# A basic sampling strategy: theory and practice

5

A.H.D. Brown[1] and D.R. Marshall[2]
[1]*CSIRO Division of Plant Industry, GPO Box 1600, Canberra, ACT 2601 Australia:* [2]*Plant Breeding Institute, University of Sydney, NSW 2006 Australia.*

## Introduction

One of the most important and difficult tasks facing plant germplasm collectors is defining the most appropriate sampling strategy for a particular species and region. It is important because the plant collector acts at the crucial interface between the genetic diversity that history has left as our endowment, and what will be conserved in collections for immediate and future use (Namkoong, 1988). In the case of rapidly eroding genetic resources, the collector finally controls what will survive for the future. Defining the strategy is often difficult because species differ in crucial ways, many plant populations have complex genetic structures and samples may be used in a variety of different ways.

Marshall and Brown (1975) considered the issue of optimal sampling strategies for use in the genetic conservation of crop plants. The focus of that paper was relatively narrow, however. It defined a strategy for populations under imminent threat of extinction, in particular traditional landraces of some major crops (notably wheat, rice and barley) in danger of replacement on a very broad scale because of the spread of modern cultivars. The strategy sought to optimize the use of the resources available for emergency action programmes to save samples of such material.

Over the last 15 years, the emphasis in the collecting of crop genetic resources, at least with respect to the major world food crops, has changed radically. In particular, 'crisis-driven', broad-scale, crop-specific programmes are no longer appropriate. Much of the material that was targeted in the mid-1970s has either been collected or lost. Rather, the collecting of crop genetic resources has entered a new phase. This involves three elements. The first is a greater emphasis on the collecting and conservation of germplasm of the wild and weedy relatives of the

major crops. The second is an increased focus on the many nationally or locally important crops used for food, fibre, medicine or fuel. Examples include the tropical and subtropical fruits of southeast Asia and Latin America, the unique high-elevation crops of the Andean region, traditional medicinal plants and the leafy vegetables used worldwide in subsistence agriculture. Few of these plants are included in scientific improvement programmes. The emphasis is on collecting for direct use. The third element is the sampling and resampling of major crop germplasm to fill specific gaps, to replace lost or poorly representative samples or to meet current specific needs of breeders, for which the material in hand is deficient. The important point is that the sampling of crop germplasm needs to be more discerning, and collectors now generally need to meet specific objectives defined in advance.

Information on the kinds and amounts of genetic variation in target species populations and its distribution in the target region is critically important in developing efficient sampling strategies. Over the last 15 years, knowledge of the genetic structure of plant populations has increased markedly (e.g. Doebley, 1989; Hamrick and Godt, 1989). This provides a more secure base on which to develop robust and relevant sampling technologies. In addition, further research has been undertaken by a number of authors, in a variety of contexts, on the theory of crop germplasm sampling (e.g. Oka, 1975; Marshall and Brown, 1981; 1983; Yonezawa, 1985; Namkoong, 1988; Chapman, 1989; Brown, 1992).

The aim of this chapter is to provide an up-to-date overview of the theory and practice of crop germplasm collecting taking account of these changes. Using simple theoretical models, the impact of varying sample size and distribution on the effectiveness of germplasm sampling will be explored. This will provide a basis for advice on how to respond, in terms of sampling strategy, to the particular practical problems that collectors face in the field.

## Theoretical background

Allard (1970) first identified clearly the critical problem facing plant explorers. He stressed that most plant species contain remarkable stores of genetic variation and consist of millions of different genotypes. Indeed, in many species each plant is genetically unique. There are several exceptions to this, even in naturally occurring populations. In particular, populations of self-pollinated or clonally propagated colonizers or crops can be depauperate in genetic variation and consist of only one or a few genotypes. Yet these species store considerable variation among populations. As a result, a plant collector 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, 1983).

***Measurement of variation***

How should plant collectors conceive of the genetic variation that they are aiming to sample? The genetic variation present in a set of populations can be described by a large array of parameters (e.g. allele and genotype frequencies, gene diversities, heterozygosity levels, disequilibrium coefficients (see Weir, 1990)). However, from the standpoint of sampling genetic resources, the basic parameter for each population is the allelic richness or the number of distinct alleles at a single locus. In practice, when an estimate of this parameter is made, it is usually the average number of alleles for a large number of marker loci after the sample is taken. This parameter is the basic one for our purposes because later users of the genetic resources can adjust the frequencies of specific desired alleles at will. Breeders might use the single copy of an allele for disease resistance irrespective of its frequency in the original population or in the sample. Thus, the allelic richness of a sample is a direct measure of its value.

***Several populations***

Sampling strategies for more than one population depend on two crucial parameters. The first is the extent of genetic divergence among populations. Marshall and Brown (1975) recognized four types of alleles: (i) common, widely distributed; (ii) common, locally distributed; (iii) rare, widely distributed; and (iv) rare, locally distributed. They argued that collecting and conserving the first class of alleles present no problem. They will almost certainly be included even in small samples from a few populations. In contrast, adequate *ex situ* conservation of the last two classes will ultimately be limited by the population sizes which we are prepared to collect and conserve, and thus by the resources available. Critical to any collecting strategy are therefore traits or alleles that are locally common. Marshall and Brown (1975) thus argue that the key indicator for optimal sampling is the number of such alleles that each population possesses, i.e. the number of alleles that attain appreciable frequencies in only one population or in a few adjacent populations. Their conservation will largely depend on the identification of populations closely adapted to specific environments and agricultural practices. This definition is similar to that of Slatkin's 'private alleles' (e.g. Slatkin and Takahata, 1985), except that only private alleles with frequencies exceeding some value (e.g. 0.05) are considered. Other things being equal, populations with a higher number of locally common alleles deserve priority.

The second parameter for optimal sampling from many populations is the variation among them in the level of genetic variation (Marshall

and Brown, 1975; Schoen and Brown, 1991). This aspect is reflected in the distribution (the range and pattern) of numbers of alleles per locus. As a general rule, populations that have higher values of diversity are genetically richer and merit larger samples.

*Number of alleles in the sample*

As mentioned above, the values of diversity parameters cannot be known ahead of sampling. Indeed, the actual number of alleles in a population is a difficult parameter to estimate because its value in a sample steadily increases with increasing sample size. However, theoretical computations based on hypothetical distributions of allele frequencies can be made. The results from such theory can be checked with those from numerical simulations of sampling from actual or conjectural examples to test the effect of strategy on allele recovery (Marshall and Brown, 1975; Yonezawa and Ichihashi, 1989).

The neutral allele theory of Kimura and Crow (1964) is the most useful for this purpose as its sampling theory has been well developed (Brown and Briggs, 1991). Thus, for example, the number of selectively neutral alleles ($k$) in a sample of $S$ random gametes, from an equilibrium population of size $N$, at a locus with mutation rate $u$, is approximately:

$$k \approx \theta \log_e[(S + \theta)/\theta] + 0.6 \quad (1)$$

where $\theta = 4Nu > 0.1$ and $S > 10$. This formula shows that the expected number of alleles in a sample increases in proportion to the logarithm of sample size. In contrast, the resources required to collect the individuals at a site (once the collector is at the site) increase in direct proportion to the sample size. Thus, there is a diminishing return (in terms of collecting new alleles) per unit of cost for increasing sample size, which becomes progressively more wasteful of resources.

*A benchmark criterion*

Given the diminishing rewards for effort expended on any one population, is there a reasonable objective that would guide the collector as to when a sufficient sample was in hand? It has previously been suggested that the objective should be to include in the sample at least one copy of 95% of the alleles that occurred in the target population at frequencies greater than 0.05 (Marshall and Brown, 1975). While the biological basis for this criterion is debatable (see Marshall (1989) for further discussion and references and Krusche and Geburek (1991) for an argument for more conservative values when dealing with forest trees), the point is that either increasing the certainty level higher than 95% or dropping the critical allele frequency below 0.05 drastically increases sample size, with only marginal gains. A sample of 59 random unrelated gametes from the population is sufficient to attain this objective. This would be assured by collecting and bulking seeds or vegetative material from 30 randomly chosen individuals in a fully outbreeding sexual species, or from 30 random genotypes in an apomictic species, or from 59 random

individuals in a self-fertilizing species. A sample of 50 individuals from each population will be considered as a benchmark. Factors that lead the collector to increase or decrease this sample size are discussed later.

## Basic sampling strategy

A full statement of a basic sampling strategy can be set out as follows.

***Number and location of sampling sites***

Before the mission begins, the collector should assemble the available information on the kinds of environments occurring in the target region and on the pattern of distribution of the target species. Based on these data, the region is then roughly divided into a limited number of areas, clearly distinct because of ecological (i.e. physiographic, edaphic, climatic), botanical, agricultural or cultural differences. The total time available can then be divided among the areas according to travelling convenience, prevalence of the targets and any perceived or known differences in genetic diversity among areas. More sites should be sampled in areas where the target species is more common, or where it is evidently more variable for conspicuous polymorphisms.

Delimiting the population, and thus the sampling site, can be problematic in wild species collecting (Chapters 3 and 6), but in crop collecting the site is usually taken as being the farmer's field or orchard, the farmer's store or the market stall. Is there a benchmark figure for the total number of such sampling sites for each species in a region? Analogous statistical and cost/benefit arguments conceivably apply at the mission level to those made above at the population level. This would suggest that a set of about 50 sites comprises a justifiable sample. However, this can only be a weak guide. Unlike the members of a single population, the potential sites in a given area could conceivably be completely different from each other and yield no redundancy upon sampling. Moving on and collecting an individual at a new site is usually preferable to collecting an additional individual at the current site. Nevertheless, it is helpful to start with a specific target of 50 sites per species per region, and vary this target up or down when clear reasons exist for doing so.

***Number of individual plants sampled at a site***

As already stated, the sample size at a site should be about 50 individuals. This figure should be increased to take account of the following factors: (i) any splitting and duplication of the sample that will take place (clearly, if only two seeds are collected from an individual, a three-way splitting of the sample will leave one subsample short); (ii) any suspicion that seeds from some individuals in the sample are not fully viable; and (iii) possible loss of some individuals in the sample in transport and quarantine. The aim is that each sample at the time of entry into each

of the gene banks conserving the material should trace back to at least 50 original individuals. If it is not possible to collect from 50 individuals at each site because populations are small or shattering has already started, more sites should be sampled.

*Choice of individuals*

A much discussed question in sampling technique for crop genetic resources is whether individuals should be taken strictly at random or biased as to phenotype or microsite (e.g. Marshall and Brown, 1975, 1981; Porceddu and Damania, 1992). Random sampling is generally the most reliable and desirable method, particularly for crop populations or market samples because subpopulation structure is unlikely to be present. In contrast, natural populations of wild species often evolve local subpopulation structure and hence stratified random sampling is appropriate for them (e.g. separate random sampling of individuals in different microsites). In the case of a crop field which consists of a mechanical mixture of species (e.g. durum and bread wheat), clearly a separate random sample should be made from each of the components (see below, 'Mixed populations'). Biased sampling of rare phenotypic variants in a population is to be avoided, except when such plants clearly merit separate and distinct recognition (e.g. a rare disease-free individual in a heavily diseased field). Such samples may be taken in addition to, but not instead of, population samples, and should receive separate collecting numbers.

Strict randomness of sampling requires that every plant at the site have an independent and equally likely chance of inclusion in the sample. Ward (1974), for example, has described a simple way in which two people may collect such a sample. This is often impossible or impractical to achieve, however (Marshall and Brown, 1983). In practice, collectors usually sample at systematic or random intervals along a number of transects. Systematic sampling is easiest and spreads the sample over the population but can be biased if variation is periodic in the population (Brown and Briggs, 1991). The starting-points and directions of the transects and the position of sampling points along them can be chosen according to a randomization procedure. It will often be advantageous to keep a minimum distance between sampling points to avoid excessive sampling of closely related individuals and repeated sampling of clones (Chapters 21, 22 and 23).

*Number and type of propagules per plant*

The final decision for the plant collector concerns the kind(s) and amount(s) of material to be collected from each plant chosen for sampling. The material ranges from pollen, seeds and vegetative cuttings or propagules (such as tubers, bulbs or corms) to whole individuals (Brown and Briggs, 1991). Collecting seeds differs from collecting vegetative material in a number of ways. In particular, it is more restrictive for the timing of the mission. Otherwise, seed, being the organ of dispersal and

storage, enjoys a number of advantages, for example less bulk and easier handling and storage. Pollen is even less bulky, but its storage is difficult and it is at the moment not possible to recover plants from pollen. Cuttings are often the most convenient vegetative samples to make, but they may require the use of *in vitro* techniques. In perennials, the removal of whole individuals should be avoided, especially when to do so would destroy the source population.

The genotypes in a sample of vegetative material are exactly those of the sampled parent plants, whereas those in a seed sample depend on the breeding and pollination system. With uniparental reproduction (self-fertilization or agamospermy), seeds will closely resemble the parent plant. In contrast, the seeds of an outbreeder will differ from the parent and show within-family diversity (Yonezawa and Ichihashi, 1989). In outbreeding species, seeds from several fruits should be gathered from each individual sampled, if possible, rather than from a single fruit, to increase the diversity of genes in the sample. Also, fruits should be collected from all parts of the crown of trees, because these may have been pollinated by different pollen sources. Similar numbers of seeds should be collected from each individual sampled.

The need for sufficient material, rather than genetic principles, largely determines the number of propagules per plant to sample. If available, enough seeds or cuttings should be sampled to provide for the division of the samples among collaborators and other recipients and avoid immediate multiplications. Some gene banks have lower limits on the number of seeds per sample that will be accepted for storage, for example if multiplication of the material is not likely to be prohibitively expensive. On the other hand, quarantine facilities or the space available in the collecting vehicle or in the field genebanks (for root and tuber crops and species with recalcitrant seeds, for example) may limit the total volume of material that can be collected. When this is the case, it would be better to maximize the number of plants sampled and reduce the number of propagules per plant.

## Modifications to basic strategy for different species

The basic sampling strategy developed above thus consists of the following four elements:

- sample about 50 populations in an ecogeographic area or mission;
- sample about 50 individual plants in each population;
- in general, sample individuals at random at each site, but sample separately within distinct local microenvironments if the habitat at the site is heterogeneous;
- sample sufficient seeds or vegetative material per plant to assure representation of each original plant in all duplicates.

Species – both cultivated and wild – differ in a number of life-history,

ecological and genetic attributes that will require amendment of this basic strategy. We now consider the more important of these attributes individually and their effect upon sampling in practice. Each point will assume a comparison of species differing only for the attribute in question, while other features are held in common.

***Distribution***

The spatial occurrence of a species will often play a major role in limiting or expanding the options available to collectors.

Geographic range

A species found only in a narrow geographic range would merit sampling from fewer sites, but with an increased number of individuals at each site, and an increased number of propagules per individual.

Local abundance

Likewise, for species (particularly wild species) that are locally rare, it may be very time-consuming or impossible to meet a target of 50 individuals per population. Compensation for limited numbers locally can be made by increasing the number of sites and by increasing the number of propagules per plant. Clearly, the problem of choosing which individuals to sample from the populations of species that are locally rare may not arise. Brown and Briggs (1991) give guidelines for the collecting of endangered species of ten individuals at each of up to five sites.

Interpopulation migration

Migration rates are likely to differ among species, whether the agents of migration are natural (wild species), inadvertently human (weeds) or deliberately human (crops). When migration rates appear to be high, populations are more likely to share most of their genes and less likely to diverge. Hence the sampling of fewer but more widely spread populations is appropriate.

Habitat diversity

Species that grow in a wide range of ecological situations are more likely to have diverged genetically among their different habitats (ecotypification). In such species, an increased number of populations or distinct subpopulations should be sampled at the expense of the number of individuals per population.

***Life history***

Among plant species, life-history traits are in part correlated with features of their distribution (see above) and genetic system (see below). For example, perennial fruit trees are commonly outbreeding whereas selfing is common in many annual crops. Differences in life-history traits lead to little modification of the first two elements of the basic strategy because they do not appear to have a clear-cut effect on the disposition

of genetic variation within and among populations. For instance, outbreeding annuals and outbreeding perennials tend to have similar levels of genetic divergence among populations (Hamrick and Godt, 1989). However, when considered on their own, several life-history traits affect sampling procedures by determining the kind and amount of collectable material.

Duration of life cycle

The collecting of perennial species may be less dependent on seasonal timing of the mission as vegetative material is available for collecting throughout the year. It may be possible to arrange return to a site should the collected material prove not viable or insufficient. Therefore, the number of propagules from a site need not be as high as for an annual species.

Population age structure

Populations of perennial species can either consist of individuals of the same age (e.g. orchards, plantations) or possess an age structure (as in most natural populations). In the case of age-structured populations, individuals should be sampled at random irrespective of size or age to maximize genetic diversity, because different cohorts may be divergent genetically. If distinct age cohorts are obvious, a stratified random sample can be made.

Vegetative reproduction

Species that produce organs of vegetative reproduction, or cuttings of which are viable, add further collecting options, as discussed above. Where both seeds and vegetative material are available, it is advisable to include samples of both kinds, particularly when the species is poorly known or when quarantine procedures may have uncertain outcomes. How to label such samples with appropriate collecting numbers is discussed in Chapter 19.

Fecundity

Clearly, fecundity has direct effects on the fourth element of the basic strategy, namely the number of propagules per plant. The sampling pattern may need adjustment in the case of species that produce few seeds per individual. Sampling from more individuals at a site could compensate for scarcity of seeds. Biasing the sample towards the most fecund individuals should be avoided.

Determinacy of flowering and seed maturation

Species with highly synchronized flowering (and hence maturity of fruit) require well-timed missions. On the other hand, indeterminate flowering may mean that only a portion of the population is ready for sampling at the time of the visit. As such variation may be due to genetic variation in response to photoperiod, it is important to sample from plants with

different maturities. Thus, variation in maturity would affect the choice of individuals for the sample and whether other types of propagules than seeds should be included. Some wild species (e.g. perennial *Glycine* species) combine a determinate chasmogamous flowering habit with a relatively indeterminate fruit production from cleistogamous flowers. This adds to the flexibility of sampling.

### *Genetic system*

Various aspects of the genetic system determine the apportionment of genetic variation among and within populations and therefore substantially affect sampling strategy.

#### Mating system

Differences among species in mating system and variation within the species profoundly influence all four elements of the basic strategy. In sampling outbreeding species, the number of populations in an ecogeographic region can be reduced and the number of individuals per site increased without great loss of efficiency. The spatial scale of local differentiation in outbreeders is likely to be larger than for autogamous species, and the collecting of seeds will in itself lead to the sampling of dispersed sources of pollen. Hence there is less reason for locally stratified sampling. Open-pollinated progeny arrays are likely to be genetically variable and include some inbred seeds. Seed viability and seedling vigour could vary and the sample size per plant should allow for this.

Under self-fertilization or apomixis, populations diverge for both the alleles they contain and the amount of genetic polymorphism. Hence it is important to sample a large number of populations even at the expense of the average number of individuals at a site. In addition, with selfing species, it is important to be on the lookout for populations that have an exceptionally large amount of polymorphism and to increase the sample size of these. Natural populations of autogamous species can possess local subpopulation structure that justifies stratified random sampling at collecting sites. It is important to realize that the seeds sampled from a plant of a species with a predominantly uniparental mating system are very similar to one another genetically. Hence the total sample size should not be inflated with large numbers of seeds from few original plants.

#### Pollination mode

The mode of pollination, in particular whether pollination is by animals or by wind, affects the genetic structure of progeny arrays from a single fruit. Under wind pollination, such arrays tend to be half-sibs, as they are the products of many sources of pollen. Under animal pollination, the array of seeds from a single fruit in many species largely has the same male parent. This implies that the sample from each individual in animal-pollinated species should include seeds from several randomly chosen

fruits. In addition, animal-pollinated outbreeding species tend to show more population divergence than do wind-pollinated outbreeders but less than is shown by selfers (Hamrick and Godt, 1989). As already noted, population divergence justifies an emphasis on the sampling of more populations.

Conspicuous polymorphism

Populations of some species can show a range of levels of morphological polymorphisms. As already noted, when some populations of a species appear to be much more polymorphic than others, it is sensible to increase the sample size in the richer populations. This is especially important in inbreeding species, where the range in the level of polymorphism is likely to be wider than in outbreeders and the link between genetic variation in marker genes and genetic variation throughout the whole genome is stronger.

***Mixed populations***

Fields of landraces may sometimes consist of mechanical mixtures, deliberately composed by the farmer. Examples range from a field with very different species, like wheat and barley or durum and bread wheat, or with differently maturing strains of a vegetable crop, to mixtures of clones of noble sugarcane each diagnostically marked by its own stalk colour pattern. If possible, the collector should seek from the farmer samples from the original seed lots of each component of such mixtures. Otherwise, the collector must decide between two options. Option 1 is to make a separate sample of each component in the field. Option 2 is to regard the field as a single population and therefore make a single random sample. Option 1 is clearly the appropriate one when two crop species are mixed, whereas option 2 is more appropriate if the components may have interbred to produce the seeds to be sampled, as for mixed races of maize or of an outbreeding vegetable. For cases intermediate between these two extremes, collectors should take a single random sample, because they can rely on the general statistical robustness and simplicity of such samples and avoid giving an extreme bias to any rare type. More guidance on this issue is given in Chapter 18, where the importance of the farmers' knowledge in deciding how to collect is stressed.

## Modifications to the basic strategy when sampling for specific goals

Sometimes, a collecting mission is undertaken to meet specific goals, such as the collecting of further genes for resistance to a particular disease from a region where the disease is known to occur. When collecting objectives are specific, how does the sampling strategy differ from that appropriate for generalized collecting?

*The rare variant*

The first point of departure is the treatment of rare alleles, as they may be more important than in the case of the sampling theory for generalized collecting. A rare allele might be the basis of the desired phenotype. Alternatively, a population that would meet the explicit objective of the mission might occur at only one kind of site (e.g. a waterlogged or saline habitat), which could be rare in the target region. In such cases, more individuals per site, or more sites, should be sampled.

Consider the case when the nominated target is coded by a specific allele. Let us assume the frequency of this allele in the populations is $p$. The size of sample in terms of the number of random gametes ($S$) required to be 95% certain of including at least one copy of the target allele or genotype is:

$$S \approx -3/\log_e[1-p] \tag{2}$$

From this formula, the sample sizes required for an allele or character that is increasingly rare ($p = 0.05, 0.03, 0.01, 0.001$) are 59, 99, 299 and 2999. The size increases at an exponential rate as the allele is progressively rarer. The same computations apply for the number of sites if $p$ represents the frequency of the desired site and $S$ the required number of sites to sample to be 95% certain of a successful mission.

Thus, the collector may well decide to increase substantially the sample size at a site that evidently meets a specific objective of the mission, especially if the desired variants are likely to be rare and restricted to that site.

*More than one copy*

The second major point of departure in sampling for specific goals compared with generalized collecting is the adequacy of a strategy that assures just one copy of the desired variant. If the mission is seeking certain genetic variants, it may need more than just one copy of a desired allele. The sampling should provide the allele in sufficient numbers to guard against its later loss, and provide it in a variety of genetic backgrounds. The size required to be assured of a specified number of copies of an allele can be calculated using the relevant sampling theory as follows.

Sedcole (1977) has computed the sample size ($S$) required to be 95% certain of recovering a minimum number ($r$) of plants with a trait that occurs in a population with frequency $p$. Table 5.1 lists a selection of these values of $S$. An approximate formula for $S$ for relatively infrequent traits ($p < 0.20$) is:

$$S \approx \{r + 1.645\sqrt{r} + 0.5\}/p \tag{3}$$

The values in Table 5.1 show that the required sample size increases with increasing required number of copies, as would be expected. However, the increase is less than proportionate. Thus, increasing the sample size

**Table 5.1.** The sample size ($S$) required to be 95% certain to recover a minimum number ($r$) of plants with a trait that occurs in a population with frequency $p$. (From Sedcole, 1977.)

| | Number of copies be recovered ($r$) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 | 15 |
| 0.25 | 11 | 18 | 23 | 29 | 34 | 40 | 50 | 60 | 84 |
| 0.125 | 23 | 37 | 49 | 60 | 71 | 82 | 103 | 123 | 172 |
| 0.0625 | 47 | 75 | 99 | 122 | 144 | 166 | 208 | 248 | 347 |
| 0.03125 | 95 | 150 | 200 | 246 | 291 | 334 | 418 | 500 | 697 |
| 0.015625 | 191 | 302 | 401 | 494 | 584 | 671 | 839 | 1002 | 1397 |
| 0.05 (from formula (3)) | 63 | 97 | 127 | 156 | 184 | 211 | 264 | 314 | 438 |

fivefold (for $p = 0.05$, from 63 to 314) increases the assured number of copies tenfold.

A relatively large sample is thus justified in missions that are aimed at finding several examples of specific rare and valuable variants. The final figure will depend on the practical limits on handling larger samples weighed against the number of copies that are actually needed.

***Representativeness***

Rather than simply being 95% certain of the presence of a given number of copies of common alleles in the sample, the collector may want the sample to accurately reflect the frequencies of alleles in the source population. This may be the case when the material is for direct use as an adapted population, rather than for indirect use as the source for alleles to be incorporated in a breeding programme. Marshall and Brown (1983) suggest that under these conditions sample size should be considerably larger, about 200 individuals. Whereas for inbreeders the number of seeds collected per individual has little effect on the fidelity of the sample, for outbreeders increasing the number of seeds per plant reduces the variance of sample allele frequencies, though with diminishing returns.

## Modifications to the basic strategy when resampling a region

The 'basic sampling strategy' (see above) was devised for the situation where little or no detailed information is available on the genetic structure of the target populations. This may be because no previous samples have been taken from the region, or no reports of the nature of its populations have been published. However, in many cases, partial information is at hand about the ecological distribution of the target species and its

genetic variation from previous sampling. The sources of information, which are discussed in other chapters, include gene-bank databases, Floras[1] and monographs, the route maps of previous collectors and their herbarium specimens with associated field records.

A specific example of when resampling of a region is particularly justified is the replacement of samples in a gene bank that have become degraded. Resampling is also appropriate in populations that are undergoing temporal coevolutionary changes in response to changes in cropping systems or in the prevalence or genetic structure of other interacting organisms such as predators, pathogens and weeds. Monitoring genetic erosion will also require repeated sampling.

Several workers have noted that sampling in two or more episodes can be a very efficient strategy, though it is clearly not always feasible. For example, Jain (1975) advocated a two-step sequential method with the first collecting season on a coarse grid and a second round on a finer scale in areas of particular interest. However, just how the collector is to modify the sampling strategy in the light of previous sampling is open to discussion.

In the case of resampling, the general strategic questions facing the collector (modified from Nabhan, 1990) are as follows:

- Should the collector give particular emphasis to uncollected or largely unstudied areas within the region? The risk of doing this is that the species may be rare or absent from such areas. The benefit is that any new samples could add new ecogeographic representation to the collection.
- Should the collector emphasize areas where the previous data and samples indicate high genetic diversity? The risk here is that the new samples may contain much that is redundant with the accessions already in gene banks. The benefit is that the collector will be assured of getting a very rich collection.
- Should the collector aim for specific ecotypes or landraces reported to exist but no longer available from gene banks? The disadvantage of this strategy is that such specific localities are often very scattered and the material may have subsequently disappeared from the site. For crop populations in particular, the field situation is likely to be relatively labile. The advantage is that the specific-target strategy has a good chance of locating at least some of the targets.

Nabhan (1990) has given a detailed case-study of how to use the data from previous explorations in devising a strategy for the sampling of wild *Phaseolus* species in northern Mexico. In this example, the primary data are herbarium records of species occurrence in various geographic subregions. Graphs of species richness against the number of samples from each subregion indicate that the subregions that may be under-

[1] See footnote p. 96.

collected and would repay further visits. This amounts to making relatively more samples in the areas of higher diversity. Chapter 15 discusses areographic methods such as the ones used by Nabhan (1990) in more detail. But what should be the degree of bias or weighting towards more diverse areas in sampling?

From the standpoint of genetic conservation, the objective of a resampling project is to obtain a maximum amount of new genetic diversity additional to that already available in gene banks. Specific guidance as to the appropriate weighting of effort among the various subregions can be sought from the sampling theory of selectively neutral alleles in finite populations in terms of its basic parameter $\theta(=4 \times$ effective population size $\times$ mutation rate). The question is: What is the optimum allocation of a fixed total sampling effort among several independent populations with varying levels of genetic diversity? It can be shown that the number of alleles expected in the total sample is a maximum when the total sampling effort is divided among the populations in direct proportion to the value of $\theta$ in each population (D.J. Schoen and A.H.D. Brown, in prep.). The value of $\theta$ should be the average over many loci, the same loci being tested in all the populations.

How is it possible to estimate the values of $\theta$ in practice? If no genetic data are available and identity by descent does not vary greatly among populations, then an approximate relative estimate of $\theta$ is the population size. Simply put, this implies taking samples from each subregion in proportion to the commonness of the species in that subregion. If data on genetic polymorphism are available on a comparable set of loci, the estimates of $\theta$ for each locus are obtained by the formula $\theta = h/(1 - h)$, where $h$ is the gene diversity or probability that two gametes sampled at random from the population (subregion) differ at the locus (Weir, 1990). The number of samples from a subregion is then taken in proportion to the average value of $\theta$ in that subregion. Finally, if comparative estimates of genetic variance are available for quantitative characters, the number of samples from each subregion should be in proportion to its genetic variance. This follows from the theory that links such variance with the parameter $\theta$.

## Conclusions

Sampling is a critical step in the conservation of plant genetic resources. Therefore, the onus is on the collector to obtain the richest collection for a given expenditure of effort. This is ensured not by collecting large quantities of material, but by a judicious division of the target region into ecogeographical areas and collecting a limited number (about 50) of random individuals from many populations in each area, up to a total of about 50 populations for the mission. These guidelines can readily be adapted to take account of biological differences among species. If the mission has specific goals, it will be appropriate to increase the sample

size when the material at a locality appears to meet (or is likely to meet) one of these goals. In the case of resampling, evidence as to genetic divergence among populations can be used in the division of the region into relatively homogeneous areas, and evidence of genetic richness can be used to adjust sample size in terms of the number of samples for each area. In this way, the samples will recover a high percentage of the alleles on offer.

## References

Allard, R.W. (1970) Population structure and sampling methods. In: Frankel, O.H. and E. Bennett (eds) *Genetic Resources in Plants - Their Exploration and Conservation*. pp. 97–107. Blackwell Scientific Publications, Oxford.

Brown, A.H.D. (1992) Human impact on plant gene pools and sampling for their conservation. *Oikos* 63:109–118.

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

Chapman, G.C. (1989) Collection strategies for the wild relatives of field crops. In: Brown, A.H.D., O.H. Frankel, D.R. Marshall and J.T. Williams (eds) *The Use of Plant Genetic Resources*. pp. 263–279. Cambridge University Press, Cambridge.

Doebley, J.W. (1989) Isozymic evidence and the evolution of crop plants. In: Soltis, D.E. and P.S. Soltis (eds) *Isozymes in Plant Biology*. pp. 165–191. Dioscorides Press, Portland.

Hamrick, J.L. and M.J. Godt (1989) Allozyme diversity in plant species. In: Brown, A.H.D., M.T. Clegg, A.L. Kahler and B.S. Weir (eds) *Plant Population Genetics, Breeding and Genetic Resources*. pp. 43–63. Sinauer Associates, Sunderland.

Jain, S.K. (1975) Population structure and the effects of breeding system. In: Frankel, O.H. and J.G. Hawkes (eds) *Crop Genetic Resources for Today and Tomorrow*. pp. 15–36. Cambridge University Press, Cambridge.

Kimura, M. and J.F. Crow (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.

Krusche, D. and Th. Geburek (1991) Conservation of forest gene resources as related to sample size. *Forest Ecology and Management* 40:145–150.

Marshall, D.R. (1989) Crop genetic resources - current and emerging issues. In: Brown, A.H.D., M.T. Clegg, A.L. Kahler and B.S. Weir (eds) *Plant Population Genetics, Breeding and Genetic Resources*. pp. 370–391. Sinauer, Sunderland.

Marshall, D.R. and A.H.D. Brown (1975) Optimum sampling strategies in genetic conservation. In: Frankel, O.H. and J.G. Hawkes (eds) *Crop Genetic Resources for Today and Tomorrow*. pp. 53–80. Cambridge University Press, Cambridge.

Marshall, D.R. and A.H.D. Brown (1981) Wheat genetic resources. In: Evans, L.T. and W.J. Peacock (eds) *Wheat Science - Today and Tomorrow*. pp. 21–40. Cambridge University Press, Cambridge.

Marshall, D.R. and A.H.D. Brown (1983) Theory of forage plant collection. In: McIvor, J.G. and R.A. Bray (eds) *Genetic Resources of Forage Plants*. pp. 135–148. CSIRO, Melbourne.

Nabhan, G.P. (1990) *Wild* Phaseolus *Ecogeography in the Sierra Madre Occidental,*

*Mexico*. Systematic and Ecogeographic Studies on Crop Genepools 5. IBPGR, Rome.

Namkoong, G. (1988) Sampling for germplasm collections. *HortScience* 23:79–81.

Oka, H.I.(1975) Consideration on the population size necessary for conservation of crop germplasm. In: Matsuo, T. (ed.) *Gene Conservation - Exploration, Collection, Preservation and Utilization of Genetic Resources*. pp. 57–84. JIBP Synthesis Volume 5. University of Tokyo Press, Tokyo.

Porceddu, E. and A.B. Damania (1992) *Sampling Strategies for Conserving Variability of Genetic Resources in Seed Crops*. Technical Manual No. 17. ICARDA, Aleppo.

Schoen, D.J. and A.H.D. Brown (1991) Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proceedings of the National Academy of Sciences* 88:4494–4497.

Sedcole, J.R. (1977) Number of plants necessary to recover a trait. *Crop Science* 17:667–668.

Slatkin, M. and N. Takahata (1985) The average frequency of private alleles in a partially isolated population. *Theoretical Population Biology* 28:314–331.

Ward, D.B. (1974) The 'ignorant man' technique of sampling plant populations. *Taxon* 23:325–330.

Weir, B.S. (1990) *Genetic Data Analysis*. Sinauer Associates, Sunderland.

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

Yonezawa, K. and H. Ichihashi (1989) Sample size for collecting germplasm from natural plant populations in view of the genotypic multiplicity of seed embryos borne on a single plant. *Euphytica* 41:91–97.