Theor Appl Genet (1993) 86:673–678

THEORETICAL AND APPLIED GENETICS



# Statistical genetic considerations for maintaining germ plasm collections

J. Crossa[1], C. M. Hernandez[2], P. Bretting[3], S. A. Eberhart[4], S. Taba[1]

[1] International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27, Apdo. Postal 6-641, México D.F., México
[2] University of Colima, Apdo. Postal 22, Colima, México
[3] Plant Introduction Station, USDA, ARS, Ames, IA 50011, USA
[4] National Seed Storage Laboratory, USDA, ARS, Fort Collins, CO 80523, USA

Received: 27 October 1992 / Accepted: 9 December 1992

**Abstract.** One objective of the regeneration of genetic populations is to maintain at least one copy of each allele present in the original population. Genetic diversity within populations depends on the number and frequency of alleles across all loci. The objectives of this study on outbreeding crops are: (1) to use probability models to determine optimal sample sizes for the regeneration for a number of alleles at independent loci; and (2) to examine theoretical considerations in choosing core subsets of a collection. If we assume that $k-1$ alleles occur at an identical low frequency of $p_0$ and that the $k^{th}$ allele occurs at a frequency of $1-[(k-1)p_0]$, for loci with two, three, or four alleles, each with a $p_0$ of 0.05, 89–110 additional individuals are required if at least one allele at each of 10 loci is to be retained with a 90% probability; if 100 loci are involved, 134–155 individuals are required. For two, three, or four alleles, when $p_0$ is 0.03 at each of 10 loci, the sample size required to include at least one of the alleles from each class in each locus is 150–186 individuals; if 100 loci are involved, 75 additional individuals are required. Sample sizes of 160–210 plants are required to capture alleles at frequencies of 0.05 or higher in each of 150 loci, with a 90–95% probability. For rare alleles widespread throughout the collection, most alleles with frequencies of 0.03 and 0.05 per locus will be included in a core subset of 25–100 accessions.

**Key words:** Genetic resources conservation – Sample size – Allele frequency – Probability models – Core subsets

Communicated by A. R. Hallouer
*Correspondence to:* J. Crossa

## Introduction

Managers of genetic resources strive to maintain allelic diversity during regeneration by attempting to retain at least one copy of each allele present in the original population. Crossa (1989) pointed out that the effectiveness of regeneration for maintaining the allelic diversity is related to proper sampling procedures, random genetic drift due to sample size, and optimum seed viability. When sample sizes are large, regeneration is difficult and expensive. Small sample sizes may result in the loss of those alleles present at low frequencies due to random genetic drift.

Genetic diversity within populations depends on the number and frequency of all alleles across all loci, plus the genetic structure of the population. Marshall and Brown (1975) suggested that the most important measure of genetic diversity is the average number of alleles per locus. Weir (1990) defined genetic diversity at a single locus as one minus the sum of squares of allelic frequencies. For outbreeding species, allelic diversity and the proportion of heterozygosity are equivalent. In contrast, self-pollinating species may have much allelic diversity among accessions but few heterozygous individuals within accessions.

The concept of a core collection was introduced by Frankel and Brown (1984) and Brown (1989a, b) with the intent of using the core collection to minimize the cost of germ plasm conservation while insuring maximum genetic diversity. Later, the authors described methods for forming a core subset using information on the origin, and agronomic and morphological characteristics of the accessions. When forming a core subset curators must know: (1) the optimal number of accessions needed to retain an acceptable proportion of alleles present in a given collection, and (2) the

method used to select accessions for the core subset.

The objectives of our study on outbreeding crops were: (1) to use probability models that incorporate the number of alleles at independent loci to determine optimal sample sizes for regenerating germ plasm accessions; and (2) to assess theoretical problems in selecting accessions for the core subset.

## Optimal sample sizes for regenerating germ plasm accessions

Consider a random mating population of infinite size and in Hardy-Weinberg equilibrium that can be subdivided into many highly homozygous lines. Assume that the individuals are diploid and that there are two classes of alleles per locus ($a_1$ and $a_2$); $a_1$ occurs with frequency $p_1$ and $a_2$ with frequency $p_2$ ($p_2 = 1 - p_1$). Sampling, at random, individuals of n different lines is equivalent to drawing, at random, one gamete without replacement after another n times from the original random mating population. This represents n independent, repeated Bernoulli trials.Therefore, the number of lines with a particular allele, in a sample size n, is a random variable with a binomial distribution. The probability of including $x_i$ alleles of class $a_1$ in a random sample of size n is

$$P = \binom{n}{x_i} p_1^{xi} p_2^{n-xi} \tag{1}$$

Then the probability of including in the sample at least one allele of class $a_1$ is

$$P = 1 - p_2^n \tag{2}$$

For k ($k > 2$) classes of alleles, $a_1, a_2, \ldots, a_k$ of a locus with frequencies of $p_1, p_2, \ldots, p_k$, the number of lines in a sample of size n with a certain number of alleles from each allelic class in a sample of size n is a random variable with a multinomial distribution. This case represents independent, repeated trials that generalize from Bernoulli trials with two outcomes to trials with more than two outcomes. Therefore, the probability of obtaining each allele class $x_i$ times in a random sample of size n is

$$P = \left[ n! \prod_{i=1}^{k} (p_i)^{xi} \right] \Big/ \left[ \prod_{i=1}^{k} x_i! \right] \quad (\text{where } \sum_{i=1}^{k} x_i = n) \tag{3}$$

Thus, the probability that each of the k allele classes will be represented at least once in the sample is given by

$$P(a_1 > 0, \ldots, a_k > 0) = 1 - \left\{ \sum_{i=1}^{k} P(a_i) - \sum_{1 \le i < j \le k}^{k} P(a_i a_j) + \sum_{1 \le i < j < z \le k}^{k} P(a_i a_j a_z) - \cdots (-1)^{k+1} \times \sum_{1 \le i < j < z \cdots \le k-1}^{k} P(a_i a_j \ldots a_{k-1}) \right\} \text{ (Crossa 1989),} \tag{4}$$

where $P(a_i)$, the probability that the allele $a_i$ will not appear in the sample, is $(1 - p_i)^n$. Crossa (1989) evaluated this equation for the case of two, three, and four alleles at different frequencies, whereas Marshall and Brown (1975) evaluated it for two and four alleles.

For m independent loci, the probability that each of k allele classes will be detected at least once in each locus in a sample size n is

$$\prod_{l=1}^{m} [P(a_1 > 0, \ldots, a_k > 0)] = \prod_{l=1}^{m} \left[ 1 - \left\{ \sum_{i=1}^{k} P(a_i) - \sum_{1 \le i < j \le k}^{k} P(a_i a_j) + \sum_{1 \le i < j < z \le k}^{k} P(a_i a_j a_z) - \cdots (-1)^{k+1} \times \sum_{1 \le i < j < z \cdots \le k-1}^{k} P(a_i a_j \ldots a_{k-1}) \right\} \right]. \tag{5}$$

Although these equations can be evaluated by numerical procedures for any number of loci and alleles at any frequency, obtaining the required sample size for many alleles at different frequencies in various loci is very impractical. Hernandez and Crossa (1993) developed a computer algorithm to evaluate these equations and thereby facilitate determination of the optimal sample size. The authors specified the assumptions underlying Eq. 5 as: (1) seeds are sampled regardless to the genotype of the parents; (2) there are no associations among genes from different loci (linkage equilibrium); (3) if there are no associations between genes within individuals at any locus, then the required sample size is exactly half the sample size (n) given by Eq. 5; (4) if there is a perfect association between genes within individuals at any locus, then the required sample size equals the sample size (n) given by Eq. 5; and (5) if the degree of association between genes within individuals is unknown, then the required sample size is between n/2 and n.

However, if allele frequencies are unknown, a more specific equation for estimating an optimal sample size that will still retain at least one copy of each of k allelic

classes for a specified probability is required. A simplification of the equation can be obtained by assuming that $k-1$ alleles occur at an identical low frequency of $p_0$ and that the $k^{th}$ allele occurs at a frequency of $1-[(k-1)p_0]$. Then, Eq. 4 can be reduced to the following much simpler expression

$$P(a_1 > 0, \ldots, a_k > 0) = 1 - \left\{ \sum_{r=1}^{k-1} (-1)^{r-1} \binom{k-1}{r} (1 - rp_0)^n \right\}. \quad (6)$$

where r denotes the number of terms in the summation (see Appendix A).

By considering only the first term of Eq. 6 [i.e., $(k-1)(1-p_0)^n$], log transforming, and solving for n, the resulting equation is

$$n > [\log(1-P) - \log(k-1)]/[\log(1-p_0)] \quad (7)$$

It can be shown, for this case, that the other terms of Eq. 6 are negligible (see Appendix B). This expression shows that the sample size required to retain, with probability P, at least one copy of each of the k allelic classes at one locus depends on the number and the frequency of the alleles. A. H. D. Brown in Frankel and Soule (1981) evaluated this formula only for the case of two alleles per locus [i.e., $\log(k-1) = 0$].

For the case of m independent loci and the same number of allelic classes (k) at each locus, Eq. 7 can be written as

$$n > \{\log[1-(P)^{1/m}] - \log(k-1)\}/\log(1-p_0) \quad (8)$$

The formula used by Chapman (1984) resembles Eq. 8, but the former considered only two alleles per locus.

In general, these formulas suggest that the optimal sample size is much more strongly affected by the frequency of the rare alleles than it is by either number of alleles or number of loci. For example, for loci with two, three, or four alleles, each at a frequency of 0.05 ($p_0$), 89, 102, and 110 individuals are required to retain at least one allele in each respective class at each of the 10 loci with 90% probability (Table 1); if 100 loci are involved, 45 additional individuals are required in the sample. For two, three, or four alleles, when the frequency of a particular allele declines to 0.03 ($p_0$) in each of 10 loci, 150, 172, and 186 individuals are required to retain at least one allele in each respective class at each of the 10 loci with 90% probability (Table 1). Similarly, for 100 loci the respective sample size increases by 75 individuals. For classes of loci with two, three, or four alleles, each at a frequency of 0.05 ($p_0$), 134, 148, and 156 individuals are required to retain at least one allele in each respective class at each of the 50 loci, with 95% probability (Table 2). If 150 loci are involved, 22 more individuals are required. To capture, with 90%–95% probability, at least one allele at a 0.03 frequency in each allele class at each of 150 loci, sample sizes of 238–349 individuals are required. For preserving alleles at a 0.01 frequency with 90%–95% probability, the regenerating sample size should include between 722 and 1057 plants.

These results indicate that sample sizes of 160–210 plants are required to maintain alleles which occur at frequencies of 0.05 or higher in each of 150 loci with a 90–95% probability. This sample size will restrict in-

**Table 1.** Sample sizes required to achieve a 90% probability of including at least one copy of alleles with $p_0$ of 0.05, 0.03, and 0.01 from each allele class for several alleles at each locus

| Number of k alleles | Number of loci 1 | 2 | 5 | 10 | 50 | 100 | 150 |
|---|---|---|---|---|---|---|---|
| | | | | $p_0 = 0.05$ | | | |
| 2 | 45 | 58 | 75 | 89 | 120 | 134 | 142 |
| 3 | 58 | 71 | 89 | 102 | 134 | 147 | 155 |
| 4 | 66 | 79 | 97 | 110 | 142 | 155 | 163 |
| 10 | 88 | 101 | 118 | 132 | 163 | 177 | 184 |
| 15 | 96 | 109 | 127 | 140 | 172 | 185 | 193 |
| | | | | $p_0 = 0.03$ | | | |
| 2 | 76 | 97 | 127 | 150 | 202 | 225 | 238 |
| 3 | 98 | 120 | 150 | 172 | 225 | 248 | 261 |
| 4 | 112 | 134 | 163 | 186 | 238 | 261 | 274 |
| 10 | 148 | 170 | 199 | 222 | 274 | 297 | 311 |
| 15 | 162 | 184 | 214 | 236 | 289 | 312 | 325 |
| | | | | $p_0 = 0.01$ | | | |
| 2 | 229 | 295 | 385 | 454 | 613 | 682 | 722 |
| 3 | 298 | 364 | 454 | 523 | 682 | 751 | 791 |
| 4 | 338 | 405 | 494 | 563 | 723 | 791 | 832 |
| 10 | 448 | 514 | 604 | 672 | 832 | 901 | 941 |
| 15 | 492 | 558 | 648 | 716 | 876 | 945 | 985 |

**Table 2.** Sample sizes required to achieve a 95% probability of including at least one copy of alleles with $p_0$ of 0.05, 0.03, and 0.01 from each allele class for several alleles at each locus

| Number of k alleles | Number of loci 1 | 2 | 5 | 10 | 50 | 100 | 150 |
|---|---|---|---|---|---|---|---|
| | | | | $p_0 = 0.05$ | | | |
| 2 | 58 | 72 | 89 | 103 | 134 | 148 | 156 |
| 3 | 72 | 85 | 103 | 116 | 148 | 161 | 169 |
| 4 | 80 | 93 | 111 | 124 | 156 | 169 | 177 |
| 10 | 101 | 115 | 132 | 146 | 177 | 191 | 198 |
| 15 | 110 | 123 | 141 | 154 | 186 | 199 | 207 |
| | | | | $p_0 = 0.03$ | | | |
| 2 | 98 | 121 | 151 | 173 | 226 | 249 | 262 |
| 3 | 121 | 143 | 173 | 196 | 249 | 271 | 285 |
| 4 | 134 | 157 | 187 | 209 | 262 | 285 | 298 |
| 10 | 170 | 193 | 223 | 245 | 298 | 321 | 334 |
| 15 | 185 | 207 | 237 | 260 | 313 | 335 | 349 |
| | | | | $p_0 = 0.01$ | | | |
| 2 | 298 | 366 | 456 | 525 | 685 | 754 | 794 |
| 3 | 367 | 435 | 525 | 594 | 754 | 823 | 863 |
| 4 | 407 | 475 | 565 | 634 | 794 | 863 | 903 |
| 10 | 517 | 584 | 675 | 744 | 903 | 972 | 1013 |
| 15 | 561 | 628 | 719 | 787 | 947 | 1016 | 1057 |

breeding to less than 1% regeneration (for two alleles per locus the rate of inbreeding is given by F = 1/2Ne, where Ne is the effective population size), and thereby avoids inbreeding depression. For most quantitative traits, alleles rarer than 0.05 would probably contribute little to the mean or the variance of the character in the population and so could not easily be measured or evaluated. Therefore, a 0.05 allelic frequency seems to be an appropriate level for calculating requisite sample sizes.

The multinomial models used here allow the generalization of various equations to k rather than to two alleles. Although isozyme analysis in general can resolve two or fewer alleles per locus, modern molecular markers techniques such as restriction fragment length polymorphism (RFLP) and random amplified polymorphic DNA (RAPD) analyses frequently uncover greater numbers of alleles per locus.

The probability of losing one or several alleles at a single locus when diploid individuals are sampled has been considered by Gregorius (1980), who compiled a table of the minimum sample sizes required to assure that all alleles for specified frequencies are included at a certain probability. When the allelic frequency decreases from 0.05 to 0.03, or from 0.02 to 0.01, the required sample size approximately doubles.

A sample size of 90 individuals was suggested by Namkoong (1988) for an average loss of one allele at a frequency of 0.05 at any of 100 loci. Required size increases to 458 individuals when the allelic frequency drops to 0.01.

Except for synthetics derived from inbred lines, the minimum frequency of alleles in the population will be unknown. And even in this case, the number of loci controlling a trait is unknown. However, the formulas derived above serve as general guidelines for the required sample size for accession regeneration.

In a non-ideal breeding population of size N not all of the individuals produce progeny. Only those progenitors that leave offspring influence the genetic constitution of the next generation and therefore contribute to the effective size of the population (Ne). Accordingly, the effective size of the population (Ne) is smaller than the total population size (Ne < N). However, from a practical perspective Ne depends on: (1) the crossing system and (2) the manner that male and female gametes are sampled (Hallauer and Miranda 1981). A practical procedure for regenerating a maize accession should control the male and female gametes. The number of male gametes is controlled through hand pollination (plant-to-plant crosses or chain crosses), whereas the number of females gametes is controlled by taking equal number of seeds from each pollinated ear. In this case the effective population size for the next generation is twice the size of the original population (Ne = 2N) (Crossa 1989).

## Core subsets

The ever increasing number and size of collections stored in germ plasm banks and the complexities of adequately managing and using them have generated considerable concern within the world plant germ plasm community. However, the formation of core subsets appear to offer opportunities for significant improvement in germ plasm management and utilization (Brown 1989b). The core subsets would provide managers of genetic resources and breeders with a manageable number of accessions for their work.

The main purpose for defining core subsets is to ensure that plant germ plasm collections will be used in such a way that they provide efficient access to the entire range of genetic variation. This would facilitate the efficiency of preliminary germ plasm evaluations for needed traits.

Core subsets comprise specific accessions from an existing collection and therefore do not constitute a separate collection per se. As such, they should be fully integrated with the "reserve subset" or non-core subset so that the collection is curated as an essential whole.

Brown (1989a) considered four classes of alleles in a given germ plasm collection: (1) common, localized, (2) rare, localized, (3) common, widespread, and (4) rare, widespread. For the first three classes of alleles, Brown (1989a) studied the distribution of neutral alleles and obtained the expected number of alleles retained in a given size of the core subset. He recommended including 5–10% of the total collection and at least 3000 accessions per species. The last class consists of alleles that are always rare but that occur across most accessions of one collection. For the last two classes of alleles, each accession can be considered to be a random sample from the collection.

### *Determining the optimal number of accessions for a core subset. Assuring that widespread, common, or rare alleles are captured*

We are concerned with the minimum number of accessions required to retain a large proportion of alleles that occur at low frequencies (less than 0.10) in most accessions of the collection (rare and widespread alleles). For this case, the expected number of alleles ($n_A$) in a sample of n accessions can be estimated as follows. Consider A alleles at a locus with frequencies of $p_1$, $p_2$, $p_3, \ldots, p_j, \ldots, p_A$ across all accessions. Let the event $x_j = 1$, if the $j^{th}$ allele is included in a sample of n accessions, and $x_j = 0$ if not. The probability that the $j^{th}$ allele is absent from the sample is $P(x_j = 0) = (1 - p_j)^n$, and the probability of the $j^{th}$ allele occurring in the

sample is

$$P(x_j = 1) = 1 - P(x_j = 0) = 1 - (1 - p_j)^n \quad (9)$$

Then, let

$$S = x_1 + x_2 + \cdots + x_A \quad (10)$$

($S = \sum_{j=1} x_j$, where $j = 1,2,\ldots, A$) be the number of alleles per locus present in the sample of n accessions. The expected value of $x_j$ is given by

$$E(x_j) = \sum_{x_j = 0} x_j P(x_j) = (0)(1 - p_j)^n + (1)[1 - (1 - p_j)^n]$$
$$= 1 - (1 - p_j)^n. \quad (11)$$

Thus, the expected number of alleles ($n_A$) per locus in a sample of n accessions is

$$E(S) = n_A = E\left(\sum_{j=1}^{A} x_j\right) = \sum_{j=1}^{A} E(x_j)$$
$$= \sum_{j=1}^{A} [1 - (1 - p_j)^n] = A - \sum_{j=1}^{A} (1 - p_j)^n \quad (12)$$

The expected number of alleles captured in all loci is the sum of $n_A$ for each locus (Brown 1989a)

Some numerical examples employing Eq. 12 are presented in Table 3. When four of five alleles have frequencies of 0.05 and one has a frequency of 0.80, a subset of 25 accessions would be expected to include four out of five alleles. When two of three alleles have frequencies of 0.05 and one has a frequency of 0.90, a subset of 15 accessions would retain, on the average, two of the three original alleles. For rare alleles widespread throughout the collection and having the range of allelic frequencies considered here, most alleles with frequencies of 0.03 and 0.05 per locus will be included in a subset of 25–100 accessions. These results apply only to alleles that are widespread throughout the collection.

**Table 3.** Number of accessions ($n$) required for core subset such that $n_A$ alleles per locus are retained for loci with three, four, and five alleles at different frequencies

| Allelic frequency | | | | | | |
|---|---|---|---|---|---|---|
| $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $n$ | $n_A$ |
| 0.0001 | 0.0001 | 0.9998 | | | 2000 | 1 |
| 0.01 | 0.01 | 0.98 | | | 70 | 2 |
| 0.01 | 0.01 | 0.98 | | | 300 | 3 |
| 0.03 | 0.03 | 0.94 | | | 90 | 3 |
| 0.05 | 0.05 | 0.90 | | | 15 | 2 |
| 0.0001 | 0.0001 | 0.0001 | 0.9997 | | 4000 | 2 |
| 0.01 | 0.01 | 0.01 | 0.97 | | 100 | 3 |
| 0.01 | 0.01 | 0.01 | 0.97 | | 350 | 4 |
| 0.03 | 0.03 | 0.03 | 0.91 | | 100 | 4 |
| 0.05 | 0.05 | 0.05 | 0.85 | | 25 | 3 |
| 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.9996 | 6000 | 3 |
| 0.01 | 0.01 | 0.01 | 0.01 | 0.96 | 150 | 4 |
| 0.01 | 0.01 | 0.01 | 0.01 | 0.96 | 400 | 5 |
| 0.03 | 0.03 | 0.03 | 0.03 | 0.88 | 100 | 5 |
| 0.05 | 0.05 | 0.05 | 0.05 | 0.80 | 25 | 4 |

In general, a useful strategy for forming core subsets in maize race collections would be to use a stratified sampling strategy. For example, subdividing the total number of accessions into non-overlapping groups based on racial complex and/or ecogeographical criteria. Then, within each racial complex select 25–100 accessions. A subset of this size will preserve, on the average, alleles with frequencies higher than 0.03 in each race collection. If possible, races represented by fewer accessions in the collection should be collected more thoroughly to sample alleles with frequencies higher than 0.03. Within each race complex accessions can be grouped by region or elevation.

Accessions should be placed in multilocational replicated trials, and several morphological and agronomic attributes should be measured. Classification techniques such as cluster analysis and ordination methods such as principal components analysis have proven to be useful for assessing genetic diversity and therefore could help the curator identify similar accessions within racial or ecogeographical subgroups of the core.

*Acknowledgement.* The authors thanks Drs. B. Johnson, K. R. Lamkey, and E. E. Roos for their helpful comments on the manuscript.

## Appendix A

The term $\sum_{i=1}^{k} P(a_i)$ of Eq. 4 includes $k - 1$ sub-terms without the allele $a_k$ and one sub-term with $a_k$. The probability that any of the $k - 1$ alleles will be absent from the sample is $(1 - p_0)^n$, and the probability that the $k^{th}$ allele be absent from the sample is $[(k - 1)p_0]^n$. Thus, the term $\sum_{i=1}^{k} P(a_i)$ can be written as

$$(k - 1)(1 - p_0)^n + [(k - 1)p_0]^n$$

The term $\sum_{1 \le i < j \le k} P(a_i a_j)$ of Eq. 4 includes $\binom{k-1}{1}$ sub-terms that contain the allele $a_k$, each with a probability of $[(k - 2)p_0]^n$, and $\binom{k-1}{2}$ sub-terms that do not include the allele $a_k$, each with a probability of $(1 - 2p_0)^n$. Therefore, the term $\sum_{1 \le i < j \le k} P(a_i a_j)$ can be summarized as follows

$$\binom{k-1}{1}[(k - 2)p_0]^n + \binom{k-1}{2}(1 - 2p_0)^n$$

The term $\sum_{1 \le i < j < z \le k} P(a_i a_j a_z)$ of Eq. 4 comprises $\binom{k-1}{2}$ sub-terms that contain the allele $a_k$, each with a probability $[(k - 3)p_0]^n$, and $\binom{k-1}{3}$ sub-terms that do not include the allele $a_k$, each with a probability of $(1 - 3p_0)^n$. Then, the $\sum_{1 \le i < j < z \le k} P(a_i a_j a_z)$ term is reduced to

$$\binom{k-1}{2}[(k - 3)p_0]^n + \binom{k-1}{3}(1 - 3p_0)^n$$

So, in general, the $r^{th}$ term of Eq. 4 can be expressed as follows

$$\binom{k-1}{r-1}[(k-r)p_0]^n + \binom{k-1}{r}(1-rp_0)^n$$

Therefore, Eq. 4 is reduced to

$$P(a_1 > 0, \ldots, a_k > 0) = 1 - \left\{ \sum_{r=1}^{k-1} (-1)^{r-1} \left[ \binom{k-1}{r-1}[(k-r)p_0]^n + \binom{k-1}{r}(1-rp_0)^n \right] \right\}$$

Since the term $[(k-r)p_0]^n$ is so small that is negligible,

$$P(a_1 > 0, \ldots, a_k > 0) = 1 - \left\{ \sum_{r=1}^{k-1} (-1)^{r-1} \binom{k-1}{r}(1-rp_0)^n \right\}$$

## Appendix B

Substitution of the value of n from Eq. 7 into any of the other summation term of Eq. 6 it can prove that the value of that term is so small as to be negligible. The $r^{th}$ term of the summation can be written as

$$\binom{k-1}{r}(1-rp_0)^a \quad (\text{for } r \geq 2),$$

where $a = n = [\log(1-P) - \log(k-1)]/\log(1-p_0)$ from Eq. 7. That is, $(1-rp_0)^a = \exp[\{(\log(1-P) - \log(k-1))/\log(1-p_0)\}\{\log(1-rp_0)\}]$. Because $\{[\log(1-P) - \log(k-1)]/\log(1-p_0)\}\{\log(1-rp_0)\}]$ is negative, the maximum value of $(1-rp_0)^a$ occurs when the expression $\exp[\{(\log(1-P) - \log(k-1))/\log(1-p_0)\}\{\log(1-rp_0)\}] = 1$. Therefore, minimum values of $p_0$, k, and r that make the quantity $\{[\log(1-P) - \log(k-1)]/\log(1-p_0)\}\{\log(1-rp_0)\}$ approach 0 are required. When $p_0 \to 0$, the limit of $\log(1-rp_0)/\log(1-p_0)$ approaches r. Then

$$\begin{aligned}(1-rp_0)^a &= \exp\{[\log(1-P) - \log(k-1)](r)\} \\ &= \exp[(r)\log\{(1-p)/(k-1)\}] \\ &= [(1-P)/(k-1)]^r\end{aligned}$$

which is minimized when $r = 2$ and $k = 3$. Therefore, the maximum value of $(1-rp_0)^a$ is at $[(1-P)/2]^2$ and for $P = 0.9$ and 0.95, $(1-rp_0)^a = 0.0025$ and 0.000625, respectively. These are the maximum possible values of the second summation term of Eq. 6.

## References


Brown AHD (1989a) The case for core collections. In: Brown AHD, Frankel DH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 136–156

Brown AHD (1989b) Core collections: A practical approach to genetic resources management. Genome 31:818–824

Chapman CGD (1984) On the size of a genebank and the genetic variation it contains. In: Holden JHW, Williams JT (eds) Crop genetic resources: conservation and evaluation. Allen and Unwin, London, pp 102–108

Crossa J (1989) Methodologies for estimating the sample size reequired for genetic conservation of outbreeding crops. Theor Appl Genet 77:153–161

Frankel OH, Brown AHD (1984) Plant genetic resources today: a critical appraisal. In: Holden JHW, Williams JT (eds) Crop genetic resources: conservation and evaluation. Allen and Unwin, London, pp 149–257

Frankel OH, Soule ME (1981) Conservation and evolution. Cambridge University Press, Cambridge

Gregorius HR (1980) The probability of losing an allele when diploid genotypes are sampled. Biometrics 36:643–652

Hallauer AR, Miranda Fo JB (1981) Quantitative genetics in maize breeding. Iowa State University Press, Ames, Iowa

Hernandez CM, Crossa J (1993) A program for estimating the optimum sample size for germplasm conservation. J Hered 84(1):85–86

Marshall DR, Brown AHD (1975) Optimum sampling strategies in genetic conservation. In Frankel OH, Hawkes JD (eds) Crop genetic resources for today and tomorrow. Cambridge University Press, Cambridge, pp 53–80

Namkoong G (1988) Sampling for germplasm collection. HortScience 23:79–81

Weir BS (1990) Sampling properties of gene diversity. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding, and genetic resources. Sinauer Assoc, Sunderland, Mass., pp 23–42