Theor Appl Genet (1989) 77:153–161

THEORETICAL AND APPLIED GENETICS

Review

# Methodologies for estimating the sample size required for genetic conservation of outbreeding crops *

**J. Crossa**

Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT),
Apdo. Postal 6-641, 06600 Mexico D.F., Mexico

Received May 10, 1988; Accepted September 8, 1988
Communicated by A. R. Hallauer

**Summary.** The main purpose of germplasm banks is to preserve the genetic variability existing in crop species. The effectiveness of the regeneration of collections stored in gene banks is affected by factors such as sample size, random genetic drift, and seed viability. The objective of this paper is to review probability models and population genetics theory to determine the choice of sample size used for seed regeneration. A number of conclusions can be drawn from the results. First, the size of the sample depends largely on the frequency of the least common allele or genotype. Genotypes or alleles occurring at frequencies of more than 10% can be preserved with a sample size of 40 individuals. A sample size of 100 individuals will preserve genotypes (alleles) that occur at frequencies of 5%. If the frequency of rare genotypes (alleles) drops below 5%, larger sample sizes are required. A second conclusion is that for two, three, and four alleles per locus the sample size required to include a copy of each allele depends more on the frequency of the rare allele or alleles than on the number. Samples of 300 to 400 are required to preserve alleles that are present at a frequency of 1%. Third, if seed is bulked, the expected number of parents involved in any sample drawn from the bulk will be less than the number of parents included in the bulk. Fourth, to maintain a rate of breeding (F) of 1%, the effective population size ($N_e$) should be at least 150 for three alleles, and 300 for four alleles. Fifth, equalizing the reproductive output of each family to two progeny doubles the effective size of the population. Based on the results presented here, a practical option is considered for regenerating maize seed in a program constrained by limited funds.

**Key words:** Seed regeneration – Sample size – Random genetic drift – Effective population size – Allele frequency

## Introduction

One of the main purposes of germplasm banks is to preserve the existing genetic variability in crop species by conserving as many as possible of the genes that account for this variation. The effectiveness of a bank in performing that task depends on three main factors that are related to each other: (1) sampling procedures, (2) random genetic drift, and (3) seed viability.

The issue dealt with in this paper is the choice of sample size used for regeneration and maintenance of seed. If the sample size is too large, the collection may become difficult and expensive to manage, but if it is too small, valuable alleles may be lost through random extinction due to genetic drift. Since we do not know which alleles will prove to be useful in the future, it is essential that sampling be done efficiently and that populations be of sufficient size to maintain as much genetic diversity as is practical.

The following are some of the issues that a germplasm bank manager must face when regenerating seed: (1) sample size in order to obtain a certain number of desired genes (or genotypes); (2) seed from different individuals is bulked, sample size must include most of the individuals; (3) two or more alleles occur at one locus, sample size must include at least one copy of them; and (4) the effect of finite population size on drift and inbreeding depression.

In this study we examine the use of probability models to answer the first three questions, which deal

* Part of this paper was presented at the Global Maize Germplasm Workshop, CIMMYT, El Batan, Mexico, March 6–12, 1988

with the choice of sample size in regeneration of seed stocks, and then we address the fourth question by employing population genetic theory to consider the genetic consequences of random drift in small populations.

## Probability theory and sampling of parents for regeneration

Drawing at random $n$ seeds or gametes from a population of $N$ seeds (or $2N$ gametes) is an experiment whose possible outcomes are random samples of size $n$. A random sample of size $n$ can be taken in two different ways: with or without replacement. For large populations ($N$) and relatively small samples ($n$), the two ways are approximately equivalent.

### *The number of desired genotypes (genes) in the sample*

The following is a basic problem in the theory of sampling. Let us suppose a finite population of M stored seeds, in which M1 represents the number of desired genotypes (genes) and $\mathrm{M2} = \mathrm{M} - \mathrm{M1}$ are the ones not desired. If $n$ seeds are sampled for regeneration, what is the probability that the sample will include $k$ desired genotypes and $n-k$ undesired genotypes?

The random variable considered here (that is, the number of desired genotypes in the sample) will follow one of two different probability distributions, depending on whether the size of the population is small or large.

*Small population.* When the amount of stored seed is small, the number of desired genotypes ($k$) in the sample follows a hypergeometric distribution, in which case the probability of obtaining at least one desired genotype in the sample is

$$P(k) = 1 - \frac{\binom{\mathrm{M2}}{n}}{\binom{\mathrm{M}}{n}} = 1 - \frac{\mathrm{M2}!\,(\mathrm{M}-n)!}{\mathrm{M}!\,(\mathrm{M2}-n)!}\,.$$

The required sample size depends on how certain one wants to be of including at least one desired genotype. Take the case, for example, of having a 500-seed collection in storage and assume that the desired genotypes occur at a frequency of about 0.05. With a sample size of 90 seeds, one can be 99% sure that at least 1 desired genotype will be included in the sample.

The average number of desired genotypes [(E($k$)] that are expected in a sample of size $n$ in $\mathrm{E}(k) = n(\mathrm{M1}/\mathrm{M})$. For example, if the desired genotypes are rare (1 out of 20), a sample size of 100 will retain, on the average, 5 of them.

*Large population.* When the number of seeds in storage is large, the number of desired genotypes in the sample follows a binomial distribution. We then want to determine the sample size required to give a specific probability of including at least one desired genotypes in the sample. Let $A = \mathrm{M2}/\mathrm{M}$ be the proportion of not desired genotypes and $B = \mathrm{M1}/\mathrm{M}$ be the proportion of desired genotypes, where $A + B = 1$.

A sample of size $n$ can be regarded as $n$ draws of one element. The probability of getting an undesired genotype in a draw is $A = \mathrm{M2}/\mathrm{M}$. The probability of obtaining only undesired genotypes in $n$ draws is $\mathrm{A}^n$. Thus, the probability that at least one desirable genotype will occur in a sample of size $n$ is given by the binomial distribution as follows:

$$P = \sum_{k=1}^{\mathrm{M1}} \binom{n}{k} \mathrm{B}^k \mathrm{A}^{n-k} = 1 - \binom{n}{0} \mathrm{B}^0 \mathrm{A}^{n-0} = 1 - \mathrm{A}^n.$$

Therefore, $n = [\log(1-\mathrm{P})]/[\log(\mathrm{A})]$ (Mainland 1923).

If desired genotypes are rare (occurring at frequencies of 0.03 to 0.05, for example), it is very likely that a regeneration sample consisting of 100 seeds will include at least 1 of these genotypes (Table 1). However, if the proportion of rare genotypes drops to 0.01, the size of the seed sample required to obtain at least one of them increases rapidly (Fig. 1). Even if the sample size is doubled, the probability of retaining at least one desired genotype increases by only 9% (Table 1). However, it is still worthwhile to sample more than 200 seeds if the proportion of desired genotypes is less than 0.03.

Since each individual is a random sample of two gametes, the formulae developed above can be applied to gametes by multiplying the total number of individuals by two.

### *Obtaining R desired genotypes in the sample*

Another way of viewing the sample size required for seed regeneration and maintenance is from the perspective of the negative binomial distribution. According to proba-

**Table 1.** Required sample sizes ($n$) to achieve certain probability of including at least one rare desired genotype ($P$) for a given genotype proportion ($B$)

| Probability $P$ | $n$ | | | | |
|---|---|---|---|---|---|
| | Proportion of rare genotype ($B$) | | | | |
| | 0.01 | 0.03 | 0.05 | 0.07 | 0.11 |
| 0.90 | 229 | 76 | 45 | 32 | 20 |
| 0.93 | 265 | 87 | 52 | 37 | 23 |
| 0.96 | 320 | 106 | 63 | 44 | 28 |
| 0.99 | 458 | 151 | 90 | 63 | 40 |

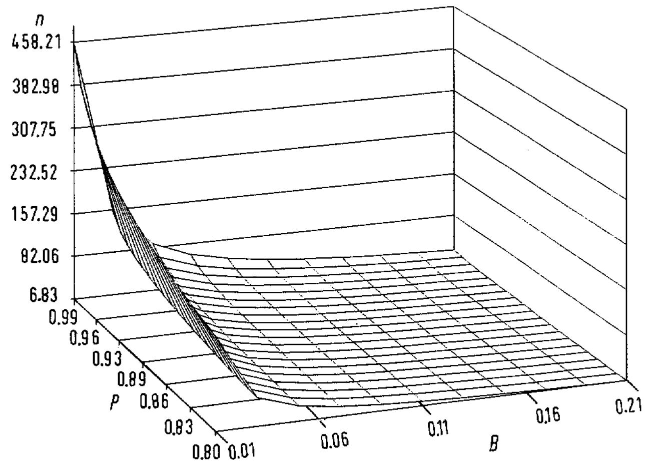

**Fig. 1.** Diagram of the sample size ($n$) required to achieve a probability ($P$) of including at least one desired genotype for various frequencies of the desired genotype ($B$)

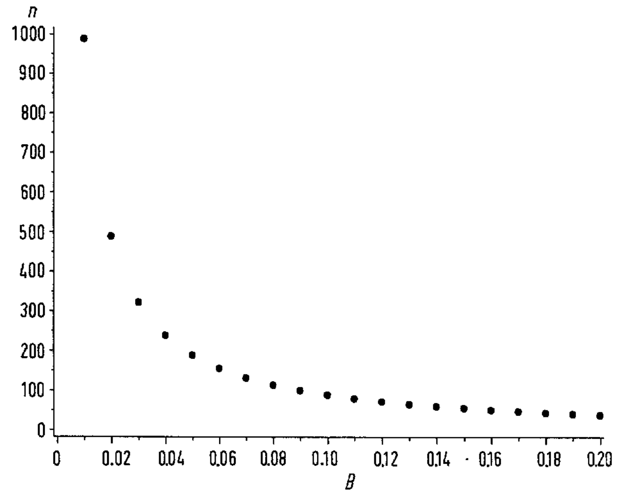

**Fig. 2.** Plot of the expected seed sample size ($n$) required to obtain ten desired genotypes and different frequencies of the desired genotype ($B$)

bility theory, a variable defined as the number of trials required before the first success has a geometric distribution. Furthermore, if individuals in the original collection are sampled until exactly $R$ of them with certain desired characteristics are found, then the number of undesired individuals sampled before obtaining the $R^{\text{th}}$ desired one is a random variable with a negative binomial distribution. In other words, how large a sample does one have to take before obtaining the $R^{\text{th}}$ desired individuals?

For a geometric random variable ($x$), where $B$ is the probability of the desired genotype and $A$ the probability of the undesired one (where $A+B=1$), the ratio $A/B$ is the expected number of undesired genotypes drawn before the first desired individual is obtained. Therefore, the expected sample size needed to obtain $R$ desired individuals is equal to $\mathrm{E}(x)=n=R(A/B)$. For example, if the frequency of the desired genotype is $B=0.01$ and we want to have one ($R=1$) of them in the sample then, on the average, 99 undesired individuals will be drawn before the first desired one is obtained. If the proportion of rare individuals in the collection is larger than 0.05, a sample of 200 or less is needed to obtain, on the average, 10 desired genotypes (Fig. 2). However, if the proportion of rare individuals is lower (0.01 or less), the required $n$ is 1000 or larger for obtaining 10 desired individuals in the sample.

### *Expected number of parents represented in the sample*

A classical problem in probability theory is that of occupancy (Feller 1957). If $G$ balls are distributed among $N$ urns, what is the expected number of occupied urns?

In regenerating maize accessions, one is faced with essentially the same problem when: (1) equal quantities of seed from $N$ ears are bulked and a sample of size $G$ is taken from the bulk, and (2) pollen from plants in one row are bulked to pollinate each female in another row. Some of the parents included in the bulk may not appear in the sample. How many parents can we expect to be represented in the sample?

For $i=1,2,\ldots,N$ parents included in the bulk, let $x_i=1$ if parent $i$ is represented in the sample or $x_i=0$ if parent $i$ is not represented in the sample. Then, let $S=x_1+x_2+\ldots+x_N$ be the number of parents represented in the sample of $G$ seeds. Thus, $\mathrm{E}(S)$ is the number of parents that we expect to be represented in the sample. The probability of parent $i$ being represented in the sample is $P(x_i=1)=1-P(x_i=0)$. Now, $P(x_i=0)=(N-1/N)^G$ where $(N-1/N)^G$ represents the number of ways of taking a sample of $G$ seeds such that parent $i$ does not occur. Therefore, $P(x_i=1)=[1-(N-1/N)^G]=\mathrm{E}(x_i)$ and the expected number of parents represented in a sample is $\mathrm{E}(S)=N-N(N-1/N)^G$. According to that equation, the disadvantage of: (1) bulking equal quantities of seed from each parent and taking a sample from the bulk, and (2) bulking pollen from plants in one row to pollinate female plants in another row is that the number of parents one can expect to include in the sample might be less than the number of parents included in the bulk. Hammond and Gardner (1974) use the above formula to determine the required sample size in maize genetic studies involving varieties or other segregating populations. They also mention the disadvantage of using bulk seed or bulk pollen and suggest that the best procedure is to: (1) use one plant as a pollen parent per female, and (2) sample seed from each female to form a balanced composite.

Suppose that an equal number of seeds is taken from each of 100 maize ears and bulked. If a sample of 100 seeds is taken from the bulk, on the average only 63 parents will be included (Fig. 3). A sample of size 240 will include approximately 90% of the parents. As the number of seeds taken from the bulk grows, the quantity $N(N-1/N)^G$ goes to zero, and $\mathrm{E}(S)$ approaches asymptotically to $N$. For examples, if $N = 100$ and $G = 500$ then $\mathrm{E}(S) = 99$ (Fig. 3).

### *Number of alleles represented in the sample*

Suppose we have a random mating population that can be subdivided into a large number of homozygous lines and in which there are two alleles at one locus, the favorable allele with frequency p and the unfavorable allele with frequency q (where $\mathrm{p}+\mathrm{q}=1$). If a set of $n$ lines (gametes) is drawn at random, the probability of obtaining at least one line (gamete) with the favorable allele is $P=1-\mathrm{q}^n=1-(1-\mathrm{p})^n$, and the required sample is $n=[\log(1-P)]/[\log(1-\mathrm{p})]$.

A. H. D. Brown in Frankel and Soule (1981) used this formula in a study of the following three options for handling redundant entries: (1) delete a random number of entries; (2) discard a random amount of seed from each entry and combine equal quantities of seed into a bulk; and (3) preserve an equal quantity of seed of each entry and maintain each entry separately at its reduced size. If only one entry contains the favorable allele with frequency p and recognition of the plants carrying the desired allele is easy, the last two options give the same sample size required for a given probability that the favorable allele will be preserved. The first option (deletion of entries) is not recommended. The second (bulking of seed) also has a major disadvantage, which was discussed above.

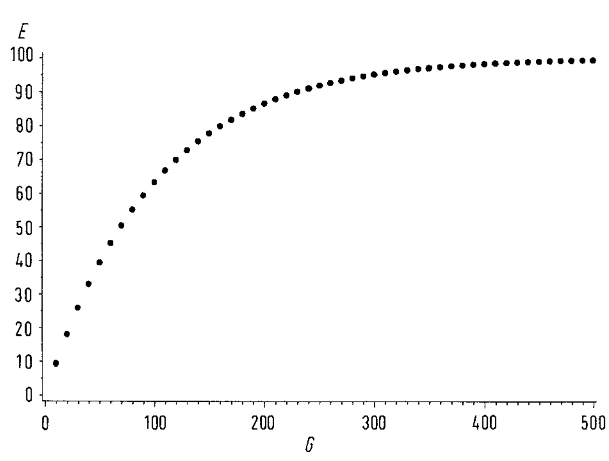

**Fig. 3.** Plot of the expected numbe of parents ($E$) represented in the sample (when equal number of seed is sampled from each of 100 maize ears) and the number of seeds ($G$) taken from the bulk

If there are $x$ independent diallelic loci with the same gene frequencies ($\mathrm{q}=1-\mathrm{p}$) and a set of $n$ lines (gametes) are taken at random, the probability of including the favorable allele at each of the $x$ loci in at least one or more lines is $P=[1-(1-\mathrm{p})^n]^x$. Thus, $(P)^{1/x}=1-(1-\mathrm{p})^n$ and $n=[\log(1-P^{1/x})]/[\log(1-\mathrm{p})]$ (Chapman 1984).

In this equation the value of $n$ depends largely on p. Thus, for one locus we would have to include 458 lines in the sample to be 99% sure of obtaining at least one of them with a rare favorable allele ($\mathrm{p}=0.01$). For 100 loci we would need a collection of 916 lines to be 99% certain of including the desired allele ($\mathrm{p}=0.01$) at each of the 100 loci in at least one line. However, when the allele occurs at a frequency of 5% at all 100 loci, a sample size of only 179 will have a 99% probability of including the desired allele at each of the 100 loci in at least one line.

A more realistic model would allow for the segregation of many alleles at a locus. It is important, therefore, to estimate the sample size of gametes (or lines) required to preserve at least one copy of each allele for the next generation.

Let us first consider two alleles, $B_1$ and $B_2$ with frequencies $\mathrm{p}_1$ and $\mathrm{p}_2$, respectively ($\mathrm{p}_1+\mathrm{p}_2=1$). Two possible outcomes are:

$k_1=\mathrm{B}_1$ is not represented in the sample of $n$ gametes; and $k_2=\mathrm{B}_2$ is not represented in the sample of $n$ gametes.

Thus, the probability that $B_1$ is not represented in the sample of $n$ gametes is $P(k_1)=\mathrm{p}_2^n=(1-\mathrm{p}_1)^n$; likewise, $P(k_2)=\mathrm{p}_1^n=(1-\mathrm{p}_2)^n$ is the probability that the allele $B_2$ will not appear in the sample. The event that at least one copy of each allele occurring in the sample of $n$ gametes is given by $K_1^c \cap k_2^c$, and its probability is

$$\begin{aligned} P(k_1^c \cap k_2^c) &= P[(k_1 \cup k_2)^c] \\ &= 1-\{P(k_1)+P(k_2)-P(k_1 \cap k_2)\} \\ &= 1-(1-\mathrm{p}_1)^n-(1-\mathrm{p}_2)^n. \end{aligned}$$

Note that $P(k_1 \cap k_2)=0$ because $k_1 \cap k_2$ is the null event of neither $B_1$ nor $B_2$ being represented in the sample. It can be shown that for $x$ alleles at a locus, the probability of obtaining at least one of them in the sample is given by

$$\begin{aligned} &P[(k_1 \cup k_2 \cup \ldots k_{x-1} \cup k_x)^c] \\ &\quad = 1-\left\{\sum_{i=1}^{x} P(k_i) - \sum_{1<i<j<x}^{x} P(k_i\, k_j) \right. \\ &\qquad + \sum_{1<i<j<l<x}^{x} P(k_i\, k_j\, k_l) \ldots \\ &\qquad \left. \ldots (-1)^{x+1} \sum_{1<i\ldots<x-1}^{x} P(k_i \ldots k_{x-1}) \right\}. \end{aligned}$$

Where $P(k_i \cap k_j \cap \ldots k_{x-1})=(1-\mathrm{p}_i-\mathrm{p}_j \ldots -\mathrm{p}_{x-1})^n$ is the probability that the $i$th, $j$th, …, and $x$th alleles are not represented in the sample.

**Table 2.** Sample sizes ($n$) required to achieve 95% probability of obtaining at least one copy of each allele for a locus with two, three, and four alleles at various allelic frequencies

| Allelic frequencies | | | | $n$ |
|---|---|---|---|---|
| $p_1$ | $p_2$ | $p_3$ | $p_4$ | |
| Two alleles | | | | |
| 0.99 | 0.01 | | | 297 |
| 0.95 | 0.05 | | | 58 |
| 0.90 | 0.10 | | | 28 |
| Three alleles | | | | |
| 0.98 | 0.01 | 0.01 | | 365 |
| 0.90 | 0.05 | 0.05 | | 72 |
| 0.80 | 0.10 | 0.10 | | 35 |
| 0.50 | 0.49 | 0.01 | | 300 |
| 0.47 | 0.48 | 0.05 | | 58 |
| 0.45 | 0.45 | 0.10 | | 28 |
| Four alleles | | | | |
| 0.33 | 0.33 | 0.33 | 0.01 | 300 |
| 0.49 | 0.49 | 0.01 | 0.01 | 366 |
| 0.97 | 0.01 | 0.01 | 0.01 | 406 |

The last term of that formula was not included because $k_1 \cap \ldots k_x$ refers to the null event of having no allele in the sample, and accordingly has a probability of zero. The above formula was used by Marshall and Brown (1975) to examine the case of four alleles at one locus.

Table 2 gives the sample size required to give a 95% probability of including at least one copy of each allele in cases where two, three, and four alleles are considered at different frequencies. Regardless of the number of alleles per locus and the number of rare alleles, when the frequency of the least common gene or genes is on the order of 0.01, $n$ must be at least 300 to be 95% sure of obtaining at least one copy of each gene in the sample. For one, two, and three rare genes at frequencies of 0.01, we can be sure of including at least one copy of each allele only if $n$ is as large as 300 to 400. But for one or two rare genes that appear at frequencies of about 0.05 at one locus, we can be 95% certain that at least one copy of each will be included in the sample if $n$ is between 60 and 70.

The results indicate, then, that the required sample size is determined by two factors: the frequency and the number of rare alleles operating at a locus. However, it is more dependent on the frequency of the rare allele or alleles than it is on their number.

Considering $x$ alleles at frequencies of $p_1, p_2, \ldots, p_j, \ldots, p_x$ and knowing that $1-(1-p_j)^n$ is the probability that includes all possible events, where the $j$th allele occurs more than once in a sample of size $n$, then the expected number of alleles [E (a)] retained in the sample is given by

$$E(a) = x - \sum_{j=1}^{x} (1-p_j)^n .$$

This formula can be derived by using the same probability procedure utilized to obtain the expected number of parents represented in the sample.

For example, with the sample sizes and cases considered in Table 2, 98% or more of the alleles will be retained. In other words, for two, three, and four alleles, the size of the sample will include, on the average, 1.96, 2.94, and 3.92 alleles, respectively. Brown (1988) has used this formula in determining the required sample size to form a core collection. He described the concept of core collection as being the central and most important part of the collection. For a locus with five alleles, four of which occur at frequencies of $10^{-4}$, a core collection formed with a sample of 3000 individuals would preserve, on the average, 41% of them, that is, 2.04 of the five alleles. Recently Petters (1988) found that number to be a realistic one for forming a subcollection of barley from a large germplasm collection of worldwide origin.

### *Sampling neutral alleles*

The expected number of neutral alleles per locus ($n_a$) maintained in a finite population depends on the effective population size ($N_e$) and the mutation rate (u) to an entirely new allele. Crow and Kimura (1970) used the diffusion approximation to estimate the expected number of neutral alleles in a population. For an allelic frequency lying between p and q ($0<p<q<1$), $n_a$ is as follows

$$n_a = \theta \int_p^q (1-x)^{\theta-1} x^{-1} dx$$

where $\theta = 4 N_e$ u.

Marshall and Brown (1975) and Brown (1988) have calculated values of $n_a$ for various allele frequencies at $\theta = 0.5$, $\theta = 1.0$, and $\theta = 2.0$. They found that large collections contain two to four alleles per locus at frequencies greater than 5% and three to seven alleles at frequencies greater than 1%.

Ewens (1972) translates the mathematical theory of the expected number of neutral alleles per locus in a population to the average number of neutral alleles per locus in the sample. For a population of size $N$ and a sample of $n$ individuals ($2n$ genes), the mean number of alleles in the population *not* observed in the sample ($n_s$) is approximated by the following equation

$$n_s = \theta \log(N/n) - \theta(\theta-1)[(2n)^{-1} - (2N)^{-1}] .$$

Suppose that sample sizes of $n = 50$, $n = 100$, and $n = 200$ ($2n$ genes) are drawn from a collection of $N = 3000$ individuals. Table 3 summarizes the average number of alleles in the collection *not* observed in the sample, assuming different value of $\theta$. The number of alleles in the collection with frequencies greater than 1% are 2.99, 4.61, and 7.2 for $\theta = 0.5$, $\theta = 1.0$, and $\theta = 2.0$, respectively

**Table 3.** Average number of alleles ($n_s$) in the collection not observed in the sample for various $\theta$ and sample sizes ($n$)

| $n$ | $n_s$ | | |
|---|---|---|---|
| | $\theta = 0.5$ | $\theta = 1.0$ | $\theta = 2.0$ |
| 50 | 0.89 | 1.78 | 3.54 |
| 100 | 0.74 | 1.48 | 2.94 |
| 200 | 0.59 | 1.18 | 2.35 |

(Brown 1988). Therefore, if 100 individuals are taken from a collection of 3000 and assuming that $\theta = 0.5$, one can expect that, on the average, about 75% of the neutral alleles occurring at a frequency greater than 1% in the population will be retained in the sample. If the sample size is increased to 200 individuals, only 5% more of the collection's alleles occurring at a frequency greater than 1% will be observed in the sample.

In summary, for preserving and renewing stored germplasm, the greatest possible reduction in loss of favorable alleles through sampling is important. In general, results from the probability models indicate that the recommended size of the sample depends largely on the frequency of the least common allele or genotype. Samples sizes of 100 will preserve rare genotypes (or alleles) that occur with frequencies of about 0.05 with a probability of about 95%. Genotypes appearing at frequencies of more than 10% can be recovered with sample sizes of 40 individuals. However, if the frequency is below 0.05, much larger samples sizes are required to maintain a high probability of including some of the rare individuals in the sample.

Bulking equal quantities of seed from each maize ear and then taking a sample from the bulk is appropriate only for large seed samples. If that procedure is applied to small samples, all seed from some ear included in the bulk will likely be lost in the sampling process.

For two, three, and four alleles at a locus, the sample size required to include at least one copy of each allele with 95% probability depends more on the frequency of the rare genes than on the number of rare alleles. It is expected that sample sizes of 300 to 400 will retain two, three, or four alleles even when some of them occur at a frequency of 1%. For 100 independent loci with the desired allele at a frequency of 5% in each locus, a sample of 179 lines will include the favorable allele at each of 100 loci in at least one line with 99% probability.

## The effect of random genetic drift in seed regeneration

The ultimate aim of seed conservation is to preserve the genetic variability still existing in crop species. Since that variability is a function of the alleles number and frequency, knowledge of the effect of changes in population size on these parameters is needed.

For large populations and constant environmental conditions, gene frequency will not change from generation to generation. However, in small samples, random fluctuation in allele frequency can occur, alleles can become fixed, and favourable alleles may be lost as a result of random sampling of gametes. This phenomenon is known in population genetics as random genetic drift. Changes in allele frequency caused by sampling are unpredictable and depend on the size of the sample. However, it is possible to calculate the variance of such random deviation of gene frequency and the probability that a desired gene will be lost.

If $2N$ gametes are drawn from a large pool of gametes, the probability that an allele with frequency p will occur in $i$ gametes of the sample is given by the binomial distribution. The variance of the number of occurrences of the allele in $2N$ trials, expressed as the variance of the proportion of occurrences, is as follows

$$V(p) = pq/2N.$$

The variance of p (or its standard deviation) gives a measure of the magnitude of the random fluctuation of p in a finite population. Using the normal distribution to approximate binomial distribution, calculation of the probability of the various values of p for the next generation after random sampling of gametes is possible. For example, consider a large parental population for which $p = 0.05$. The probability distribution of p in the next generation of $N = 100$ offspring indicates that 2.1 % of the loci will have $0.05 + 2(0.015) < p < 0.05 + 3(0.015)$ and 2.1% of them will have $0.05 - 2(0.015) < p < 0.05 - 3(0.015)$, where 0.015 is the standard deviation of p.

### *Inbreeding and seed regeneration in small populations*

The consequence of unpredictable changes in gene frequency caused by random genetic drift in a population are well known. In a small population, this leads to continuous fixation and loss of alleles which, in turn, causes a decrease in the proportion of heterozygous individuals in the population. For a population with two alleles at a locus, the reduction in heterozygosity (or F = rate of inbreeding) per generation of random genetic drift is $1/2\,Ne$ (Wright 1931), where $Ne$ represents the effective size of the breeding population. This means that in each generation an average of $1/2\,Ne$ of the previously unfixed loci become fixed or lost; that is, the rate of change from a population with two alleles to a population with one allele is $1/2\,Ne$ per generation of sampling. In general, the rate of change from $x$ alleles to $x-1$ alleles is $F = x(x-1)/4Ne$ (Kimura 1955). This formula is relevant to problems of regenera-

**Table 4.** Effective population sizes (*Ne*) required to maintain F = 1% for various numbers of alleles per locus (x)

| No. of alleles *x* | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| *Ne* | 50 | 150 | 300 | 500 | 750 |

tion and maintenance because the effective population size is related to the rate of inbreeding by $Ne = x(x-1)/4\,F$.

Frankel and Soulé (1981) suggest that natural selection for performance and fertility can balance inbreeding depression only if F = 1%. They called this the basic rule of conservation genetic.

With two allele per locus, an effective population size of 50 will ensure a minimum loss in heterozygosity of 1% (Table 4) and a 95% probability that at least one copy of each allele will be recovered if one of them is rare ($p_2 = 0.05$) (Table 2). In the case of three alleles, an effective sample size of 150 is required to maintain a rate of inbreeding of 1%. This is more than two times the sample size required to get at least one copy of each allele when one or two alleles occur at frequencies of 0.05 (Table 2). For the case of four alleles, an effective sample size of 300 will ensure that (1) the rate of inbreeding will be maintained at 1%, and (2) 99% of the alleles, on the average, will be retained in the sample (see expected number of alleles in the sample). We can conclude that the estimated sample sizes discussed previously will ensure a minimum loss in heterozygosity of 1%.

### *Effective population size and seed regeneration*

Not all of the individuals in a population of size *N* will produce offspring that survive to maturity in the next generation. Consequently, the effective number of progenitors (*Ne*) producing offspring that will constitute the next generation may be much lower than the total number of individuals (*N*) in the total population. The degree of sampling error or genetic drift of an ideal breeding population of size *Ne* is the same as that affecting a particular real population of *N* individuals with different numbers of males and females. Departures from the definition of an ideal population, such as variation in the sex ratio of breeding individuals, in the number of breeding individuals at different times, in the number of offsprings per family, and in the type of reproduction influence the genetic contribution to the next generation and therefore the effective size of a population. In general, all these factors reduce the actual size of the breeding population ($Ne < N$).

For example, in the multiplication of maize accessions, suppose that 3000 individuals are planted in isolation and that the pollen from most of the plants in the field is bulked and used for pollination. If only 100 plants are harvested, *Ne* is only 387 and the rate of decrease in heterozygosity will be F = 0.13%. Because each sex provides half of the alleles, a sex ratio that is not 1:1 will increase the likelihood of genetic drift. In general, for a population with a limited number of females (*Nf*), *Ne* is smaller than four times *Nf*.

One of the aims of any seed regeneration procedure should be to use the same number of breeding individuals through the different cycles of regeneration. Because the effective population size is the harmonic mean of the population sizes in the different cycles of regeneration, changing *Ne* leads to a population size which is closest to the minimum number.

Controlling the number of progeny per family may produce a much larger *Ne* than the real size of the population. This case is of much interest in the rejuvenation of seed stored in germplasm banks.

In an ideal population of size *N*, with $Nm = N/2$ males and $Nf = N/2$ females mating at random, the variance of the random deviation of gene frequency is given by $pq/2N$. The number of progeny ($ki$), randomly distributed among families, follows a Poisson distribution. As a consequence, the mean number of progeny ($\bar{k}$) and the variance in the number of progeny [$V(k)$] are equal.

Let *N* be the actual number of parents, $ki$ the number of gametes (offspring) contributed by the $i$th parent, $\bar{k} = \sum ki/N$ the mean, and $V(k)$ the variance of the number of gametes (offspring) left by a parent. Then,

$$Ne = \{N\bar{k}(N\bar{k}-1)\}/\{N-1)\,V(k) + N\bar{k}(\bar{k}-1)\}\,.$$

The derivation of this formula is given in C. C. Li (1976).

If every parent contributes the same number of gametes (offspring) to the next generation, then $ki = \bar{k}$, $V(k) = 0$, and $Ne = (N\bar{k}-1)/(\bar{k}-1)$. When two progeny from each individual are allowed to reproduce, such that $ki = \bar{k} = 2$ and $V(k) = 0$, then $Ne = 2N-1$, which is almost twice as large as the real population. This result can be used in seed conservation. By equalizing the reproductive output among families to two progeny, $V(k)$ is reduced to zero, and the effective size of the population is nearly double what it would be with random mating. Furthermore, special mating systems that minimize random genetic drift, such as maximum avoidance of inbreeding and circular half-sib, have no advantage over the equalization of progeny number among families (Frankel and Soule 1981). Any procedure that makes *Ne* as large as possible will be useful in preserving genotypes (alleles) that appear at frequencies of 0.01 or less.

### *The effect of finite population on two linked loci*

Conservationists may be interested in maintaining not only individual genes, but also groups of linked genes and in breaking up linkage groups when desired and undesired alleles are in close association.

Although a theory comparable to the one locus model has not been developed, genetic drift may cause nonrandom associations between linked loci. The linkage disequilibrium between two loci caused by random sampling of gametes in a finite population has been studied by Hill and Robertson (1968) and Ohta and Kimura (1969).

The squared correlation coefficient $r^2$ between gene frequencies at two segregating loci is used to show the effect of finite population size on gametic disequilibrium. For an effective population size of *Ne* and an amount of recombination "*c*" between two loci, the expected value of $r^2$ is

$$\mathrm{E}(r^2) = 1/(1 + 4\,Ne\,c)\,.$$

For small values of *Ne c*, the gametic disequilibrium approaches the maximum value of 1.0. This expression can be used to study the effect of different population sizes on the measure of gametic disequilibrium ($r^2$) for a given recombination factor (*c*). A substantial amount of gametic disequilibrium (between 10%–20%) is obtained for effective population sizes of 200 to 100, respectively (Table 5). If desired linked groups are to be maintained, it is expected that with an effective propulation size of 50–100, 30%–20% of the gametes will contain the original blocks of genes. However, if recombination of linked genes is desired, then population sizes of more than 250 individuals would be necessary.

## A practical system for maize seed regeneration

An ideal system for maize seed regeneration should (1) allow for equalizing the genetic contributions of the parents, and (2) avoid small population sizes at any regeneration cycle. In practice, however, resources are limited and ideal procedures may not be practical. A practical system for maize seed regeneration is described as follows.

Suppose that we have a collection of 100 maize ears. Take at random one kernel from each ear and put it in a packet; repeat until two packets are completed. Then, plant a block from each seed packet (a single seed per hill); each block will have 100 plants (one plant from each original maize ear). Make 50 random plant to plant crosses within a block, using each plant only once as either male or female, but not both. Finally, the number of progeny per original family is equalized to two by harvesting the two blocks. From the 100 ears collected (50 from each block), take at random one kernel and put it in a packet; repeat until 2 packets are completed. These will be used for the next cycle of regeneration. With this procedure *Ne* = 200. This is similar to the biparental mating scheme recommended by Gale and Lawrence (1984).

**Table 5.** Percentage gametic disequilibrium ($r^2$%) for various effective population sizes (*Ne*) *

| *Ne* | 50 | 100 | 150 | 200 | 250 | 300 |
|---|---|---|---|---|---|---|
| $r^2$% | 33 | 20 | 14 | 11 | 9 | 7 |

* Assume $c = 0.01$

Optimum conditions for growing and drying the seed as well as optimum long-term seed-storage facilities will ensure the maintenance of high seed viability and genetic integrity of every accession. To monitor the seed viability of collections stored in a germplasm bank is an important procedure for deciding whether or not to regenerate an accession (Ellis and Roberts 1984). Furthermore, the viability test will serve to adjust the number of seeds put into a packet, that is, a decline in seed germination will have to be compensated for by planting more seeds in the field.

*Acknowledgements.* The author thanks Drs. C. O. Gardner, S. A. Eberhart, M. Goodman, G. Wilkes, S. Taba, D. Eaton, G. Han, D. C. Jewell, and the anonymous reviewers for their helpful comments on the manuscript.

## References

Brown AHD (1988) The case for core collection. In: Brown AHD, Frankel DH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge (in press)

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 102–118

Crow JF, Kimura M (1970) An introduction to population genetic theory. Burgess Publishing, Minnesota

Ellis RH, Roberts EH (1984) Procedures for monitoring the viability of accessions during storage. In: Holden JHW, Williams JT (eds) Crop genetic resources: conservation and evaluation. Allen and Unwin, London, pp 63–75

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Bio 3:87–112

Feller W (1957) An introduction to probability theory and its application, 2nd ed, vol 1. Wiley, New York

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

Gale JS, Lawrence MJ (1984) The decay of variability. In: Holden JHW, Williams JT (eds) Crop genetic resources: conservation and evaluation. Allen and Unwin, London, pp 77–100

Hammond JJ, Gardner CO (1974) Effect of genetic sampling technique on variation within populations derived by crossing, selfing or random-mating other population. Crop Sci 14:63–66

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Kimura M (1955) Random genetic drift in a multi-allelic locus. Evolution 9:429–435

Li CC (1976) First course in population genetics, 2nd ed. The Boxwood Press, California, p 561

Mainland GB (1923) Muller's method of calculating population sizes. J Heredity 42:237–240

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

Ohta T, Kimura M (1960) Linkage disequilibrium due to random genetic drift. Genet Res 13:47–55

Petters JP (1988) The emergence of new centres of diversity: evidence from barley. Theor Appl Genet 76:17–24

Wright S (1931) Evolution in mendelian populations. Genetics 16:97–159