Chapter 5: Basic sampling strategies: theory and practice

Jose Crossa
Biometrics and Statistics Unit
International Maize and Wheat Improvement Center (CIMMYT)
Apdo. Postal 6-641, 06600 Mexico, DF, Mexico
E-mail: j.crossa@cgiar.org

Roland Vencovsky
Dept. of Genetics
ESALQ/Universidade de Sao Paulo
Cx P. 83,13400-970, Piracicaba, Sao Paulo, Brazil
E-mail: rvencovs@esalq.usp.br

Abstract

Programmes for conserving genetic resources have collected, received and stored hundreds of thousands of accessions of different cultivated species and their wild and weedy relatives. Collection and regeneration protocols must consider the species (i.e., allogamous, partially allogamous, autogamous and dioecious) to ensure that the sample is representative of the population. Previous studies have used allelic richness as the basic parameter for determining sample sizes for genetic resource conservation. The concept of variance effective population size is important to the measurement of genetic representativeness and has been successfully used in genetic conservation (regeneration and collection). The aim of this chapter is to show how to practically apply the theory developed earlier and to demonstrate its use for answering practical questions that a manager of genetic resource conservation might pose when collecting and regenerating plant genetic resources. This chapter explains strategies for determining efficient sample size in order to maintain the representativeness of the original diversity when collecting and regenerating genetic resources.

Introduction

The great genetic complexity of most plant populations and the many possible ways that genetic resources may be used in the future makes it difficult to provide simple and efficient sampling schemes and optimal sample sizes for the maintenance of all species (Namkoong 1986). Programmes for the conservation of genetic resources and genebanks around the world have collected (in their centres of origin or elsewhere), received and stored hundreds of thousands of accessions of different cultivated species and their wild and weedy relatives. These accessions represent a wide spectrum of population diversity. Collection and regeneration protocols must consider the species (including allogamous, partially allogamous, autogamous and dioecious species, plus the type of materials being collected and regenerated) to ensure that the sample is representative of the population.

This chapter is a synthesis of new knowledge, procedures, best practices and references for collecting plant diversity since the publication of the 1995 volume *Collecting Plant Diversity: Technical Guidelines*, edited by Luigi Guarino, V. Ramanatha Rao and Robert Reid, and published by CAB International on behalf of the International Plant Genetic Resources Institute (IPGRI) (now Bioversity International), the Food and Agriculture Organization of the United Nations (FAO), the World Conservation Union (IUCN) and the United Nations Environment Programme (UNEP). The original text for Chapter 5: A Basic Sampling Strategy: Theory and Practice, authored by A. H. D. Brown and D. R. Marshall, has been made available online courtesy of CABI. The 2011 update of the Technical Guidelines, edited by L. Guarino, V. Ramanatha Rao and E. Goldberg, has been made available courtesy of Bioversity International.

Previous studies have developed equations for collection and regeneration to be applied when using practical breeding schemes or methods for conserving genetic resources, mainly of monoecious species. The studies and research previously done on conservation of genetic resources concentrated on developing sample sizes based on maximizing allelic richness or the number of distinct alleles at a single locus (Marshall and Brown 1975).

Another important measurement of representativeness is related to how stochastic changes in allelic frequency – caused by sampling error in small populations (random genetic drift) – lead to continuous fixation and loss of alleles, and reduce the proportion of heterozygous individuals in the population (Crow and Kimura 1970; Wright 1931). These random changes in allelic frequency affect the genetic representativeness of the finite population. In a large mating population of N individuals, the reduced number of progenitors whose offspring will constitute the next generation causes random genetic drift, which is quantified and predicted using the parameter called "effective population size" (N_e). The effective population size is the size of an ideal population that has the same amount of drift in allelic frequency or the same rate of decrease in heterozygosity as the actual population. Several factors affect the effective size of a population, including the number of parents per generation, the number of corresponding offspring, the number of male and female gametes contributed per individual in the parental population, the mating system of the species, and so on. The effective population size, taken as a measure of the genetic representativeness of a seed sample, can be adapted to specific aspects of plant breeding and conservation of genetic resources, such as seed regeneration and collection.

During the process of collecting, storing, regenerating and restocking germplasm, genetic drift occurs and affects a population's genetic integrity in a number of ways. First, when collecting germplasm in the field, an appropriate sampling strategy should be used to avoid, as much as possible, dramatically reducing the size of the population (bottleneck). Second, when accessions are stored, the different survival rates of genotypes and the accumulation of mutations affect the genetic integrity of the accessions. Third, when seed of an accession is drawn for regeneration, the sample needs to be optimum in order to avoid genetic drift due to sampling or to differential survival (germination) or fecundity, leading to changes in the genetic constitution of the accession. Fourth, when plants of accessions are being regenerated in the field, insects, diseases, and other environmental factors can affect the plant stand, thereby limiting the accessions' gametic contribution (offspring) to the next generation. Since we do not know which alleles will prove 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. Large samples are expensive and difficult to manage, but if samples are too small, genetic diversity may be affected by the loss of valuable alleles through random changes in allelic frequency. Maintaining allelic diversity during regeneration depends on factors such as sampling procedures, seed viability and mating systems, all of which cause random genetic drift.

Concerning the sample size for collecting germplasm, Allard (1970) pointed out that most plant species contain remarkable stores of genetic variation and consist of millions of different genotypes. Plant collectors can hope to sample only a fraction of the variation that occurs in nature. It is important that this fraction be as large as possible and contain the maximum amount of useful (now and in the future) variation. Allard (1970) also recognized that collectors as well as end-users of germplasm have limited time and resources at their disposal. Thus, the problem is to define a sampling procedure that yields the maximum amount of useful genetic variation, within a specified and limited number of samples (Marshall and Brown 1975).

The aim of this chapter is to show how to apply the theory and equations developed earlier and to demonstrate their use for answering the practical questions that a manager of genetic resource conservation might pose when collecting and regenerating plant genetic resources. It includes theoretical considerations for probability models that compute the required sample size for conserving alleles at loci and for developing strategies for efficient sample size to maintain the representativeness of the original diversity when collecting and regenerating genetic resources. It also provides theoretical concepts on the variance effective population size ($N_{e(v)}$), and describes the derivation of manageable equations for collection and regeneration of germplasm. We give practical examples of how to (1) compute the required sample size for maintaining, with a certain probability, at least one copy of each allele at independent loci in the accession to be regenerated, (2) assess practical procedures for the collection and regeneration of genetic resources

that will increase the $N_{e(v)}$ of random-mating populations and the $N_{e(v)}$ of species with mixed self- and random-mating systems of reproduction and different natural rates of self-fertilization, and (3) examine practical collection and regeneration procedures that will increase the $N_{e(v)}$ of dioecious species with different proportions of male and female plants.

Theory

Probability models for number of alleles in the sample

We are interested in finding a sample size n_g such that the probability of detecting at least one allele of each allele class is greater than a quantity 1- α or $P(n)>1-\alpha$ (for small α). Except in the case of synthetics derived from inbred lines, in real situations, allele frequencies are unknown; therefore, some simplification is necessary. Assuming that k-1 alleles occur at an identical low frequency of p_0 and that the kth allele occurs at frequency of 1-[(k-1)p_0], Crossa et al. (1993) showed that the sample size (n_g gametes) required to retain with a probability $P=1-\alpha$ that at least one copy of each of the k allelic classes in each of the m loci should be larger than

$$n_g > \frac{\log\left[1-(P)^{\frac{1}{m}}\right]-\log(k-1)}{\log(1-p_0)} \quad (1)$$

where it is assumed that loci are independent and that diploid individuals (not gametes) are sampled. The assumption of loci independence is unrealistic, since genetic drift may cause non-random associations between linked loci. If the population is under random mating (a cross-pollinated species), linkage equilibrium can be assumed for all possible pairs of loci, so it is expected that, for each pair of loci, a similar number of coupling and repulsion combinations occur. This obviously does not hold for self-pollinated species or species with mixed self- and random-mating systems of reproduction. Therefore equation 1 only offers approximate guidelines for the range of sample sizes, which are greatly dependent on the frequency of rare alleles. Furthermore, sampling $n_g= 2N$ gametes is equivalent to sampling (½)$n_g= N$ individuals (diploid zygotes) only for panmictic populations (i.e., idealized random-mating populations of infinite size with no association between any two genes within individuals). In this case, the number of diploid individuals that need to be sampled is exactly half the number of gametes (n_g).

It should be pointed out that the range of sample sizes given by equation 1 is based solely on probability models and does not consider the genetic structure of the population or specify how well a particular sample represents the reference population in terms of genetic parameters such as variance of allelic frequency, inbreeding, random genetic drift due to sampling error, genetic linkage, etc. Furthermore, probability models do not, by themselves, specify appropriate mating and reproduction systems such as panmixia, mixed self- and random mating, self-fertilization, and so forth, which could, under specific circumstances, maximize the genetic representativeness of the sample.

Therefore, it is clear that for an initial census of a population of size N, it is very likely that $N_e < N$. For studying factors that will make $N_e > N$, aspects of population genetics theory must be incorporated into the probability models in order to facilitate the estimation of parameters such as the mean and variance of male and female contributed gametes and the mating systems that would facilitate control of male and female gamete contributions and thus allowing to increase N_e.

Measure of representativeness: variance effective population size [$N_{e(v)}$]

The concept of effective population size $N_{e(v)}$ is useful when studying the breeding structure and genetic representativeness of an actual population, as related to an ideal population in which individuals mate at random with no variation in fertility: the number of progeny per parent has a binomial distribution (which approaches Poisson for large sample sizes) and its size is kept constant through time (Wright 1931). The effective population size determines the amount of sampling error between generations that causes random fluctuations in allelic frequency. On the other hand, natural and breeding populations fluctuate in size from

one generation to another, and some individuals might not produce gametes or might have different natural inbreeding levels and different natural self-fertilization rates than others. Thus, measures of effective size often deviate from the size of the actual population.

Crow and Kimura (1970) derived models for studying the genetics of finite populations and assessed the relationship between the actual size of a population, *N*, and the distribution of progeny among parents by estimating its mean and variance. They derived an expression for the variance effective population size ($N_{e(v)}$) when a random sample of individuals is taken from an original population of size *N*, assuming that all the individuals potentially contribute male and female gametes,

$$N_{e(v)} = \frac{2n}{\frac{S_k^2}{\bar{k}}(1+\alpha)+(1-\alpha)} \tag{2}$$

where α measures the departure from Hardy Weinberg equilibrium in generation $t-1$, $\alpha = -\frac{1}{2N_{t-1}-1}$; n is the number of offspring in generation t (or total number of seeds collected from the parents); and k is the number of gametes contributed by a particular parent. This can be considered a binomial random variable in the case of a random sampling of gametes with no variation in fertility, where $\bar{k} = \frac{2n}{N}$ and $S_k^2 = \frac{N}{(N-1)}V_k$ are the mean and variance of the gametes contributed by the parents, respectively.

For large *N*, $S_k^2 = \frac{N}{(N-1)}V_k \approx V_k \approx \bar{k}$. Crow and Denniston (1988) provided expressions of the inbreeding and the variance effective number that are general for populations under random mating, and presented formulas as functions of means, variances and covariances of the number of gametes contributed by parents to the offspring.

Genetic models based on the number of male and female gametes contributed by individuals of monoecious plant species have been developed for variance effective population size applied in the context of artificial selection (Vencovsky 1978) and to specific aspects of genetic resource conservation (i.e., collection and regeneration) (Crossa and Vencovsky 1994; Vencovsky 1978). Later, Crossa and Vencovsky (1997, 1999) and Vencovsky and Crossa (1999a, 1999b, 2003) showed the theoretical developments and practical applications of variance effective population size when drift occurs at two stages: when sampling parents for reproduction and when sampling gametes (offspring) from those parents for monoecious populations and for populations under mixed self- and random mating.

Two-stage sampling model for random-mating species

Crow and Kimura (1970) assumed that all *N* individuals in the parental generation potentially contribute gametes to the next generation. However, in real-life situations of plant breeding, regeneration and collection of plant genetic resources, some plants may fail to produce gametes (offspring) due to external factors such as poor seed viability or germination, insects and diseases in crossing blocks or the systematic exclusion of a fraction of parents, etc. Therefore, a more realistic model would consider that the overall effect of sampling on allelic frequency drift is the summation of drift occurring at two stages: (1) when sampling parents from the original population (first stage) and (2) when sampling gametes from these parents (second stage).

Vencovsky (1978) first proposed models for computing $V_{(k)}$ and $N_{e(v)}$ for the two-stage sampling model for monoecious species that considers that some plants might not contribute gametes to the next generation. Crossa and Vencovsky (1994) provided approximate equations for computing $V_{(k)}$ and $N_{e(v)}$ for germplasm collection and regeneration with and without control of contributed male and female gametes. Crossa and Vencovsky (1997) provided the theoretical development of the two-stage model, gave an alternative derivation based solely on the theory of random sampling within a finite population, and demonstrated practical procedures for seed collection and regeneration.

Consider an initial set of N diploid monoecious plants (reference population), where P plants contributing male and female gametes are randomly selected ($0 < P \leq N$). From the remaining N-P plants, R contributing only male gametes are also sampled at random [$0 \leq R \leq (N\text{-}P)$]. Therefore, $M = P+R$ plants contribute male gametes, P plants contribute female gametes, and N-M plants do not contribute any gametes at all. Thus, the proportion of functional seed parents is $u=P/N$ (or $P = uN$) (where $0< u \leq 1$) and the proportion of functional pollen parents is $v= (P+R)/N = M/N$ (or $M = vN$) ($0 < v \leq 1$). With these proportions, it is possible to simulate several important practical situations that arise when regenerating and collecting genetic resources.

For example, the condition given by Crow and Kimura (1970) when developing equation 2 (i.e., that all N individuals potentially produce male and female gametes) is achieved by considering $P=M=N$ or $u=v=1$, that is, a perfect stand of plants in the field. For seed regeneration, this means that all plants in the field are used as male and female parents. On the other hand, if $P<N$, $M<N$ such that $u<1$, $v<1$, some plants fail to produce gametes due to poor germination, poor seed viability or a poor stand of plants in the field because of insects, diseases or other environmental factors.

In seed collection activities, parameters u and v are fractions of seed and pollen parents, respectively, that effectively contribute gametes for generating the sample of seeds collected. They should be measured in relation to the size of the reference population under natural conditions. For monoecious species, Crossa and Vencovsky (1994) derived the following $N_{e(v)}$ expressions for different alternative sampling schemes of female and male gametes. As shown in Annex A, we considered three different cases, depending on how male and female gametic control is performed.

- **Case 1:** Plant-to-plant hand pollination is practiced and equal numbers of seeds are taken from each plant (female plus male gametic control; *FGC + MGC* scheme) (equation 3).
- **Case 2:** Pollination is random and the same number of seeds is taken from each seed parent (female gametic control: *FGC* scheme) (equation 4).
- **Case 3:** Pollination is random and n seeds are randomly sampled from a bulk of seeds stemming from P seed parents (*RS* schemes) and M pollen parents (equation 5).

For seed collection, it can be assumed for random-mating species that the number of pollen parents (M) is very large; then potentially $M \approx N$ and $v \approx 1$. This over-simplification is used for calculating an upper limit for $N_{e(v)}$. Also, the number of seed parents can be considered to be much smaller than N ($P << N$) and $u \approx 0$. Thus, assuming that M is very large, $M \approx N$, *and* $v \approx 1$, the approximation to $N_{e(v)}$ is given by equations 6 and 7. Also, annex A shows how equation 6 can be written when the average number of seeds per seed parent is given by n/P.

Two-stage sampling model for mixed self- and random-mating species

Many self-compatible species have a high natural rate of self-fertilization (s) and thus may be considered as having a mixed self- and random-mating reproduction system. Logically, the genetic structure of a mixed self- and random-mating species is complex because not all plants have the same level of natural inbreeding. Furthermore, the progeny of a population of a mixed self- and random-mating species will have a mixture of selfed seed with proportion s and half-sibs or outcrossed seeds with proportion $1-s$. This model excludes biparental crosses between parents. Proportion s can be artificially manipulated by hand-pollination so that it becomes $s = 1$ when selfing all plants or $s = 0$ when crossing all plants. However, s varies from 0 to 1 in natural populations, which are assumed to be in inbreeding equilibrium when they reproduce naturally. In this case, $f = s/(2-s)$ [or $s = 2f/(1+f)$]. Vencovsky and Crossa (1999a) extended the two-stage sampling models for random-mating species to include mixed self- and random-mating species ($0 \leq s \leq 1$). They also developed estimates and direct expressions for computing $V_{(k)}$ and $N_{e(v)}$ with applications to specific aspects of germplasm regeneration and collection. The authors considered an initial set of N diploid plants where P plants (contributing male and female gametes) are randomly sampled from N and, subsequently, R plants (contributing only male gametes) are randomly sampled from $N-P$. Proportions u and v are defined as in two-stage sampling for random-mating monoecious species.

The model allows a correlation between the numbers of female and male gametes contributed per individual within the set of *P* plants (a correlation assumed to be zero for random-mating species). Therefore, the natural or artificial rate of self-fertilization (assumed to be constant over parents) is *s*, the proportion of crossing plants is 1−*s*, and *n* is the total number of seeds collected from the set of *P* parents. Then, 2*ns* is the expected total number of female and male gametes contributed from selfing and 2*n*(1−*s*) is the expected total number of male and female gametes contributed from crossing: the overall total of contributed gametes is 2*ns*+2*n*(1−*s*) = 2*n*.

Vencovsky and Crossa (1999a) considered random sampling of seed (*RS*) and female gametic control (*FGC*) under unrestrictive inbreeding and under inbreeding equilibrium. Here, we will consider only *RS* and *FGC* under inbreeding equilibrium. The general equations for a fraction *u* of potential parents contributing female gametes and *v*, the fraction of potential parents contributing male gametes, are given in Annex B. Equations 8 and 9 can be adapted for seed collection with the assumptions that the number of seed parents (*P*) is a very small fraction of the entire population ($u \approx 0$) and that the number of pollinating parents (for $s < 1$) is sufficiently large, allowing the assumption that $M \approx N$, such that $v \approx 1$. This last assumption tends to inflate the corresponding $N_{e(v)}$ value. In practical situations, the number of pollen parents for each seed parent can be estimated using molecular markers, but it requires a sample of adult plants and the corresponding maternal offspring to be genotyped.

Based on these assumptions, equations 10 and 11 were obtained for collecting *n* seeds from *P* seed parents (Vencovsky and Crossa 1999a) of a population in inbreeding equilibrium with a natural rate of self-fertilization *s*. A fundamental $N_{e(v)}$ expression is given by Cockerham (1969) for a group of *n* individuals or seeds differentiated from a reference population of infinite size solely by drift. See equation 12, which can be adapted for measuring the degree of representativeness contained in a single maternal descendant (equation 13), which is equivalent to equation 10 or 11 when $P = 1$.

Seed collection from several subpopulations

Under natural conditions, there may be situations in which a large population of a species, or meta-population, is made up of a group of subpopulations. When seeds are collected in such a situation, the number of subpopulations from which seeds should be sampled is an additional sampling unit to be considered and incorporated for estimating $N_{e(v)}$. Annex C shows some of the equations obtained by Vencovsky and Crossa (2003) when extending the theory developed by Cockerham (1969).

Vencovsky and Crossa (2003) derived $N_{e(v)}$ equation 14 (Annex C), where S^* is the total number of subpopulations in the region and C^2 is the squared coefficient of variation of the number of seeds collected from the *S* subpopulations, which is n_i for subpopulations *i*, such that $n = \sum n_i$ ($i = 1, 2, \ldots, S$). With equal n_i, $C^2 = 0$. Parameters F_{ST} and F_{IT} are Wright's measures of the divergence among subpopulations and the total inbreeding coefficient, respectively (Weir, 1996). These *F* statistics can be estimated for the parental generation using codominant genetic markers.

When seeds are not collected in bulk but on a progeny basis, an additional sampling unit is incorporated, namely the number of seed parents. Now the expression for estimating $N_{e(v)}$ is equation 15 (Vencovsky et al. 2007), where C_s^2 is the square of the coefficient of variation of the number of seeds among subpopulations and C_m^2 is the square of the coefficient of variation of the number of seeds among seed parents. Also, θ_{m} is the average coancestry among offspring within progenies, which can also be estimated using codominant molecular markers. Equations 14 and 15 (annex C) are useful for planning the collection sampling strategies necessary to obtain a desired $N_{e(v)}$ value.

Two-stage sampling model for dioecious species

Originally Crow and Kimura (1970) derived $N_{e(v)}$ expressions for species with separate sexes based on simplified assumptions. Later Crow and Denniston (1988) made the necessary correction of this earlier work. The aim of Vencovsky et al. (2011) was to consider the statistical properties of the number of contributed gametes in various practical situations. To this end, they transformed and adapted the formulas of Crow and Denniston (1988) in such a way that they are easy to apply in practical breeding situations and when collecting and regenerating genetic resources. Vencovsky et al. (2011) consider a finite reference

population such that all individuals are potentially functional parents; they also assume that, of the reference population, only a fraction of individuals is taken as functional parents. The derivations and results given by Vencovsky et al. (2011) refer only to dioecious species, and only the variance effective population size is considered. The reference population is considered under random mating, a condition also assumed by Crow and Denniston (1988).

Some of the derivations of Vencovsky et al. (2011) are shown in Annex D and refer to populations of arbitrary size, excluding some parents, that is, a fraction of parents is systematically discarded or rejected, such that only a subset of the initial N_f female and N_m male parents effectively participate in the reproduction process (Annex D, equations 16–18). Consider that, of the initial N_f female parents, only P participate in reproduction such that the fraction is $u = \frac{P}{N_f}$ $(0 < u \leq 1)$. For male parents (M), the fraction is $v = \frac{M}{N_m}$ $(0 < u \leq 1)$. For the rejected $N_f - P$ females and $N_m - M$ males, the number of gametes contributed will be zero. Consider that f (female) and m (male) offspring are sampled, with a total of $t = f + m$, and also that the sex ratio is $r = \frac{f}{(f+m)}$.

In another case, all the parents are included (Annex C, equations 19–21). In both cases, when some parents are excluded or all the parents are included, three cases of gametic control are considered, no gametic control or random sampling (*RS*), female gametic control (*FGC*), and female gametic control and male gametic control (*FMC+MGC*)

Practical applications

Probability models for number of alleles in the sample

The statistical genetic models used by Crossa et al. (1993) and based on equation 1 indicate that sample sizes of 160 to 250 plants of a random-mating population are required for capturing alleles at frequencies of 0.05 or higher in each of 150 loci, with 90%–95% probability. These formulas consider the sample size for conserving at least one allele per locus, but they do not quantify the probability of conserving two, three or more alleles per locus. The genetic diversity of a population depends on the number and frequency of alleles at a locus and across loci, and determining a sample size depends on whether estimates of allele frequencies are available (Hernandez and Crossa 1993).

However, when the required probability for conserving alleles at different loci increases or the frequency of a rare allele (p_0) drops to 1%, larger sample sizes than those specified earlier are required. For example, for unknown associations between genes within individuals and $p_0 = 0.001$, sample sizes of between 533 and 1708 individuals will retain, with a probability of 0.9999, at least one allele from each class, considering a wide range of alleles, $k = 2$ to 20, and a number of loci ranging from $m = 5$ to 1000. Sample size is always more affected by low allele frequency than by the number of alleles or the number of loci. To maintain alleles at 3% frequency with a 0.9999 probability, assuming unknown associations between genes within individuals, the required sample sizes are between 177 and 564 individuals for 2 to 20 alleles per locus and for 5 to 1000 loci. For retaining alleles at a frequency of 5%, 105 to 335 individuals are required. For 6 alleles per locus with allele frequencies between 0.03 and 0.10 and for 1 to 20,000 loci, a sample size of between 84 and 750 individuals will preserve at least one copy of each allele class in each locus with a probability of 0.9999999.

Assuming two alleles at each of the 20,000 loci and one of them at a 0.05 frequency, 186 individuals will preserve this allele at each locus with a 95% probability. However, a sample size of 172 individuals will only retain an allele at 5% frequency, assuming no association between genes within individuals at any locus. However, this does not seem to be sufficient if associations between genes within individuals exist at some loci and/or if maintaining alleles at frequencies between 3% and 1% is required.

Assuming that k-1 alleles occur at an identical low frequency for all loci makes the required sample size estimated by equation 1 conservative, because it is likely that some alleles will have higher frequencies, at least at some loci. The assumption of loci independence is unrealistic, since genetic drift may cause non-random associations between linked loci. If the population is under random mating (a cross-pollinated

species), linkage equilibrium can be assumed for all possible pairs of loci, so that a similar number of coupling and repulsion combinations can be expected to occur for each pair of loci. This obviously does not hold for self-pollinated species or species with mixed self- and random-mating systems of reproduction.

Recommendations for estimating general sample size using probability models

General recommendations and/or guidelines concerning sample sizes should be given

- within ranges (or intervals), depending on model assumptions and biological considerations such as number of alleles, loci, genetic linkage, etc.

Furthermore, it is important to point out that

- certain sample sizes (n) obtained from the probability models might or not provide a similar value for $N_{e(v)}$.

Two-stage sampling model for random-mating species – regeneration

For germplasm regeneration of N (finite) plants in the field, equations 3, 4 and 5 (annex A) can be adapted for different situations and mating designs, for monoecious species. Assuming that accession size is kept constant ($n = N$), these expression are much simplified, as illustrated in the following cases.

Two general cases are considered.

Case 1

The integrity of the accession is affected ($u = P/N \leq 1$), and there is loss of seeds from the original accession, but pollen and seeds are contributed by all remaining plants. Then $P = M = uN$ ($P = M$; $u = v$). This is similar to the case where seeds of an accession are lost due to environmental causes, during storage in the genebank, or while accession plants are growing in the field. The exception is when $u = 1$, as shown in Table 5.1.

Table 5.1: Variance Effective Population Size ($N_{e(v)}$) for Regenerating Accessions of a Monoecious Species Using Expressions Given for Cases 1a-c and Cases 2b-c for n=N=1000 and Different Proportions of Parents Contributing Male and Female Gametes to the Next Generation (u), Assuming Constant Accession Size

u	Case 1a* $N_{e(v)}=N[2u/(2-u)]$	Case 1b $N_{e(v)}=N[4u/(4-u)]$	Case 1c $N_{e(v)}=Nu$	Case 2b** $N_{e(v)}=N[4u/(1+2u)]$	Case 2c $N_{e(v)}=N[4u/(1+3u)]$
0.1	105	103	100	333	308
0.2	222	211	200	571	500
0.3	353	324	300	750	632
0.4	500	444	400	889	727
0.5	667	571	500	1000	800
0.6	857	706	600	1091	857
0.7	1077	848	700	1167	903
0.8	1333	1000	800	1231	941
0.9	1636	1161	900	1286	973
1.0	2000	1333	1000	1333	1000

Source: Extracted from Crossa and Vencovsky (1994).
* **Case 1a-c:** Pollen: $P=uN$, Seed: $M=uN$.
****Case 2b-c:** Pollen: $M=N$, Seed: $P=uN$.
(a) *FGC* + *MGC* (full sibs).
(b) *FGC* (half sibs).
(c) *RS* (random mating and random sampling).

Case 2

The integrity of the accession is not affected; that is, pollen is produced by all plants ($M = N$) but seed is collected from only P plants. Then, $M = N$, $v = 1$, but $P = uN$. This case simulates a situation where only a portion of the total number of pollinated plants in the field is harvested.

For **Case 1**, three mating designs based on different types of gametic control are considered:

1. female and male gametic control (full-sib mating system), *FGC + MGC*
2. female gametic control only (half-sib mating system), *FGC*
3. no female and male gametic control, that is, random open pollination, and unequal numbers of seeds are randomly taken from the set of *P* parents, *RS*

For **Case 2**, only schemes *FGC* and *RS* (b and c) are considered here because a scheme with control of both types of gametes (*FGC+MGC*) is not possible because, with *Nu* female parents available, only *Nu* pollen parents will be taken for obtaining full-sib progenies and not *N*.

In **Case 1a** (seed and pollen from $P = M = uN$), where female and male gametic control involves full-sib crosses and taking equal numbers of seeds from $P = M = uN$ plants, $N_{e(v)} = N[2u/(2-u)]$. In **Case 1b** (seed and pollen from $P = M = uN$), where random pollination involves half-sib crosses among $P = M = uN$ plants, and female gametic control implies taking equal numbers of seeds from $P = uN$ plants, $N_{e(v)} = N[4u/(4-u)]$. In **Case 1c** (seed and pollen from $P = M = uN$), with random pollination and unequal numbers of seeds randomly taken from P parents, $N_{e(v)} = Nu$.

In **Case 2b** (seed from $P = uN$ and pollen from $P = M$), where random pollination includes half-sibs among all plants, but female gametic control implies taking seeds from only $P = uN$ pollinated plants, $N_{e(v)} = N[4u/(1+2u)]$. In **Case 2c**, with random pollination and unequal numbers of seeds randomly taken from $P = uN$ parents, $N_{e(v)} = N[4u/(1+3u)]$.

The above expressions for **Cases 1 and 2** and the three mating systems, considering a range of accession losses (from 90% [u=0.l] to 0% [u=1, full stand of plants in the field]), are given in Table 5.1. For **Case 1**, with a perfect stand of plants (u=1), taking equal numbers of seeds per pollinated plant (female gametic control) and using hand-pollination (full-sibs) (male gametic control) produces $N_{e(v)} = 2N$; with only female gametic control and random pollination (half-sibs), $N_{e(v)} = (4/3)N$; and without male and female gametic control, $N_{e(v)} = N$. Values of $N_{e(v)}$ for **Case 2** (last two columns in Table 5.1) tend to be larger than their corresponding columns for **Case 1** (the first three columns in Table 5.1) because the pollen pool is larger when M=N than when M=uN (except when u=1).

> *How can a genebank curator regenerate an accession of a monoecious plant species that has lost a percentage of its seeds or plants, and still attain* $N_{e(v)} = N$*?*

In **Case 1**, a loss of 30% ($u = 0.7$) of the plant accessions can be recovered by male and female gametic control ($N_{e(v)}$=1077 when $n = 1000$), whereas a loss of 20% ($u = 0.8$) can be recovered by simply taking equal numbers of seeds from the 800 randomly pollinated plants that remained in the field ($N_{e(v)}$ =1000). Furthermore, in **Case 2**, if all 1000 plants are left to pollinate randomly (half-sibs), but equal numbers of seeds are taken from only 500 of them ($u = 0.5$), $N_{e(v)}$=1000 = n. This indicates the importance of controlling female gametes as a mechanism for recovering the desired magnitude of $N_{e(v)}$ because of the possible loss of seeds and/or plant accessions before and during the growing cycle (**Case 1**) or the potential problems encountered when harvesting only a portion of the total number of pollinated plants in the field (**Case 2**).

Table 5.1 shows that $N_{e(v)}$ is larger for female gametic control with half-sibs (**Case 2** [fifth column in Table 5.1]) than for female and male gametic control with full-sibs (**Case 1** [second column in Table 5.1]). This is true up to an accession loss equal to or greater than 30% ($u \leq 0.7$) (1167 versus 1077 for $u = 0.7$). When $u > 0.7$ under female and male gametic control with full-sib families from $P = M = u\,N$ plants, **Case 1** is

superior to that of half-sib families from $M = N$ plants, but female gametic control on only $P = uN$ families (1333, 1636 and 2000 versus 1231, 1286 and 1333 for u = 0.8 , 0.9 and 1.0, respectively). The half-sib mating system of **Case 2** is the best option for compensating for insufficient sampling of female parents (smaller values of u) through increasing $N_{e(v)}$, in contrast to the full-sib alternative of **Case 1**, which becomes important only for smaller accession losses ($u > 0.8$). On the other hand, female and male gametic control with the full-sib system in **Case 1** (second column in Table 5.1) is always superior to female gametic control with the half-sib system in **Case 1** (third column in Table 5.1) for all values of accession loss, as well as for the situation of no gametic control (fourth column in Table 5.1).

As already pointed out, when comparing half-sibs with female gametic control (**Case 2**) versus full-sibs and total gametic control (**Case 1**), the former is preferred if deterioration of the accession is intermediate to severe ($0 < u < 0.7$). Accession losses in **Case 1** increase drift that is attributable to sampling parents (first stage of sampling), but controlling female and male gametes should decrease drift due to sampling of gametes (second sampling stage) and, therefore, should help to control overall drift. The advantage of the half-sib system is that it does not require hand-pollination because pollen from all plants in the field can be used. It is a valuable alternative for decreasing the contribution to drift attributable to the sampling of parents (first stage).

Recommendations for regenerating random-mating monoecious species

Generally speaking, in the regeneration process discussed so far, two aspects are fundamental for maintaining adequate representativeness of accessions:

- avoiding excessive deterioration of the accession
- practicing female gametic control when sampling seeds after reproduction

Two-stage sampling model for random-mating species – collection

In collection activities, obtaining exact estimates of $N_{e(v)}$ when collecting seeds is not possible in most cases. This is due to the fact that the real size of the reference population is not known, and fractions u and v, therefore, cannot be estimated. Equations 6 and 7 (Annex A) are only approximations, since the number of female parents sampled for seed collection (P) is considered to be a very small fraction of the entire population ($u=P/N \approx 0$) and the number of pollen parents is admittedly very large, such that $v=M/N \approx 1$. Effective numbers, consequently, are overestimated. In any case, equations 6 and 7 can be considered for showing that under these assumptions, $N_{e(v)}$ is dominated by the number of seed parents (P). As an example, consider a sample of 1000 seeds taken from P = 100 (10 seeds per parent), P = 200 (5 seeds per parent) and P = 500 (2 seeds per parent). With female gametic control, the resulting $N_{e(v)}$ values are 308, 500 and 800, respectively (equation 6). If seeds are bulked and then sampled, the resulting $N_{e(v)}$ values for n = 1000 are 286, 447 and 667, respectively (equation 7). Once again, results demonstrate the importance of female gametic control in these activities.

The alternative form of equation 6 can be used for planning collection strategies. If, for instance, collecting 15 seeds per parent ($\bar{n}$=15) is feasible, and a value of $N_{e(v)}$=1000 is required for representing the reference population, then $N_{e(v)} = 4P\frac{\bar{n}}{\bar{n}+3} = 4P\frac{15}{18} = 1000$, and P = 300 is the number of seed parents necessary to achieve $N_{e(v)}$=1000.

The total number of seeds required in this case is n = 4500, and the average $N_{e(v)}$ per maternal progeny is $\bar{N}_{e(v)}$ = 3.33. An additional condition not mentioned earlier and required for the validity of computing P is that the seed parents should be genetically unrelated (negligible coancestry among all parents).

There should be a minimum distance among seed parents during seed collection in order to have negligible coancestry among adult individuals. Investigations on pollen distance and on the reproductive neighbourhood area are generally recommended for estimating this distance (Viegas et al. 2011). Molecular markers can be used for this propose.

Recommendations for collecting random-mating monoecious species

The main lessons for collectors of genetic resources are that $N_{e(v)}$ can be increased by

- increasing the number of seed parents (P)
- taking an equal number of seeds per seed parent, that is, exercising female gametic control

Two-stage sampling model for mixed self- and random-mating species – regeneration

Equations 8 and 9 (Annex B) show that $N_{e(v)}$ increases linearly with sample size n, for a given s, when regeneration occurs before any loss of plants in the field, with all parents contributing male and female gametes. Both equations assume inbreeding equilibrium and can be adapted to germplasm regeneration and collection, considering the effect of (1) an arbitrary rate of natural self-fertilization (s) and (2) random sampling of seeds or female gametic control (*FGC*). We will examine the same two cases we studied when considering the two-stage model for random-mating species, that is,

- **Case 1:** loss of accessions given by $u < 1$ and $v = u$, with seed parents $P = uN$, and pollen parents $M = uN$
- **Case 2:** no loss of accessions but seeds taken from uN pollinated plants; that is, seed parents $P = uN$ and pollen parents $M = N$, when $s < 1$. An additional effect that can be considered is given by the fact that the number of collected seeds (n) that can be taken is equal to N (constant $n=N$) or arbitrary ($n \neq N$).

Results obtained by Vencovsky and Crossa (1999a) indicated that *FGC* gives higher values of $N_{e(v)}$ than *RS*. Female gametic control combined with adequate levels of u ($u > 0.7$) becomes an effective combination for maintaining values of $N_{e(v)}$ equal to or higher than n, when $n=N$; whereas, with *RS*, values of $N_{e(v)}$ are always smaller than sample size (n). In Cases 1 and 2 (here and below), *RS* of cross-fertilizing species ($s = 0.0$) is less affected by accession deterioration than are more self-fertilizing species at any level of accession deterioration (u). However, in **Case 1** and **Case 2**, *FGC* of cross-pollinating species gives slightly higher values of $N_{e(v)}$ than that of more self-fertilizing species but only for severe values of accession deterioration; the reverse is true for mild accession loss. Vencovsky and Crossa (1999a) found that $N_{e(v)} = n$ for high values of u and when the species approaches panmixia ($s = 0.0$), under *RS*.

Other findings from Vencovsky and Crossa (1999a) relative to germplasm regeneration can be pointed out, as given below.

Female gametic control (*FGC*) always gives higher $N_{e(v)}$ values than random sampling (*RS*) in all circumstances. This practice produces a rapid increase in $N_{e(v)}$ for autogamic species when $u > 0.7$.

In **Cases 1 and 2**, when seeds are taken randomly (*RS*), accession regeneration is more efficient for panmictic species than for autogamic ones, under any level of accession deterioration. This is also true with *FGC* and severe deterioration loss. On the other hand, autogamic species with *FGC* produce higher $N_{e(v)}$ in regeneration than allogamic species when u is mild.

For germplasm collection, the results obtained by Vencovsky and Crossa (1999a) contradict those for regeneration; that is, attaining adequate $N_{e(v)}$, even with *FGC* for collecting autogamic species, is more difficult than regenerating accessions of this category when $u > 0.8$. The reverse is true for panmictic species.

For constant accession size ($n=N$) and *FGC*, considered here as a reference point, the minimum value of u (u^*) acceptable for having $N_{e(v)}=n$ is given by $u^*=4/(s^2+5)$. Evaluating this quantity gives $u^*= 0.8$, 0.76 and 0.67 for $s = 0.0$, 0.5 and 1.0, respectively. This indicates that more autogamic species permit slightly higher levels of accession deterioration, compared to more panmictic ones. Values given in Table 5.2 confirm this.

Increasing sample size (n) during regeneration increases $N_{e(v)}$, but only when $s < 1$. For perfectly autogamic species ($s = 1$), increasing n has no effect on $N_{e(v)}$, as can be seen in Table 5.2.

Table 5.2: Variance Effective Population Size ($N_{e(v)}$) When Regenerating N= 500 Plants and Harvesting an Arbitrary Number of Seeds (n) for Different Values of Accession Loss or Proportion of Functional Parents (u) and Natural Rate of Self-Fertilization (s), Assuming Inbreeding Equilibrium

u	*s*	*n*	**Case 1 – *RS* seed: *P* = *uN* pollen: *M* = *uN***	**Case 1 – *FGC* seed: *P* = *uN* pollen: *M* = *uN***	**Case 2 – *RS* seed: *P* = *uN* pollen: *M* = *N***	**Case 2 – *FGC* seed: *P* = *uN* pollen: *M* = *N***
0.5	0	500	250.3	285.9	400.5	500.3
0.5	0	1000	333.6	363.8	667.3	800.3
0.5	0	3000	428.7	444.5	1200.7	1333.6
0.7	0	500	350.4	424.5	452.1	583.7
0.7	0	1000	538.9	622.5	824.4	1037.6
0.7	0	3000	840.3	903.4	1827.4	2154.6
0.9	0	500	450.5	581.0	487.0	643.3
0.9	0	1000	818.9	1029.2	948.3	1242.2
0.9	0	3000	1801.2	2118.5	2573.8	3274.5
0.5	0.5	500	187.8	261.0	240.4	375.1
0.5	0.5	1000	250.2	307.8	353.4	480.1
0.5	0.5	3000	321.6	349.5	514.6	590.2
0.7	0.5	500	262.9	433.2	302.6	552.8
0.7	0.5	1000	404.3	579.5	506.7	815.8
0.7	0.5	3000	630.3	747.9	920.4	1194.5
0.9	0.5	500	338.0	683.9	353.4	750.4
0.9	0.5	1000	614.4	1137.3	667.6	1333.9
0.9	0.5	3000	1351.2	2038.2	1638.2	2770.1
0.5	1	500	125.3	250.0	—	—
0.5	1	1000	166.9	250.0	—	—
0.5	1	3000	214.4	250.0	—	—
0.7	1	500	175.4	583.3	—	—
0.7	1	1000	269.6	583.3	—	—
0.7	1	3000	420.3	583.3	—	—
0.9	1	500	225.5	2250.0	—	—
0.9	1	1000	409.8	2250.0	—	—
0.9	1	3000	901.2	2250.0	—	—

Note: Cases 1 and 2 with random sampling of seeds (*RS*), female gametic control (*FGC*).
Equations 8 and 9 used for *RS* and *FGC* $N_{e(v)}$ values, respectively.
For Case 2, v=1.

How can a genebank curator regenerate an accession of a mixed self- and random-mating species that has lost a percentage of seeds or plants and still attain $N_{e(v)}$=n*?*

We can study the advantages of *FGC* over *RS* systems under **Cases 1 and 2**. For this situation, we consider an arbitrary value of *n* ($n \neq N$). Values of $N_{e(v)}$ for regenerating a hypothetical accession of *N*=500 plants in the field for **Cases 1 and 2**, under *RS* and *FGC* for various values of accession loss (1-*u*), natural rate of self-fertilization (*s*) and sample size (*n*) are shown in Table 5.2, assuming that plants are in equilibrium with respect to the mating system.

In Table 5.2, effective sizes for **Case 2** are always higher than the corresponding ones for **Case 1**, because of the larger pollen pool in **Case 2**, where *M*=*N* as compared to *M*=*uN* in **Case 1**. With complete autogamy (*s* = 1), the condition *M*=*N* is impossible because there is complete self-fertilization. This situation (*M*=*N* with *s* = 1) was wrongly included by Vencovsky and Crossa (1999a). The advantage of **Case 2** over **Case 1** can be visualized by comparing $N_{e(v)}$ values within the *RS* and *FGC* systems.

As already mentioned, considering only **Case 1**, Table 5.2 shows that for small losses in accessions (large *u*) it is easier to regenerate more autogamic species, whereas for large accession losses (small *u*) it is easier to regenerate more panmictic species.

If we collect *n*=*N*=500 seeds from 250 plants (*u* = 0.5) by taking two seeds per plant (*FGC*) of a panmictic species (*s* = 0) but using pollen from all *M*=*N*=500 plants (**Case 2**), the required condition $N_{e(v)}$=*N*=500 is achieved ($N_{e(v)}$=500.3 in the last column of Table 5.2). On the other hand, if we collect *n*=500 seeds from 250 plants, by taking two seeds per plant (*FGC*) of an autogamous species (*s* = 1) (**Case 1**), the required condition is never achieved because $N_{e(v)}$=(1/2)*N*=250, irrespective of sample size *n*. As expected, this indicates that severe accession loss has a greater impact on autogamic species than on panmictic ones. If pollen from all 500 plants of a panmictic species (*s* = 0) is used and 12 seeds (*FGC*) are taken from only 250 plants (*u* = 0.5 and *n* = 3000) (**Case 2**), $N_{e(v)}$ = 1333.6 = 2.67*N* and $N_{e(v)}$ = 1200.7= 2.40*N* with *RS* (Table 5.2). In **Case 1**, where the pollen pool is smaller than in **Case 2** (for *s* < 1), with *FGC*, the condition that $N_{e(v)} \geq N$ is only achieved for a smaller accession loss (*u* > 0.7) than that allowed for **Case 2** and *u* = 0.5.

In **Case 1**, when there is an accession loss of 10% (*u* = 0.9) and 3000 seeds are collected under *FGC* from a panmictic species (*s* = 0), $N_{e(v)}$= 2,118.5 > 4*N*. In **Case 2**, a 30% accession reduction on the female side (*u* = 0.7) and collecting *n* = 3000 will produce $N_{e(v)}$ = 2154.6 = 4.31*N*.

For the three types of species considered, the condition that $N_{e(v)} \geq N$ is reached for $u \geq 0.7$, under **Case 1**. In Table 5.2, it can be seen that the highest $N_{e(v)}$ values are obtained by practicing *FGC* and that fractions of accession loss are very small (*u* = 0.9). In general terms, the combination of female gametic control with the possibility of having a larger pollen pool, as in **Case 2**, produces important increases in $N_{e(v)}$ for *u* = 0.7, 0.8 and 0.9, especially for sample size *n* >*N*. It is worth mentioning that the ideal situation for regenerating an accession occurs with *FGC* in autogamic species (*s* = 1) and no accession loss (*u* = 1). In such a case, there is no drift between the original and the regenerated accession and $N_{e(v)} = \infty$ (equation 9). Practicing random sampling in such a situation is inadequate.

Recommendations for regenerating mixed self- and random-mating species

- Taking an equal number of seeds (*FGC*) gives higher values of $N_{e(v)}$ than a random number of seeds (*RS*).
- Female gametic control combined with adequate percentage germination (u >0.7) becomes an effective combination for maintaining values of $N_{e(v)}$ equal to or higher than *n*, when *n*=*N*; whereas, with *RS*, values of $N_{e(v)}$ are always smaller than sample size *n*.
- Random sampling of cross-fertilizing species (*s* = 0.0) is less affected by accession deterioration than it is with more self-fertilizing species at any level of accession decrease in percent germination.
- The combination of female gametic control with the possibility of having a larger pollen pool produces important increases in $N_{e(v)}$. The ideal situation for regenerating an accession occurs with *FGC* in autogamic species (*s* = 1) and no accession loss (*u* = 1). In such a case, there is no drift between the original and the regenerated accession and $N_{e(v)} = \infty$. Practicing random sampling in such a situation is inadequate.

Two-stage sampling model for mixed self- and random-mating species – collection

For collection activities, equations 10 and 11 (Annex B) are used to perform some numerical evaluations. It is assumed that n = 1000 seeds are collected from P=100 and P=20 adults, assuming random sampling (*RS*) and female gametic control (*FGC*), for different values of the natural rate (s) of self-fertilization. In natural conditions, the population is assumed to be in equilibrium relative to the mating system. Corresponding $N_{e(v)}$ values are given in Table 5.3.

Table 5.3: Effective Sizes $N_{e(v)}$ of n=1000 Seeds Collected from P=100 and P=20 Seed Parents, for Random Sampling (*RS*) and Female Gametic Control (*FGC*), from a Population in Inbreeding Equilibrium and Rate of Self-Fertilization (s)

s	*P*=100		*P*=20	
	RS	***FGC***	***RS***	***FGC***
0.0	285.9	307.7	74.1	75.5
0.3	162.8	177.0	38.5	39.2
0.7	79.1	86.6	17.5	17.9
1.0	45.5	50.0	9.8	10.0

Note: Equations 10 and 11 used for *RS* and *FGC* $N_{e(v)}$ values, respectively.

As can be seen in Table 5.3, the dominating factors in the effective size values, for n =1000 seeds collected, are the number of seed parents and the population's rate of self-fertilization. Practicing gametic control has a positive effect for increasing $N_{e(v)}$; however, that diminishes as the number of seed parents decreases.

Values such as those given in Table 5.3 can be used as a guide for planning sampling procedures when a specific $N_{e(v)}$ value is desired. With *FGC*, P=100 (10 seeds per parent) and under panmixia (s = 0), we have an average effective size value of $N_{e(v)}$=3.1 per seed parent. To attain a value of $N_{e(v)}$=1000, in this case, P = 322 seed parents would be necessary under the supposition that they are genetically unrelated (negligible or zero coancestry among them). For autogamy (s =1.00) and P =100, we have an average $\bar{N}_{e(v)}$= 0.5 per seed parent, meaning that P =2000 would be required to reach a total $N_{e(v)}$ of 1000. This specific example reinforces the fact that obtaining seed samples with adequate representativeness is more difficult in more autogamic species.

Effective size values for a single maternal progeny, under the assumption of mixed self- and random mating, can be obtained directly from equation 13 (Annex B). Table 5.4 gives an $N_{e(v)}$ that is applicable to this case for several sample sizes (n) and self-fertilization rates (s). Considering the same example as in the preceding paragraph, with n =10 seeds per seed parent and s = 0.00, $N_{e(v)}$ = 3.1, as found earlier. The following numbers of seed parents (P) would be necessary for having a joint value of $N_{e(v)}$= 1000, for s = 0.0, 0.25, 0.5, 0.75 and 1.0, respectively, with n =10 seeds collected per seed parent: P =322, 526, 833, 1250 and 2000. For s =1.00, $N_{e(v)}$= 0.5 per progeny, which corresponds to a single gamete sampled from an idealized population.

In a broad review involving 30 tree species conducted by Sebbenn (2006), the number of seed parents (P) was computed for attaining an effective size of $N_{e(v)}$= 150. Values were obtained based on estimates of the average coancestry ($\bar{\theta}$) and the inbreeding coefficient (F), using molecular markers. The number of seeds collected per tree was considered large, such that (n-1)/n » 1 and (1+ $\bar{F}$)/2n is negligible. Values of the number of seed parents (P) varied considerably among species, between 44 and 144, with an average of $\bar{P}$ = 67. Most species showed a predominantly allogamic mating system with s varying between 0.0 and 0.25 and an average value of $\bar{s}$ = 0.08.

Table 5.4: Variance Effective Population Size ($N_{e(v)}$) Values of a Single Maternal Progeny for Several Sample Sizes (n) and Natural Rates of Self-Fertilization (s), Reference Population of Infinite Size in Equilibrium under Mixed Self- and Random Mating

s	*n*				
	5	**10**	**50**	**100**	**limit**
0.00	2.5	3.1	3.8	3.9	4.0
0.25	1.7	1.9	2.2	2.2	2.2
0.50	1.2	1.2	1.3	1.3	1.3
0.75	0.8	0.8	0.8	0.8	0.8
1.00	0.5	0.5	0.5	0.5	0.5

Note: $N_{e(v)}$ values obtained using equation 13.

Recommendations for collecting mixed self- and random-mating species

The dominating factors in the effective size values are

- the number of seed parents
- the population's rate of self-fertilization (s)

While practicing gametic control has a positive effect for increasing $N_{e(v)}$, that diminishes as the number of seed parents decreases.

Two-stage sampling model for dioecious species

Several dioecious plant species are economically important, such as some domesticated dioecious arboreal species (Bandel and Gurgel 1967). Accounting for approximately 6% of angiosperms, dioecious species evolved from hermaphrodites in multiple independent events (Barret et al. 2010). Bawa (1980) mentions that dioecy is not as rare as is generally assumed. Bawa et al. (1985) found that there is a high level of dioecy (> 20%) among 333 tropical tree species. A similar percentage (22%) was found by Bawa (1974) among 130 lowland tropical tree species, and in the tropical savannah, it was found that about 15% of woody species are dioecious (Oliveira 1996; Oliveira and Gibbs 2000). Studying tropical coastal vegetation, Matallana et al. (2005) found 14% dioecy among 566 species and a higher percentage (35%) among dominant woody plants. In general, the percentage of dioecy among tropical forest trees and shrubs varies between 16% and 28% (Queenborough et al. 2007).

Regenerating dioecious species

Table 5.5 gives the results obtained by studying the effect of reducing seed viability or the germination rate (u; v) of an accession initially containing $N_f = N_m = 100$ individuals or seeds. Functional female and male parents are $P=uN_f$ and $M=vN_m$, respectively (Annex D, equations 16–18).

Suppose enough resources are available for a plant-to-plant crossing scheme between the $N_f = 100$ females and males, aimed at achieving a total number of offspring of at least $t = 200$. After planting the seed, it is verified that only 50 of the male and female plants germinated. If female and male gametic control is practiced (by hand-pollinating female plants using plant-to-plant crosses and by taking an equal number of seeds per female plant), the resulting effective size is $N_{e(v)}= 133.3$ (Table 5.5) for $r = 0.5$ in the offspring generation. Equation 18 gives this value by taking $t = 200$, $M = P = 50$, $u=v=0.5$ and $r = 0.5$. To achieve $N_{e(v)} = 200$ under these conditions, the number of parents should be higher, namely $P=M = 66.7$, such that $u=v=0.67$. For $P=M=70$ and $u= v = 0.7$, $N_e= 215$ for $r = 0.5$ (Table 5.5). To reach $N_{e(v)}= t = 200$ and the initial sizes, the following number of functional parents of each sex ($P=M$) would be necessary: 66.7, 68.6, and 76.4, for sex ratios $r = 0.5$, $r = 0.4$ and $r = 0.3$, respectively. These results indicate that losses of 25% to 30% of plants in the field due to poor germination or to any environmental cause might be compensated for, at least in part, by female and male gametic control. If all female and male parents are used in the

crossing block and no loss of plants has occurred, the effective population size is doubled for $r = 0.5$ and reaches $N_{e(v)} = 369.2$ and $N_{e(v)} = 289.6$ for $r = 0.4$ and $r = 0.3$, respectively, with gametic control on both sexes. All values given here refer to a sample size of $t = 200$. As expected, when the sex ratio deviates from $r = 0.5$, there is a reduction in $N_{e(v)}$.

Table 5.5: Variance Effective Population Size ($N_{e(v)}$) Values for Examples of Genebank Accession Regeneration with an Original Dioecious Reference Population with $N_f = N_m = 100$, under Random Mating, Assuming That $t = 200$ Offspring Are Sampled; Reducing Values of u and v ($u = v$); Three Sex-Ratio Levels ($r = 0.3$, $r = 0.4$ and $r = 0.5$) and Three Procedures of Gametic Control Random Sampling (*RS*), Female Gametic Control (*FGC*) and Female Plus Male Gametic Control (*FGC + MGC*)

			$N_{e(v)}$		
$N_f = N_m$	$P = M$	$u = v$	*RS*	*FGC*	*FGC + MGC*
----------	----------	----------	$r = 0.3$	----------	----------
100	100	1	168.7	213.2	289.6
100	70	0.7	124.1	146.5	178.7
100	50	0.5	91.7	103.3	118.3
100	30	0.3	57.0	61.2	66.1
100	10	0.1	19.1	20.2	20.6
100	5	0.05	10.0	10.0	10.2
----------	----------	----------	$r = 0.4$	----------	----------
100	100	1	192.9	253.4	369.2
100	70	0.7	136.7	164.4	206.1
100	50	0.5	98.4	111.9	129.7
100	30	0.3	59.6	64.2	69.6
100	10	0.1	20.0	20.5	21.0
100	5	0.05	10.0	10.1	10.2
----------	----------	----------	$r = 0.5$	----------	----------
100	100	1	201.0	267.6	400.0
100	70	0.7	140.7	170.2	215.4
100	50	0.5	100.5	114.6	133.3
100	30	0.3	60.3	65.0	70.6
100	10	0.1	20.1	20.6	21.0
100	5	0.05	10.0	10.2	10.2

Source: Extracted from Vencovsky et al. (2011).
Note: RS, *FGC* and *FGC + MGC* based on equations 16, 17 and 18, respectively.

As an alternative to the previous examples, consider that of the $N_f = 100$ only 33 are available. With a germination rate of only 50%, the remaining functional parents will be $P = 16$ and $M = 50$, such that $u = 0.16$ and $v = 0.5$. Since plant-to-plant crosses are not possible, male gametic control is not practicable and the option is to apply female gametic control. With these values and using equation 17, for $t = 200$ and $r = 0.5$, for the offspring generation, the resulting effective size is $N_{e(v)} = 51.7$, which is relatively small due to the considerable reduction in the number of female parents and the impossibility of controlling the number of contributed male gametes. With the basic quantities remaining the same but a germination rate of 70%, $P = (0.7)(33) \approx 23$, $M = (0.7)(100) = 70$, $u = 0.23$ and $v = 0.70$, for a result of $N_{e(v)} = 75.9$.

When all individuals of the reference population are functional parents, the following values are obtained for $t = f + m$ and $r = 0.5$, $N_{e(v)} = t$ for *RS*, $N_{e(v)} = 1.33t$ for *FGC*, and $N_{e(v)} = 2t$ for *FMC+MGC* (see Annex D, equations 19, 20 and 21).

Recommendations for regenerating dioecious species

Maintaining high levels of germination and viability is fundamental for preserving adequate $N_{e(v)}$ values of accessions in genebanks. In the example,

- a 20% increase in the germination rate leads to an increment of 24.2 units, or 46.8%, in $N_{e(v)}$;
- deviations of the sex ratio from $r = 0.50$ only have sizable effects on $N_{e(v)}$ for larger values of u and v or when the viability of the accession is higher than 50% or 60%.

Collecting dioecious species

When collecting germplasm, the number of seed parents from which the seed samples are taken is, in general, a small fraction of the total population size, such that u tends to be small or negligible. In relation to the pollinator plants, the corresponding fraction v can also be smaller or larger than u, depending on the factors affecting pollen dispersal, the structure of the population and the size of the reference population.

Suppose that resources are available for collecting only $t = 200$ seeds, stemming from $P = 100$ female parents randomly crossed with $M = 100$ males within an ecogeographic area. Assume that the population of females and males is very large ($N_f = N_m = 10{,}000$) such that only 1% of the total population is sampled ($u = v = 0.01$), and that female gametic control is used. The expected population size will be $N_{e(v)} = 115.1$ (equation 17). For such small fractions u and v, the effective size is smaller than the sample size ($t = 200$) even with female gametic control (Table 5.6). Control of both types of gametes is rarely applicable in collection activities, but just for comparison, let us assume complete control of gametes. Effective size would then be $N_{e(v)} = 134.2$ (equation 18), with a gain of 16.6% relative to $N_{e(v)} = 115.1$. In this example, it was assumed that the sex ratio is $r = 0.5$ for both offspring and parents.

Table 5.6 shows some applications of equations 16 and 17, with a constant number of functional parents but a decreasing degree of representativeness ($u = v$) and increasing size of the reference population (N_f; N_m). It is assumed that the sex ratio remains the same between generations. Here parameter r is included as a peculiarity of some species. As already seen in previous items, as r deviates from the value $r = 0.5$, N_e becomes smaller. It should also be noted that sampling 0.1% or 5.0% of individuals in the parental generation has little effect on $N_{e(v)}$. Female gametic control is required for obtaining $N_{e(v)} = t$, but the effect is noticeable only when u and v are large.

Estimating $N_{e(v)}$ in dioecious species requires distinguishing the sex of the offspring. Since this is generally difficult or impossible in plant seedlings or seeds, some strategies are necessary. An indirect procedure is estimating r at the adult stage in a natural population and transferring this value to the offspring.

Another procedure is achieving some protection through gametic control. Establishing that a sample of t seeds should have an effective size $N_{e(v)} \geq t$ and if the entire set of parents participate in reproduction, the given condition will be achieved through female gametic control for a range of r values between 0.28 and 0.72. With control of both types of gametes (*FGC+MGC*), this range is increased to $0.21 \leq r \leq 0.79$ gametic control. Therefore, gametic control is also important when the sex ratio is not known.

Recommendations for collecting dioecious species

As r deviates from the value $r = 0.5$, $N_{e(v)}$ becomes smaller.

- Female gametic control is required for obtaining increased values of $N_{e(v)}$; however, the effect is noticeable only when the percentage of seed germination is large.
- Sampling 5% or 0.1% of adult plants of the population for seed collection has negligible effect on $N_{e(v)}$.

Table 5.6: Variance Effective Population Size ($N_{e(v)}$) Values for Examples of Germplasm Collection in Dioecious Species, with Increasing Size of the Reference Random Mating Population (N_f; N_m), a Constant Total Number of Parents (P+M=200) and a Reducing Fraction of Functional Parents Sampled (u and v). Total number of seeds collected: t=200.

					$N_{e(v)}$	
N_f	N_m	P	M	$u = v$	RS	FGC
				r = 0.3		
60	140	60	140	1	168.8	213.1
86	200	60	140	0.7	129.7	154.3
120	280	60	140	0.5	112.4	130.4
200	467	60	140	0.3	99.1	112.9
600	1400	60	140	0.1	88.7	99.5
1200	2800	60	140	0.05	86.4	96.6
6000	14000	60	140	0.01	84.6	94.5
60000	140000	60	140	0.001	84.3	94.0
				r = 0.4		
80	120	80	120	1	193.0	253.3
114	171	80	120	0.7	148.3	181.5
160	240	80	120	0.5	128.4	152.6
267	400	80	120	0.3	113.3	131.7
800	1200	80	120	0.1	101.3	115.8
1600	2400	80	120	0.05	98.7	112.4
8000	12000	80	120	0.01	96.7	109.8
80000	120000	80	120	0.001	96.3	109.3
				r = 0.5		
100	100	100	100	1	201.0	267.6
143	143	100	100	0.7	154.4	190.9
200	200	100	100	0.5	133.8	160.3
333	333	100	100	0.3	118.0	138.2
1000	1000	100	100	0.1	105.5	121.4
2000	2000	100	100	0.05	102.8	117.8
10000	10000	100	100	0.01	100.8	115.1
100000	100000	100	100	0.001	100.3	114.5
				r = 0.6		
120	80	120	80	1	193.0	253.6
171	114	120	80	0.7	148.3	181.6
240	160	120	80	0.5	128.4	152.7
400	267	120	80	0.3	113.3	131.8
1200	800	120	80	0.1	101.3	115.9
2400	1600	120	80	0.05	98.7	112.5
12000	8000	120	80	0.01	96.7	109.9
120000	80000	120	80	0.001	96.3	109.3

Table 5.6: Continued.

					$N_{e(v)}$	
N_f	N_m	P	M	$u = v$	RS	FGC
---------- r = 0.7 ----------						
140	60	140	60	1	168.8	213.6
200	86	140	60	0.7	129.7	154.6
280	120	140	60	0.5	112.4	130.6
467	200	140	60	0.3	99.1	113.0
1400	600	140	60	0.1	88.7	99.6
2800	1200	140	60	0.05	86.4	96.7
14000	6000	140	60	0.01	84.6	94.6
140000	60000	140	60	0.001	84.3	94.1

Source: Extracted from Vencovsky et al. (2011).
Note: Two procedures of gametic control were used: RS (random sampling) and FGC (female gametic control).
Five sex-ratio levels (r=0.3, 0.4, 0.5, 0.6 and 0.7) were assumed equal for offspring and parents.
A sample of t=200 offspring is assumed.
RS and *FGC* are based on equations 16 and 17, respectively.

Sampling seeds from subpopulations

As already mentioned, there are situations in which seed samples are intended to represent a large population composed of subpopulations or fragments in natural conditions. In such a case, the number of subpopulations (S) is an additional sampling unit to be taken into account.

In this example, we consider that a sample of n =1000 seeds is taken from S subpopulations. We assume that an equal number of seeds (n/S) is sampled in bulk from each subpopulation and obtain $N_{e(v)}$ values by applying equation 14 (Annex C). For evaluation, four degrees of genetic divergence due to drift among subpopulations (F_{ST} = 0.02, 0.05, 0.10 and 0.15) are incorporated. These subpopulations are considered to be in equilibrium with respect to the mating system, with a constant natural rate of self-fertilization (s), such that, within each one, the natural level of inbreeding is $F_{IS} = s/(2-s)$. The relationship $(1-F_{IT})=(1-F_{ST})(1-F_{IS})$ (Weir 1996) was used to obtain the total level of inbreeding F_{IT}, which is also necessary for applying equation 14. To widen the example, we include two other potential situations, namely, (1) that the number of populations existing in the region is very large ($S^*=\infty$) or (2) that this number is relatively small (S^*=50). Pertinent $N_{e(v)}$ values obtained for this example are given in Table 5.7. In equation 14, we considered that $C^2 = 0$.

A striking outcome in this example is that $N_{e(v)}$ is strongly dependent on the number S of subpopulations sampled. This can be seen even when the divergence among subpopulations is only a small fraction of the total genetic diversity, such as when it is only 2% (F_{ST} = 0.02). Effective size barely reaches $N_{e(v)}$ =500 for $s < 50$ and $S^*=\infty$, under random mating within subpopulations (s = 0) and with a mixed mating system (s = 0.5). The degree of among-subpopulation divergence (F_{ST}) also shows a predominant influence upon $N_{e(v)}$.

Equation 14 can be used for determining the number S necessary for obtaining a given $N_{e(v)}$ value. If n is sufficiently large and $S^*=\infty$, this expression leads to $S \approx 2F_{ST}\, N_{e(v)}$. If, for instance, F_{ST} = 0.05 and the desired $N_{e(v)}$ is 1000, then the necessary number is $S \approx 100$ for large n, such that $(1+F_{IT})/2n$ is negligible.

When the reference population contains a finite set of subpopulations (S^*= 50 in the example), the general trend remains the same. Effective size values for subpopulations with a mixed mating system (s = 0.5) are always smaller than the corresponding values obtained for s = 0, as was observed for the infinite model ($S^*=\infty$). Also, in general, $N_{e(v)}$ diminishes as among-subpopulation divergence (F_{ST}) increases. An exception occurs when all subpopulations are included in the sample ($S = S^*$= 50). As expected, it is much

easier to represent a population in a sample when it is composed of a finite set of subpopulations than when a population includes a very large number of subpopulations. Such results demonstrate the importance of clearly defining the reference populations when computing an $N_{e(v)}$ value.

Table 5.7: Effective Size ($N_{e(v)}$) Values for Examples of n=1000 Seeds Collected from S Subpopulations, for Increasing Among-Subpopulation Divergence (F_{ST}), Two Levels of Self-Fertilization (s) within Subpopulations and Two Models of Natural Population Size (S*)

***n* = 1000**		***s* = 0***				***s* = 0.5**			
	S	F_{ST}=0.02	F_{ST}=0.05	F_{ST}=0.10	F_{ST}=0.15	F_{ST}=0.02	F_{ST}=0.05	F_{ST}=0.10	F_{ST}=0.15
$S^* = \infty$	2	47.7	19.6	9.9	6.6	46.9	19.5	9.9	6.6
	5	111.4	47.7	24.4	16.4	107.4	47.0	24.3	16.4
	10	200.8	91.3	47.8	32.4	188.4	88.8	47.2	32.1
	25	387.6	202.0	112.4	77.8	344.0	189.9	108.7	76.1
	40	505.1	289.9	169.5	119.8	433.5	265.5	161.3	115.8
	50	561.8	339.0	204.1	146.0	474.7	306.1	192.3	140.2
$S^* = 50$	2	48.6	20.0	10.1	6.8	47.9	19.9	10.1	6.8
	5	120.1	51.8	26.6	17.9	115.6	50.9	26.4	17.8
	10	235.6	109.7	58.0	39.5	218.7	106.0	57.1	39.0
	25	556.7	334.4	200.7	143.4	471.0	302.3	189.3	137.8
	40	844.5	684.8	520.7	420.1	661.9	562.8	450.4	375.4
	50	1020.4	1052.6	1111.1	1176.5	765.3	789.5	833.3	882.4

* $N_{e(v)}$ values obtained using equation 14.

So far, we have stressed the importance of having a sufficient number S of subpopulations as sources for seed collection. In practice, we understand that such a sample of S subpopulations should not be taken at random. The fact is that, very often, adjacent subpopulations have a certain degree of genetic resemblance, a phenomenon detectable when genetic distances between all pairs of subpopulations are correlated with the corresponding geographic distances. This resemblance can be evaluated using morphological measurements or genetic data based on molecular markers. This technique involves estimating spatial autocorrelation, as shown by Sokal and Oden (1978a,b) and originally by Moran (1950). The ideal is to collect seeds from a subset S of subpopulations having the smallest possible degree of resemblance to each other. General results obtained by applying equation 14, as well as those given in Table 5.7, are applicable as strategies for *in situ* preservation.

Recommendations

- $N_{e(v)}$ is strongly dependent on the number S of subpopulations sampled.
- The degree of among-subpopulation divergence (F_{ST}) also shows a strong influence upon $N_{e(v)}$. $N_{e(v)}$ diminishes as among-subpopulation divergence (F_{ST}) increases. An exception occurs when all subpopulations are included in the sample. As expected, it is much easier to represent a population in a sample when it is composed of a finite set of subpopulations than when a population includes a very large number of subpopulations.
- It is important to define the reference populations when computing $N_{e(v)}$.
- When possible, collect seeds from a subset S of subpopulations having the smallest possible degree of resemblance to each other.

General recommendations

Regeneration and seed collection

- Practicing female gametic control (harvesting an equal or approximately equal number of seeds per seed parent) is always recommended for all species.
- When possible and especially for valuable materials, always obtain genetic information concerning the mating system of the species, based on molecular markers. This is especially important for species with mixed mating systems.
- A computer programme for analysing marker data is available for this purpose: MLTR win (Multilocos Mating System Program) by Kermit Ritland (http://genetics.forestry.ubc.ca/ritland/programs.html).

Regeneration

- Avoid a loss of viability or germination below 60%–70% during storage of germplasm accessions.
- In addition to female gametic control, also apply male gametic control when species is a cross-fertilizer. This requires plant-to-plant hand pollination and is important for accessions that have low germination (high degrees of deterioration).

Collection

- In the case of a single population, avoid harvesting large amounts of seeds from a small number of seed parents. This is valid for all species. To attain reasonable representativeness, harvest seeds from at least 25 seed parents for cross-fertilizing species and 50 seed parents for autogamous species.
- In the case of large and genetically structured populations in natural conditions, seeds should be collected from the largest possible number of sites or subpopulations. This is recommended when no additional information is available about the genetic structure of the population.
- When possible, always obtain measures of the genetic divergence among collection sites or subpopulations of the species in a given ecogeographic region, and of the level of natural inbreeding. This can be done as an activity parallel to collection. The following computer programmes are available for this purpose: FSTAT by Jérome Goudet (www2.unil.ch/popgen/softwares/fstat.htm) and GDA (Genetic Data Analyses) by Paul O. Lewis and Dmitri Zaykin: (http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php).

Future challenges

Making general considerations about future challenges in regeneration and collection activities is not simple. Every programme in the area of genetic preservation and conservation has its own peculiarities, objectives and difficulties. There are, however, some goals all curators have in mind when collecting seeds for a genebank. One is to apply strategies that permit obtaining the best possible samples under specific circumstances. We believe that this goal can be achieved when the strategies are developed on the basis of genetic information furnished by molecular markers. It is worth remembering how important this information can be for conservation activities. The following points can be clarified using molecular techniques:

- the mating system
- the degree of natural inbreeding
- the level of genetic divergence among collection sites or subpopulations; area of seed and pollen dispersal; the degree of genetic relationship or coancestry between adult plants in natural conditions; the neighbourhood within which adults are genetically related

There are three basic units to be considered in seed collection:

- the number of seeds per seed parent
- the number of seed parents per subpopulation
- the number subpopulations

Molecular genetic information allows one to find the ideal combination of these units for maximizing the representativeness of the seed sample. It also helps to define where the samples should be taken.

We understand that the collection of valuable materials or of populations endangered by human activity should be supported by estimates of the parameters of population genetics obtained from molecular markers. Along with the use of dense molecular markers, which are becoming cheaper with time, the use of valuable bioinformatics visualization tools and efficient biometrical-statistical methods are indispensable tools that can assist genebank curators in making appropriate decisions about optimum sampling strategies for regeneration, collection and subpopulation structures.

Acknowledgements

This work was partly funded by CNPq (the Brazilian Council for Scientific and Technological Development). The authors would like to thank Evandro V. Tambarussi and Fernando H. Toledo for their assistance.

References

Allard RW. 1970. Population structure and sampling methods. In: Frankel OH, Bennett E, editors. Genetic Resources in Plants: Their Exploration and Conservation. Blackwell Scientific Publications, Oxford, UK. pp.97–107.

Bandel G, Gurgel JTA. 1967. Proporção do sexo em Pinheiro Brasileiro *Araucaria angustifolia* (Bert.) O. Ktze. Rev. Tec. Serv. Florestal do Est. de S. Paulo. Secr. Agr. Est. S. Paulo 6: 209–220.

Barret SCH, Yakimowsky SB, Field DL, Pickup M. 2010. Ecological genetics of sex ratios in plant populations. Philosophical Transactions of the Royal Society B. 365:2549–2557.

Bawa KS. 1974. Breeding systems of tree species of a lowland tropical community. Evolution 28(1):85–92.

Bawa KS. 1980. Evolution of dioecy in flowering plants. Annual Review of Ecology, Evolution, and Systematics 11:15–39.

Bawa KS, Perry DR, Beach JH. 1985. Reproductive biology of tropical lowland rain forest. I. Sexual systems and incompatibility mechanisms. American Journal of Botany 72(3):331–345.

Cockerham CC. 1969. Variance of gene frequency. Evolution 23:72–84.

Cockerham CC, Weir BS. 1984. Covariances of relatives stemming from a population undergoing mixed self and random mating. Biometrics 40:157–164.

Crossa J, Vencovsky R. 1994. Implications of the variance effective population size on the genetic conservation of monoecious species. Theoretical and Applied Genetics 89:936–942.

Crossa J, Vencovsky R. 1997. Variance effective population size for two-stage sampling of monoecious species. Crop Science 37:14–26.

Crossa J, Vencovsky R. 1999. Sample size and variance effective population size for genetic conservation. Plant Genetic Resources Newsletter 119:15–25.

Crossa J, Hernandez CM, Bretting P, Eberhart SA, Taba S. 1993. Statistical genetic considerations for maintaining germplasm collections. Theoretical and Applied Genetics 86:673–678.

Crow JF, Denniston C. 1988. Inbreeding and variance effective numbers. Evolution 42(3):482–495.

Crow JF, Kimura M. 1970. An Introduction to Population Genetics Theory. Burgess Publishing, Minneapolis, Minnesota.

Hernandez CM, Crossa J. 1993. A program for estimating the optimum sample size for germplasm conservation. Journal of Heredity 84:1.

Marshall DR, Brown AHD. 1975. Optimum sampling strategies in genetic conservation. In: Frankel OH, Hawkes JG, editors. Crop Genetic Resources for Today and Tomorrow. Cambridge University Press, Cambridge, UK.

Matallana G, Wendt T, Araujo DSD, Scarano FR. 2005. High abundance of dioecious plants in a tropical coastal vegetation. American Journal of Botany 92(9):1513–1519.

Moran PAP. 1950. Notes on continuous stochastic phenomena. Biometrika 37:17–33.

Namkoong G. 1986. Sampling for germplasm collections. Horticultural Science 23(1):79–81.

Oliveira PE. 1996. Dioecy in the cerrado vegetation of Central Brazil. Flora 191:235–243.

Oliveira PE, Gibbs PE. 2000. Reproductive biology of woody plants in a cerrado community of Central Brazil. Flora 195:311–329.

Queenborough SA, Burslem DFRP, Garwood NC, Valencia R. 2007. Determinants of biased sex ratios and inter-sex cost of reproduction in dioecious tropical forest trees. American Journal of Botany 94(1):67–78.

Sebbenn AM. 2006. Sistema de reprodução em espécies tropicais e suas implicações para a seleção de árvores matrizes para reflorestamentos ambientais. In: Higa AR, Silva LD, editors. Pomares de Sementes de Espécies Florestais Nativas. Fundação de Pesquisas Florestais do Paraná (FUPEF), Curitiba, Brazil. pp.93–108.

Sokal RR, Oden NL. 1978a. Spatial autocorrelation in biology. 1. Methodology. Biological Journal of the Linnean Society 10:199–228.

Sokal RR, Oden NL. 1978b. Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10:229–249.

Vencovsky R. 1978. Effective size of monoecious populations submitted to artificial selection. Brazilian Journal of Genetics 1(3):181–191.

Vencovsky R, Crossa J. 1999a. Variance effective population size under mixed self and random mating with applications to genetic conservation of species. Crop Science 39:1282–1294.

Vencovsky R, Crossa J. 1999b. Medidas de representatividad. Workshop O Melhoramento de Plantas na Virada do Milenio. Universidad Federal de Vicosa, MG, Brasil.

Vencovsky R, Crossa J. 2003. Measurements of representativeness used in genetic resource conservation and plant breeding. Crop Science 43:1912–1921. doi:10.2135/cropsci2003.1912.

Vencovsky R, Nass LL, Cordeiro CMT, Ferreira MAJF. 2007. Amostragem em recursos genéticos vegetais. In: Nass LL, editor. Recursos Genéticos Vegetais. Embrapa, Brasília.

Vencovsky R, Chavez L, Crossa J. 2011. Variance effective population size for dioecious species. Crop Science (in press).

Viegas MP, Silva CLSP, Moreira JP, Cardin LT, Azevedo VCR, Ciampi AY, Freitas MLM, Moraes MLT, Sebbenn AM. 2011. Diversidade genética e tamanho efetivo de duas populações de *Myracrodruon urundeuva* fr. all., sob conservação *ex situ*. Revista Árvore 35(4):769–779.

Weir BS. 1996. Genetic Data Analysis II - Methods for Discrete Population Genetic Data. Sinauer Associates, Sunderland, Massachusetts.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Internet resources

MLTR (Multilocos Mating System Program) by Kermit Ritlan (a computer programme for analysing marker data): http://genetics.forestry.ubc.ca/ritland/programs.html

Resources for obtaining measures of genetic divergence among collection sites or subpopulations of a species in a given ecogeographic region, and of the level of natural inbreeding:

FSTAT by Jérome Goudet: www2.unil.ch/popgen/softwares/fstat.htm

GDA (Genetic Data Analyses) by Paul O. Lewis and Dmitri Zaykin: http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php

Annex A: Two-stage sampling model for random-mating species

Theoretical formulas for different sampling schemes of monoecious species

For monoecious species, Crossa and Vencovsky (1994) derived the following $N_{e(v)}$ expressions for different alternative sampling schemes for female and male gametes:

Case 1: Plant-to-plant hand pollination is practiced and equal numbers of seeds taken from each plant (female plus male gametic control: *FGC + MGC*):

$$N_{e(v)} = \frac{n}{\frac{n(1-u)}{4P} + \frac{3n(1-v)}{4M} + \frac{1}{2}} \tag{3}$$

Case 2: Pollination is random and the same number of seeds is taken from each seed parent (female gametic control: *FGC*):

$$N_{e(v)} = \frac{n}{\frac{n(1-u)}{4P} + \frac{3n(1-v)-1}{4M} + \frac{3}{4}} \tag{4}$$

Case 3: Pollination is random and *n* seeds are randomly sampled from a bulk of seeds stemming from *P* seed parents (*RS*) and *M* pollen parents:

$$N_{e(v)} = \frac{n}{\frac{n(1-u)-1}{4P} + \frac{3n(1-v)-1}{4M} + 1} \tag{5}$$

For seed collection for random-mating species, it can be assumed that the number of pollen parents *M* is very large, $M \approx N$, $v \approx 1$, and the number of seed parents can be considered to be much smaller than *N* ($P \ll N$) and $u \approx 0$. This over-simplification is used for calculating an upper limit for $N_{e(v)}$. With these assumptions, the approximate expressions are

$$N_{e(v)} \approx \frac{n}{\frac{n}{4P} + \frac{3}{4}} \tag{6}$$

for female gametic control (FGC) and

$$N_{e(v)} \approx \frac{n}{\frac{n-1}{4P} + 1} \tag{7}$$

for random sampling (*RS*) (Crossa and Vencovsky 1994). Equation 6 can be rewritten if the average number of seeds per seed parent $\bar{n} = n/P$ is introduced, resulting in $N_{e(v)} = 4P\,(\frac{\bar{n}}{\bar{n}+3})$, $\bar{n}$ being constant over seed parents, due to *FGC*.

Annex B: Two-stage sampling model for mixed self- and random-mating species

The general equations for a fraction *u* of potential parents contributing female gametes and *v*, the fraction of potential parents contributing male gametes, are

Random sampling (*RS*)

$$N_{e(v)} = \frac{n(2-s)}{\frac{(1+s)^2[n(1-u)-1]}{2P} + \frac{(1-s)[n(1-v)(3+s)-1]}{2M} + 2} \quad (8)$$

Female gametic control (*FGC*)

$$N_{e(v)} = \frac{n(2-s)}{\frac{(1+s)^2[n(1-u)]}{P} + \frac{(1-s)[n(1-v)(3+s)-1]}{M} + (3+s)(1-s)} \quad (9)$$

The following equations were obtained for collecting *n* seeds from *P* seed parents (Vencovsky and Crossa 1999a) of a population in inbreeding equilibrium and natural rate of self-fertilization (*s*):

Random sampling (*RS*)

$$N_{e(v)} = \frac{n(2-s)}{\frac{(1+s)^2(n-1)}{2P} + 2} \quad (10)$$

Female gametic control (*FGC*)

$$N_{e(v)} = \frac{2n(2-s)}{\frac{(1+s)^2 n}{P} + (3+s)(1-s)} \quad (11)$$

A fundamental $N_{e(v)}$ expression is given by Cockerham (1969) for a group of *n* individuals or seeds differentiated from a reference population of infinite size solely by drift:

$$N_{e(v)} = \frac{0.5}{\left[\frac{n-1}{n}\right]\bar{\theta} + \frac{1+\bar{F}}{2n}} \quad (12)$$

where $\bar{\theta}$ is the average coancestry between all pairs of individuals and $\bar{F}$ is the average inbreeding coefficient. In practice, this would require using codominant genetic markers for estimating $\bar{\theta}$ and $\bar{F}$. Equation 12 can be adapted for measuring the degree of representativeness contained in a single maternal progeny. If the population is in equilibrium with respect to self- and random-mating, $\bar{\theta}$ and $\bar{F}$ can be expressed as functions of the natural rate of selfing (*s*) $\bar{\theta} = \frac{(1+s)^2}{4(2-s)}$ and $\bar{F} = \frac{s}{2-s}$ (Cockerham and Weir 1984).

Equation 12 can then be expressed as

$$N_e = \frac{2n(2 - s)}{(1 + s)^2(n - 1) + 4} \tag{13}$$

For a single maternal progeny of size *n*, equation 13 is an overestimation since, in natural populations, a certain amount of out-crossed seeds can be biparental. This increases $\bar{\theta}$ and reduces the corresponding $N_{e(v)}$ value. Equation 13 is equivalent to equations 10 or 11 when *P*=1.

Annex C: Seed collection from several subpopulations

Extending the theory developed by Cockerham (1969), Vencovsky and Crossa (2003) derived the following $N_{e(v)}$ expression:

$$N_{e(v)} = 1/(2D_1) \tag{14}$$

where $D_1 = F_{ST}\left[\frac{1+C^2}{S}\left(\frac{S^*}{S-1}\right) - \frac{1}{S-1} - \frac{1}{n}\right] + \frac{1+F_{IT}}{2n}$

with S >1 and S ≤ S* for a total of *n* seeds collected randomly and in bulk within *S* subpopulations.

When seeds are not collected in bulk but on a progeny basis, an additional sampling unit is incorporated, namely, the number of seed parents. Now the expression for estimating $N_{e(v)}$ is (Vencovsky et al. 2007)

$$N_{e(v)} = 1/(2D_2) \tag{15}$$

where $D_2 = F_{ST}\left[\frac{C_s^2+1}{S}\left(\frac{S^*}{S^*-1}\right) - \frac{1}{S^*-1} - \frac{C_m^2+1}{P}\right] + \theta_m\left[\frac{C_m^2+1}{P} - \frac{1}{n}\right] + \frac{1+F_{IT}}{2n}$

for a total of *n* seeds collected from *P* seed parents. In equation 15, θ_m is the average coancestry of offspring within progenies. C_s^2 and C_m^2 are the squared coefficients of variation of the number of seeds among subpopulations and among seed parents, respectively.

Annex D: Two-stage sampling model for dioecious species

Populations of arbitrary size with partial exclusion of parents

Random sampling (RS)

When there is no control of gametes, the variance effective population size is

$$N_{e(v)} \approx \frac{4t}{D_3} \tag{16}$$

where $D_3 = \frac{1}{r(1-r)} + \frac{(1-u)t-1}{P} + \frac{(1-v)t-1}{M}$ if it can be assumed that $2N_f - 1 \approx 2N_f$, $2N_m - 1 \approx 2N_m$ and $1/(2N_f - 1) = 1/(2N_m - 1) \approx 0$, for a total of t seeds sampled from P seed parent and M pollen parents.

Female gametic control (*FGC*)

The expression is

$$N_{e(v)} \approx \frac{4t}{D_4} \quad (17)$$

where $D_4 = \frac{1}{r(1-r)} + \frac{(1-u)t}{P} + \frac{(1-v)t-1}{M} - 1$, with the same assumptions considered for equation 16.

Female and male gametic control (*FGC* + *MGC*)

Now

$$N_{e(v)} \approx \frac{4t}{D_5} \quad (18)$$

where $D_5 = \frac{1}{r(1-r)} + \frac{(1-u)t}{P} + \frac{(1-v)t}{M} - 2$, with the same assumptions as before.

Populations of arbitrary size when including all parents

When no loss or rejection of parents has occurred and for sufficiently large parental populations, the following $N_{e(v)}$ expressions were derived by Vencovsky et al. (2011).

Random sampling (*RS*)

$$N_{e(v)} \approx \frac{4fm}{f+m} \quad (19)$$

Female gametic control (*FGC*)

$$N_{e(v)} \approx \frac{4fm}{f+m-\frac{fm}{f+m}} \quad (20)$$

Female and male gametic control (*FGC+MGC*)

$$N_{e(v)} \approx \frac{4fm}{f+m-\frac{2fm}{f+m}} \quad (21)$$

For sex ratio r =0.5, equations 19, 20 and 21 reduce to $N_{e(v)} = t$, $1.33t$ and $2.00t$, respectively.

These expressions are adequate when the number of female parents (N_f) and male parents (N_m) of the reference population are such that $2N_f - 1 \approx 2N_f$ and $2N_m - 1 \approx 2N_m$, and also that $1/(2N_f - 1) \approx 0$ and $1/(2N_m - 1) \approx 0$, as before.