of probabilities

in terms of genotypic frequencies, we get the final expression which can be numerically solved for n

$$\alpha = C_1^n + C_2^n - C_3^n + C_4^n - C_5^n - C_6^n - C_7^n - C_8^n - C_9^n - C_{10}^n + G_1^n + G_2^n + G_3^n + G_4^n \quad (b)$$

$$\begin{aligned} G_1 &= p_1^2(1 - F_1) + p_1F_1, & G_2 &= p_2^2(1 - F_1) + p_2F_1, \\ G_3 &= p_3^2(1 - F_1) + p_3F_1, & G_4 &= p_4^2(1 - F_1) + p_3F_1, \\ G_5 &= 2p_1p_2(1 - F_1), & G_6 &= 2p_1p_3(1 - F_1), \\ G_7 &= 2p_1p_4(1 - F_1), & G_8 &= 2p_2p_3(1 - F_1), \\ G_9 &= 2p_2p_4(1 - F_1), & G_{10} &= 2p_3p_4(1 - F_1), \\ C_1 &= 1 - G_1 - G_5 - G_6 - G_7, & C_2 &= 1 - G_2 - G_5 - G_8 - G_9, \\ C_3 &= 1 - G_3 - G_6 - G_6 - G_{10}, & C_4 &= 1 - G_4 - G_7 - G_9 - G_{10} \\ C_5 &= G_3 + G_4 + G_{10}, & C_6 &= G_2 + G_4 + G_9, \\ C_7 &= G_2 + G_3 + G_8, & C_8 &= G_1 + G_4 + G_7, \\ C_9 &= G_1 + G_3 + G_5, & C_{10} &= G_1 + G_2 + G_5. \end{aligned}$$

(iii) *Tetraploid with 3 alleles*

We can also formulate the expression for a tetraploid population with 3 alleles $A_1, A_2,$ and $A_3,$ and 15 possible genotypes as

$$\alpha = C_1^n + C_2^n + C_3^n - G_1^n - G_2^n - G_3^n, \quad (c)$$

$$\begin{aligned} G_1 &= p_1^4(1 - F_1) + p_1F_1, & G_2 &= p_2^4(1 - F_1) + p_2F_1, \\ G_3 &= p_3^4(1 - F_1) + p_3F_1, & G_4 &= 4p_1^3p_2(1 - F_1), \\ G_5 &= 4p_1^3p_3(1 - F_1), & G_6 &= 2p_2^3p_1(1 - F_1), \\ G_7 &= 4p_2^3p_3(1 - F_1), & G_8 &= 4p_3^3p_1(1 - F_1), \\ G_9 &= 2p_3^3p_2(1 - F_1), & G_{10} &= 12p_1^2p_2p_3(1 - F_1), \\ G_{11} &= 12p_2^2p_1p_3(1 - F_1), & G_{12} &= 12p_3^2p_1p_3(1 - F_1), \\ G_{13} &= 6p_1^2p_2^2(1 - F_1), & G_{14} &= 6p_2^2p_3^2(1 - F_1), \\ G_{15} &= 6p_1^2p_3^2(1 - F_1), \\ C_1 &= G_2 + G_3 + G_7 + G_9 + G_{14}, \\ C_2 &= G_1 + G_3 + G_5 + G_8 + G_{15}, \\ C_3 &= G_1 + G_2 + G_4 + G_6 + G_{13}. \end{aligned}$$

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 and Edinburgh: Blackwell) pp 97–107
- Bray R A 1983 Strategies for gene maintenance; in *Genetic resources of forage plants* (eds) J G McIvor and R A Bray (Melbourne: CSIRO)
- Bogyo T P, Poreceddu E and Perrino P 1980 Analysis of sampling strategies for collecting genetic material; *Econ. Bot.* **34** 160–174
- Chapman C G D 1984 On the size of genebank and the genetic 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, Hernandez 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 losing an allele when diploid genotypes are sampled; *Biometrics* **36** 643–652
- 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
- Quaiset C O 1975 Sampling germplasm in a center of diversity: an example of disease resistance in Ethiopian Barley; in *Crop genetic resources today and tomorrow* (eds) O H Frankel and J G Hawkes (Cambridge: Cambridge University Press) pp 81–96
- 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 Ichihashi H 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 12 June 1998; accepted 16 September 1998