Saving Seeds: Optimally Planning Our *Ex Situ* Conservation Collections to Ensure Species' Evolutionary Potential[1]

Sean M. Hoban[2,3]

In the face of ongoing environmental change, conservation and natural resource agencies are initiating or expanding *ex situ* seed collections from natural plant populations. Seed collections have many uses, including in provenance trials, breeding programs, seed orchards, gene banks for long-term conservation (live plants or seeds), restoration, reforestation, and scientific study of plant germination or other plant ecology studies. Well-known examples of *ex situ* collections include the Millennium Seed Bank Partnership, Australian Seed Bank Partnership, United Kingdom National Tree Seed Program, United States National Plant Germplasm System, and South African Regional Seed Bank. Some collections focus on rare species, species with relevance to agriculture or forestry, or regional flora. Other collections are in response to immediate threats, such as damaging insects and pathogens (e.g., emerald ash borer). In this talk I will discuss how to sample seeds to most optimally conserve the evolutionary potential of a species to ensure its long-term survival.

A useful seed collection captures as much phenotypic and genetic diversity from natural populations as possible. Choices for a collector include how many populations, maternal plants, and seeds per plant to collect. A collector wishes to achieve efficiency—to not waste limited time, resources, personnel, and storage space, but also to achieve effectiveness—to be as complete as possible in case important genetic variants are lost from natural populations.

In a series of papers starting in 1975, Brown and Marshall (1975) proposed some solutions to this general sampling problem. They used simple mathematical equations to derive a minimum number of samples needed to capture allelic variants that occur in a population at a given minimum frequency. For an arbitrary minimum frequency of 0.05, the recommended minimum sample size was 30 individuals from a fully outcrossing species or 59 individuals from a fully self-pollinating species. A 'rule of thumb' of 50 samples was suggested for practical use. This 50 sample guideline has since been integrated into many protocols for sampling seed, and is still common 4 decades later. Hoban and Strand (2015) found this guideline cited in 60 percent of protocols from major seed collecting or natural resource management organizations.

Though widely used, these guidelines may be suboptimal for genetic representation for several reasons. Principally, this approach assumes that there is no genetic structure in a population. Specifically, the assumption is that all populations have completely different sets of alleles present – no sharing of alleles among locations. An alternative approach, that of Lawrence et al. (1995), assumes all populations have the exact same allele frequencies. The corollary assumption is an absence of spatial genetic structure such as clines, barriers, or disjunctions. Most real species, however, have a substantial amount of shared alleles among populations, and feature spatial patterns such as a decline in genetic similarity between populations due to geographic or environmental distance (Sork and Smouse 2006).

The Brown and Marshall approach also assumes that every seed chosen within a population will contain a random sample of the population's genetic variation; thus, every seed has an equal chance to add new variation to the sample. In reality, however, sampling many seeds from a maternal plant does not

---

[1] A version of this paper was presented at the Gene Conservation of Tree Species – Banking on the Future Workshop, May 16-19, 2016, Chicago, IL.

[2] National Institute for Mathematical and Biological Synthesis, 1122 Volunteer Blvd, Suite 110, University of Tennessee, Knoxville, TN 37916.

[3] The Morton Arboretum, 4100 Illinois Route 53, Lisle, IL 60532.

Corresponding author: shoban@mortonarb.org.

add significant variation but rather largely duplicates variation because factors such as limited pollen dispersal and dominance of nearby fathers result in typically small paternal gene pools.

It has not yet been evaluated whether the Brown and Marshall assumptions are too simplifying, whether they truly capture the expected amount of genetic variation in the seed sample for real world species. Some authors have proposed that collection protocols should be based on a species' biological characteristics, such as population size, rarity, degree of self-pollination, and level of habitat fragmentation (CPC 1991, Guerrant et al. 2014, Hoban et al. 2015). As yet, there is no quantitative advice for how to tailor a collection protocol to a species' traits. Plant traits are incorporated into conservation tasks such as quantitative assessments of invasion or extinction risk, however (Bland et al. 2015. Murray et al. 2014), suggesting that plant traits could be incorporated, quantitatively, into collection protocols.

With a group of colleagues, I have been using a new approach to designing *ex situ* seed collections. It is based on population genetic simulations that incorporate plant traits. This approach is similar to the method of Bataillon et al. (1996), who used simulations to sample from existing seedbanks to create a 'core collection' that would contain most of the genetic variability of the whole collection. In my approach, simulations are used to create in silico datasets representing the species or set of populations of interest. These datasets are then sampled from with different potential sampling protocols. Then the genetic diversity captured in each protocol can be compared, and the protocol capturing the most variation in the most efficient manner can be chosen. Throughout, this work assumes that the goal is to calculate a minimum sample size- the fewest number of seeds or samples to capture a given amount of genetic variation. Note that this approach assumes that all seeds are viable and all will grow to a reproductively mature individual, i.e., no attrition or loss during seed storage and use. It has been shown that attrition can vary from a few percent to nearly 95 percent (Cochrane et al. 2007; Hoban and Way, unpublished data). In addition, loss of genetic variation can occur through multiple cycles of seed increase (sowing the collected seed to then gather larger amounts of seed) and other processes (Basey et al. 2015). Minimum sample sizes such as those discussed here should be adjusted upward to allow for expected losses.

In a series of studies, colleagues and I used simulations and real datasets to determine the influence of several factors on the amount of genetic variation captured: different plant traits, degree of range wide population structure, varied local sampling patterns (transect, random, etc.), and size of the paternal pollen pool. I also calculated the return on investment (ROI, defined as new alleles captured per sampling effort) of sampling from a new plant, or sampling more seeds from plants that were already sampled.

In one simulation experiment, we quantified the influence of dispersal kernel, self-pollination rate, and life history (annual or perennial habit) on genetic capture in a collection. This experiment demonstrated that each of these biological factors does influence genetic variation captured, and that simulations can be used to calculate how minimum sample sizes should differ in plants with different traits (Hoban and Strand 2015). For example, a species with high selfing rates and low dispersal required seed lots up to five times as large (about 40 seeds each from 40 plants, or 1600 total seeds) as species without those characteristics (about 10 seeds each from about 30 plants, or 300 seeds), to achieve a reasonable goal of conserving 90 percent of the rare alleles. I confirmed this general trend by sampling from existing datasets of three tree or shrub species with different traits, and found drastically different allelic capture in each species for a given sampling effort. The results of this experiment underscore that the same sampling strategy applied to different species will capture different amounts of genetic diversity, for each species.

Another characteristic that influences the amount of variation captured by sampling is the level of population fragmentation, which can range from well-connected to scattered populations. We simulated one common model of range wide population structure, the ecoregional model, in which migration among ecoregions is lower than within them. We compared two differing approaches to sampling- sampling one population from each of four regions (dispersed) and sampling four populations all in one region (constrained). This is the same number of populations but different spatial coverage. Our work showed (as predicted by theory) that it is more important to sample from each ecoregion for less well-connected populations than it is for more well-connected populations. Also, the amount of variation captured when sampling in a constrained fashion was as little as half the variation captured by dispersed sampling, depending on the type of allele of interest (Hoban and Schlarbaum 2014). This information is important

because sampling may be constrained by logistical, financial or political borders. In such cases, the genetic variation collected may be less than desired.

We also showed that population structure had a previously unrecognized effect on sampling: the minimum number of samples recommended by Brown and Marshall (50 sampled plants) may sometimes be a conservative number when sampling many populations in a many-population system. This is due to allele sharing among populations; in real populations many alleles occur in multiple populations, albeit at different (sometimes low, sometimes high) frequencies. Thus, sampling multiple populations will increase the cumulative probability an allele will be captured from at least one population. Therefore in a real system of populations, the Brown and Marshall suggestion of 50 samples per population will capture more than the amount of genetic diversity predicted by their models. In one situation we examined, sampling just 25 samples per population from many populations in a many-population system captured 97 percent of all alleles, while sampling 50 samples captured 99 percent. Sampling from many populations in a well-connected system clearly gives multiple chances to capture alleles and may allow lower minimum sample numbers per population.

We then showed that the better choice is almost always to collect from a new plant rather than to collect more seeds from the plants already visited. Indeed the return on collecting from a new plant is relatively constant across collections sizes, while the return rapidly drops to near zero when sampling more seeds on plants already visited. This is because all the maternal alleles will be collected by a small sample (half of each seed is maternal genetic material) and a small number of fathers that typically contributes to the seed set on a given plant (in trees, often less than 20 fathers, e.g., Grivet et al. 2009). Hoban and Schlarbaum (2014) found that, in several situations examined, allelic capture appears to begin to plateau after 16 seeds are taken per maternal plant, though more situations need to be examined. Hoban et al. (unpublished data) showed that sampling 1000 seeds from a plant may capture negligibly more genetic variation than 200.

Next, we examined how to sample spatially at local scales within populations. I tested this because collectors often sample opportunistically, at accessible trails or roadside patches, or by straight line transects. It would be expected that a random or systematic approach covering equal parts of the population would be best, while covering small areas would capture less genetic variation, but it has never been quantified how each strategy performs relative to the others. We showed that random and grid sampling are statistically indistinguishable, that opportunistic sampling captures about half the variation of random sampling, and that sampling a transect produced results intermediate between restricted and random sampling (Hoban and Strand 2015). With this information, a sampler can decide whether it is worth the extra effort to cover more ground.

There are interesting future directions for this type of work. Other aspects of plant biology may also be expected to influence gene distribution on the landscape and thus the effectiveness of different sampling strategies. Population density, animal vs. wind pollination, and clonality are three such aspects (Vekemans and Hardy 2004). The location of seeds within the canopy of a single plant, especially large trees, may also be important because different parts of the canopy will receive pollen from different sources. Another unexplored aspect of the distribution of genetic diversity in plant populations is the influence of location within the species' range, e.g., edge vs. center (see Gapare et al. 2008). Lastly, population history likely influences the amount and type of variation available, and thus how to sample optimally. It will be important to determine how to sample species subject to recent colonization or bottlenecks, or large disjunctions.

Of course, the degree to which these aspects can be incorporated into sampling strategies will depend on the amount of information available for a species. If little information is available, desk studies or surveys of a species may help gain information and develop an effective strategy. For an emergency collection, the Brown and Marshall guideline remains a starting point, though for a species or population under immediate threat such as destruction from development, collecting all available seed has been recommended.

The simulation approach can also be applied to real species' collections. Thus far, my investigations have been into species archetypes or broad categories as defined by traits. I can also build demographic

genetic simulations tailored to particular taxa. Along with colleagues at the United Kingdom National Tree Seed Project, I have been working to test sampling protocols for European ash (*Fraxinus excelsior*) in the United Kingdom. We will estimate how much genetic variation has been captured in the samples taken so far (three seasons of seed sampling), where to sample next, what is the relative benefit of sampling on the edge vs. the center and in the north vs. the south, and what are optimal numbers of plants per population and seeds per plant to collect. This simulation experiment will use a realistic demographic model based on observed tree abundances at fine scale (10 km grid), and a genetic model fine-tuned to produce F_{ST}s similar to those observed in a recent microsatellite study of this species.

In summary, I explained and demonstrated a new approach to optimize sampling protocols for a conservation seed collection. I used spatial, demographic and genetic data from empirical and simulated data under an individual-based model, to lead to tailored collections that maximize diversity while minimizing collection size when a collector knows some basic biological characteristics of the species. I showed that characteristics of plant reproduction and dispersal, as well as logistical factors, significantly influence the genetic diversity captured in seed collections. As one example, a highly self-pollinating, low dispersal species likely needs sample sizes five times larger than current guidelines. Results show that minimum collection protocols should be customized for the target species, rather than commonly implemented "rules of thumb" like 50 samples. There is not a single minimum sample number suitable for all species, but we can derive minimum numbers for archetypes defined by plant traits. Overall, my work shows that it is possible to improve the value of seed collections by quantitatively integrating current knowledge of plant biology, spatial distribution, and genetics into collection design. This work is important and timely knowledge for managers and policy makers because limited conservation resources demand effective, efficient investment.

Acknowledgments

I thank two anonymous reviewers, as well as collaborators on this work including Chris Richards, Michael Way, Naomi Fraga, Scott Schlarbaum, Clare Trivedi, Simon Kallow, Jill Hamilton, and Allan Strand. I thank several seed banks for funding scientific research visits including Rancho Santa Ana Botanic Gardens, NCGRP, and the Millennium Seed Bank Partnership.

Literature Cited

Basey, A.C.; Fant, J.B.; Kramer, A.T. 2015. Producing native plant materials for restoration: 10 rules to collect and maintain genetic diversity. Native Plants Journal. 16: 37–53.

Bataillon, T.M.; David, J.L.; Schoen, D.J. 1996. Neutral genetic markers and conservation genetics: simulated germplasm collections. Genetics. 144: 409–417.

Bland, L.M.; Collen, B.; Orme, C.D.L.; Bielby, J. 2015. Predicting the conservation status of data-deficient species. Conservation Biology. 29: 250–259.

Brown, A.H.D.; Marshall, D.R. 1995. A basic sampling strategy: theory and practice. In: Guarino L.; Ramanantha Rao V.; Read, R., eds. Collecting plant genetic diversity: technical guide-lines. Wallingford, UK: CAB International: 75–92.

Cochrane, J.A.; Crawford, A.D.; Monks, L.T. 2007. The significance of ex situ seed conservation to reintroduction of threatened plants. Australian Journal of Botany. 55: 356–361.

CPC. 1991. Genetic sampling guidelines for conservation collections of rare plants. In: Falk, D.; Holsinger, K.E., eds. Genetics and conservation of rare plants. New York: Oxford Press: 225–238.

Gapare, W.J.; Yanchuk, A.D.; Aitken, S.N. 2008. Optimal sampling strategies for capture of genetic diversity differ between core and peripheral populations of *Picea sitchensis* (Bong.) Carr. Conservation Genetics. 9: 411–418.

Grivet, D.; Robledo-Arnuncio, J.J.; Smouse, P.E.; Sork, V.L. 2009. Relative contribution of contemporary pollen and seed dispersal to the effective parental size of seedling population of California valley oak (*Quercus lobata* Née). Molecular Ecology. 18: 3967–3979.

Guerrant, E.O.; Havens, K.; Vitt, P. 2014. Sampling for effective ex situ plant conservation. International Journal of Plant Sciences. 175: 11–20.

Hoban, S.; Schlarbaum, S. 2014. Optimal sampling of seeds from plant populations for ex-situ conservation of genetic biodiversity, considering realistic population structure. Biological Conservation. 177: 90–99.

Hoban, S.; Strand, A. 2015. Ex situ seed collections will benefit from considering spatial sampling design and species' reproductive biology. Biological Conservation. 187: 182–191.

Hoban, S.M.; Strand. A.; Fraga, N.; Richards. C.; Schlarbaum, S. 2015. Developing quantitative seed sampling protocols using simulations: a reply to comments from Guja et al. and Guerrant et al. Biological Conservation. 184: 469–470.

Lawrence, M.J.; Marshall, D.F.; Davies, P. 1995. Genetics of genetic conservation. I. Sample size when collecting germplasm. Euphytica. 84: 89–99.

Murray, K.A.; Verde Arregoitia, L.D.; Davidson, A.; Di Marco, M.; Di Fonzo; M.M.I. 2014. Threat to the point: improving the value of comparative extinction risk analysis for conservation action. Global Change Biology. 20: 483–494.

Sork, V.; Smouse, P. 2006. Genetic analysis of landscape connectivity in tree populations. Landscape Ecology. 21: 821–836.

Vekemans, X.; Hardy, O.J. 2004. New insights from fine-scale spatial structure analyses in plant populations. Molecular Ecology. 13: 921–935.