


Simulating pollen flow and field sampling constraints helps revise seed sampling recommendations for conserving genetic diversity

Kaylee J. Rosenberger^{1,2}  | Sean Hoban^{2,3}

¹The Morton Arboretum, 4100 IL-53, Lisle, Illinois 60532, USA

²Department of Ecology and Evolutionary Biology, University of Colorado Boulder, 1900 Pleasant St., Boulder, Colorado 80302, USA

³Committee on Evolutionary Biology, The University of Chicago, 1025 E. 57th St., Chicago, Illinois 60637, USA

Correspondence

Kaylee J. Rosenberger, The Morton Arboretum, 4100 IL-53, Lisle, Illinois, 60532, USA.
Email: kaylee.rosenberger@colorado.edu

This article is part of the special issue “From Theory to Practice: New Innovations and Their Application in Conservation Biology.”

Abstract

Premise: In this study, we use simulations to determine how pollen flow and sampling constraints can influence the genetic conservation found in seed collections.

Methods: We simulated genotypes of parental individuals and crossed the parents based on three different ranges of pollen flow (panmictic, limited, and highly limited) to create new seed sets for sampling. We tested a variety of sampling scenarios modeled on those occurring in nature and calculated the proportion of alleles conserved in each scenario.

Results: We found that pollen flow greatly influences collection outcomes, with panmictic pollen flow resulting in seed sets containing 21.6% more alleles than limited pollen flow and 48.6% more alleles than highly limited pollen flow, although this impact diminishes when large numbers of maternal plants are sampled. Simulations of realistic seed sampling (sampling more seed from some plants and fewer from others) showed a relatively minor impact (<2.5%) on genetic diversity conserved compared to ideal sampling (uniform sampling across all maternal plants).

Discussion: We conclude that future work must consider limited pollen flow, but collectors can be flexible with their sampling in the field as long as many unique maternal plants are sampled. Simulations remain a fruitful method to advance ex situ sampling guidelines.

KEYWORDS

ex situ conservation, genetic diversity, population genetics, sampling guidelines, simulations

Botanic gardens and arboreta are important conservation resources, preserving species' genetic diversity ex situ. In some cases, botanic gardens can conserve a large proportion of species' wild genetic diversity, creating a reservoir of genetic material for future restoration efforts or to safeguard genetic material in case a species goes extinct in the wild (Hoban, 2019; Hoban et al., 2020; Zumwalde et al., 2022). Creating and maintaining genetically diverse collections is becoming an increasingly important goal for gardens in the changing climate (Westwood et al., 2021), but often requires large investments of both time and effort (see Zumwalde et al., 2022).

Carefully informed sampling strategies are required to create genetically diverse ex situ collections (Guerrant et al., 2014; Bragg et al., 2021). As collections are created from seed or cuttings from wild populations, they often represent a reduced

subset of wild genetic diversity. Collector decisions about which and how many populations, plants, and seeds to sample can have a large impact on the genetic diversity conserved ex situ. In addition, the collection of genetic diversity often shows diminishing returns with increased effort (Hoban et al., 2020; Rosenberger et al., 2022) due to inherent mathematical relationships (Exposito-Alonso et al., 2022).

Early recommendations used analytical models to show that a sample size of 50 individuals per population conserves at least 95% of common alleles (frequency >0.05) for crop species (Marshall and Brown, 1975), a guideline that is still widely applied today (Hoban and Strand, 2015). However, simple, rule-of-thumb sampling guidelines may not be appropriate for all collection goals. In a recent study using DNA markers, Hoban et al. (2020)

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

compared genetic diversity between in situ populations and ex situ collections and determined that many ex situ collections do not adequately conserve 95% of wild genetic diversity. The reasons for suboptimal conservation of genetic diversity include distribution of genetic diversity over multiple populations, over-representing maternal lines in sampling (sampling many seeds from one plant), and not considering the species' mating system (e.g., pollen and seed dispersal, dispersal distance, and frequency of seed production). For example, Lu-Irving et al. (2023) showed that twice as many populations should be sampled from to adequately conserve genetic diversity in a selfing vs. non-selfing *Hakea* species in Australia.

Recent research has improved sampling strategies, showing that more genetically diverse collections can be created by sampling according to species biology and geographic distribution. For example, Hoban and Strand (2015) determined that when sampling strategies account for the dispersal distance, life span, and life history of a species, a more genetically diverse sample can be obtained. Rosenberger et al. (2022) used geographic models of 14 threatened oak species to establish minimum sample guidelines tailored to each species' particular geography, history, and population sizes.

Simulation models have been a vital tool in this research to develop plant conservation sampling guidelines (Hoban, 2014, 2019; Hoban and Strand, 2015; Rosenberger et al., 2022). Because simulations represent a simplified version of reality, they are useful for exploring and quantifying complex processes that would be difficult or impossible to recreate using traditional empirical experiments (Menges et al., 2004; Yuan et al., 2012). In the context of plant conservation, simulations facilitate the creation of artificial data sets that represent expected genetic patterns among populations and individuals, based on population size, number, and migration. Such simulated data sets represent an abstracted version of reality, as though the investigator has a fully genotyped species. These in silico data sets can then be sampled in various ways to test the effectiveness of ex situ conservation strategies. The robustness of a strategy can be quickly and cheaply evaluated by simulating and sampling species with different characteristics using different sampling approaches; these approaches are termed "scenarios."

However, simulations simplify biological reality. The commonly used, computationally efficient, and flexible software SIMCOAL 2.0 makes assumptions such as random, population-wide mating under the Wright–Fisher model (Excoffier and Foll, 2011), where all individuals in the population are equally likely to mate with any other, and they produce a relatively equal (Poisson-distributed) number of offspring. Often this assumption is not observed in real species, because mating for plants is restricted based on phenological overlap (flowering at the same time), spatially limited pollen and seed dispersal, and in many plants, self-incompatibility mechanisms. Even in wind-pollinated plants, pollen flow can be highly localized, meaning the closest individuals (spatially

to a given maternal plant are most likely to pollinate it (Sork et al., 2002; Bacles and Ennos, 2008). Consequently, a sample of seeds from a single maternal plant may contain genetic contributions from many paternal lines for highly dispersing species (similar to the SIMCOAL assumption), or from only a few of the closest plants (or indeed a single plant) for species with limited dispersal (e.g., insect dispersal). This will influence the amount of genetic diversity conserved in those seeds. In practice, past simulation-based research relying on the assumption of random, population-wide mating could be overestimating the amount of diversity in a sample of seeds from one maternal plant, impacting real-world conservation outcomes.

In addition, the sampling designs used in previous simulation studies have often been simplified compared to reality. Much of the research using resampling techniques based on simulated data has assumed that only one "seed" (equivalent to taking one cutting) is sampled from each plant sampled in a population (Hoban et al., 2018; Hoban, 2019; Bragg et al., 2021; Rosenberger et al., 2022; Zumwalde et al., 2022). It is known that sampling many unique maternal lines is the most efficient and effective method of sampling genetic diversity from a population (Hoban, 2019; Hoban et al., 2020); however, in reality, more than one seed is sampled per plant. In addition, seed collectors may not be able to sample the exact same number of seeds from each plant in the population due to differential reproduction. Some plants are often more productive than others, and occasionally, some plants do not reproduce at all in a given year or for many years (Griffith et al., 2015). This can be quite common in threatened plants. For example, rare species such as *Quercus acerifolia* (E. J. Palmer) Stoyonoff & W. J. Hess and *Zamia integrifolia* L. f. often have only a few maternal individuals producing fruit in a given year, possibly due to environmental conditions, self-incompatibility alleles, Allee effects, and similar issues (Schumacher et al., 2023). Simplifications about sampling seed that were assumed in previous simulation-based resampling studies may generate recommendations resulting in significantly lower genetic diversity being conserved in practice by overrepresenting maternal lineages.

Here, we aim to correct some of these limitations of previous sampling guidelines based on simulations, by increasing the complexity of both the biology of the system and the sampling design, to quantify how simplifications assumed in previous studies impact the genetic diversity conserved in ex situ collections. In particular, we create more biologically realistic simulations of pollination systems and determine the impact of different pollen dispersal types on the diversity conserved in a given sampling strategy. Additionally, we aim to model sampling strategies that are closer to the reality faced by conservation seed collectors and determine both the relative impact of sampling more than one seed per plant and how sampling an unequal number of seeds per plant impacts the diversity conserved with a given sample size.

To accomplish these goals, we first simulate parental genotypes, then create seed sets from the parentals by crossing individuals based on three defined pollen donation types that are generalizations of real pollen dispersal mechanisms. We sample these seed sets in various ways to represent how collectors may sample seed in the wild, ranging from idealized sampling strategies previously assumed in simulation studies to scenarios that more closely resemble sampling in the field. We investigate the relationships between genetic diversity conservation and the total seed sample size, the pollen donor type, and the particular sampling strategy used (ideal or realistic) to identify which variables significantly impact the genetic diversity conserved in a sample.

We hypothesize that for a given sampling strategy, the pollen donor type will impact the genetic diversity represented in the sample relative to the number of potential pollen donors, such that fewer potential pollen donors will result in less genetic diversity being conserved (Aim 1 in Figure 1). Furthermore, we hypothesize that the difference in diversity conserved for each pollen donor type will be largest when

fewer maternal plants are sampled. Similarly, when sampling more seeds per unique maternal plant, we hypothesize that the effect of the number of potential pollen donors will have a stronger impact on the genetic diversity conserved (Aim 2 in Figure 1). Lastly, we hypothesize that sampling an equal number of seeds per plant in the population will result in more genetic diversity conserved from a given sample, compared to sampling an unequal number of seeds from each plant (Aim 3 in Figure 1). We expect this difference to be most apparent, e.g., the strongest effect, in scenarios with a limited number of potential pollen donors.

METHODS

Overview of the methodology

An overview of our methodology is outlined in Figure 2. We simulated generic *in silico* species to determine the impact of more realistic sampling strategies and pollen dispersal types on the genetic diversity conserved in a sample. First,

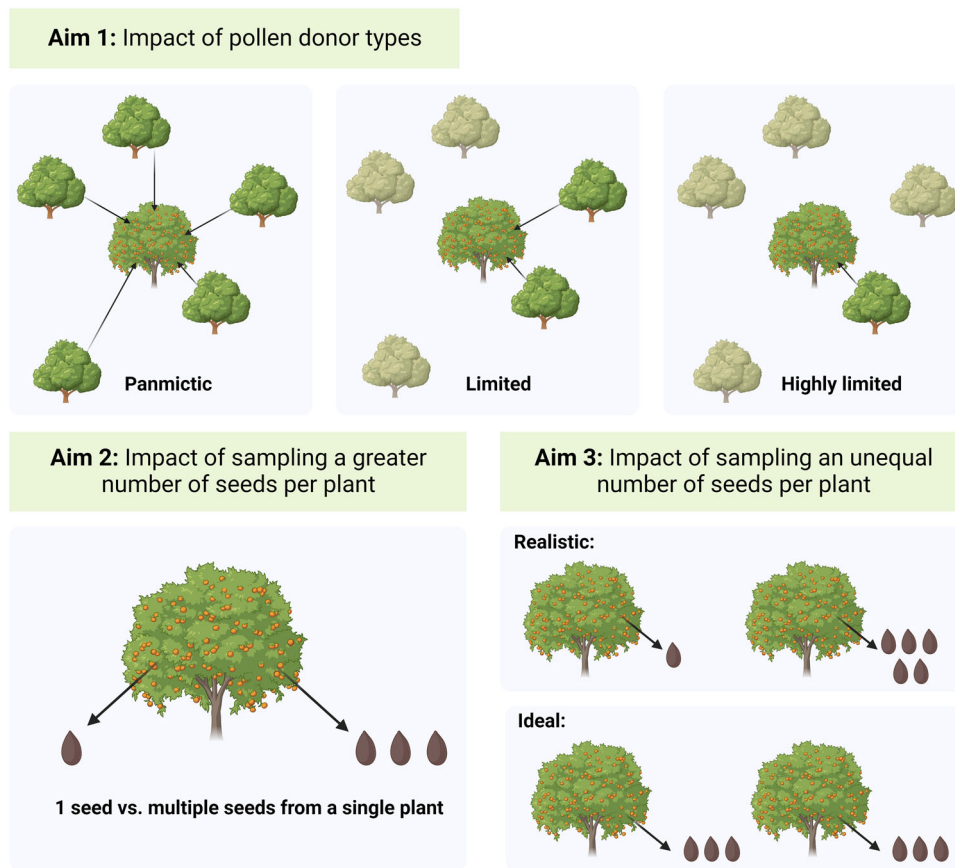


FIGURE 1 A visual representation of aims and how we test their impact on the diversity conserved in a sample of seeds. In Aim 1, we determine the impact of different types of pollen dispersal, ranging from a scenario where all plants in the population can pollinate any other, to a scenario in which only a single plant can pollinate another. Arrows indicate pollen movement. In Aim 2, we determine the impact of sampling an increased number of seeds from a given plant. Here, arrows indicate sampling seed from the plant. In Aim 3, we determine the impact of sampling an unequal number of seeds per plant, compared to an idealized equal number per plant. We test each aim over a range of total sample sizes and varying number of maternal plants sampled. Figure created with [Biorender.com](https://biorender.com).

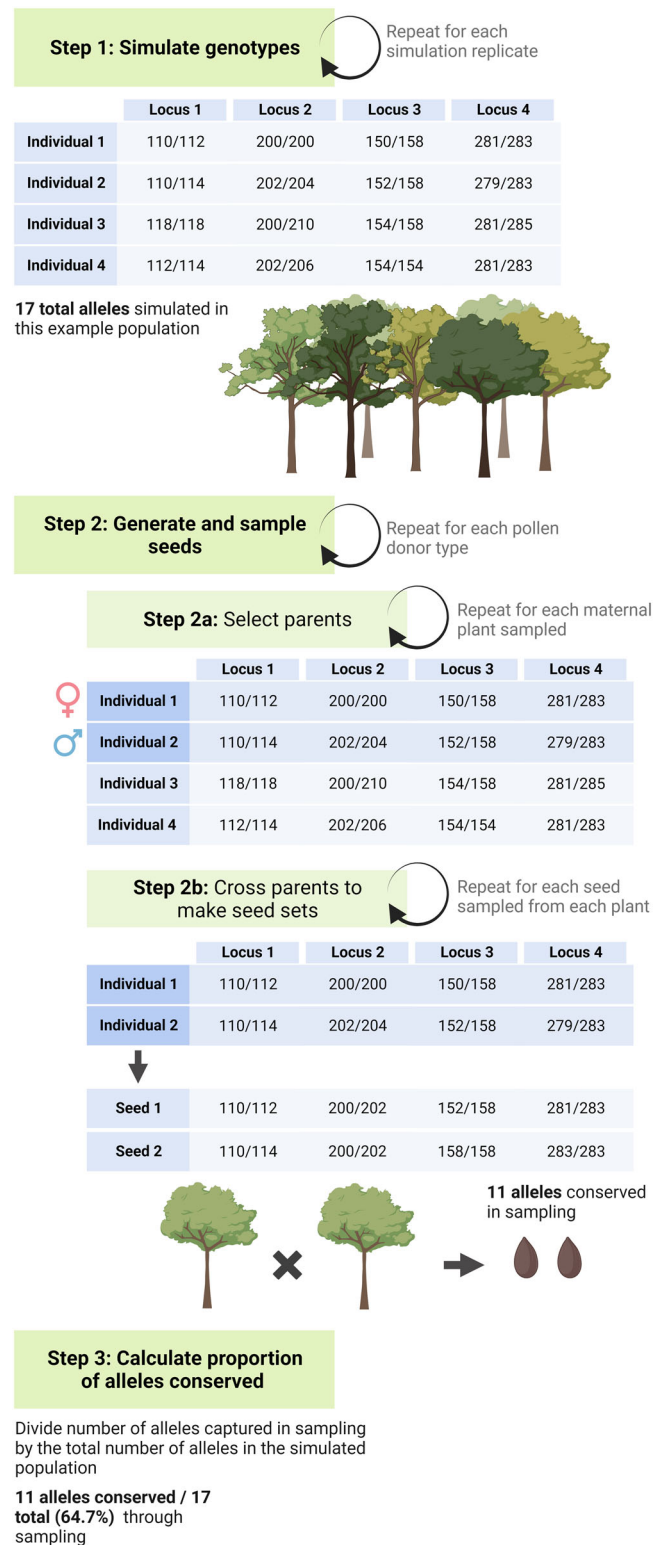


FIGURE 2 Overview of the study methodology summarizing how each step is completed. Step 1: generate an initial population of genotypes using the software SIMCOAL 2.0. Step 2: create a new population of genotypes (representing seeds) by defining different pollen donors to pollinate other plants resulting in a new seed set, then sample from this population of seeds. Step 3: calculate the proportion of the total alleles in the population that were captured to determine genetic conservation success. Figure created with [Biorender.com](https://biorender.com).

we simulated the genotypes of a hypothetical, moderately rare, or Vulnerable species based on the International Union of Conservation of Nature (IUCN) Red List criteria (see “Simulating a model species,” below, and Step 1 in Figure 2). Using these simulated data files to represent parental plants in a population, we created a set of seeds by selecting maternal plants and several pollen-donor plants, according to the three pollination scenarios defined below (see “Simulating pollen flow and creating new populations of seeds,” and Step 2 in Figure 2). Each seed was created by randomly selecting alleles from each parent to create a new individual genotype; this differs from most previous simulations of ex situ collections, which sample a maternal genotype directly rather than simulating pollination. We sampled following various seed sampling strategies as defined in more detail below (under “Sampling seeds”), representing a conservationist collecting seeds in the field. Lastly, we assessed the genetic diversity conserved in the sample by the proportion of wild alleles represented by sampling (see “Calculations,” below, and Step 3 in Figure 2). We test our hypotheses by comparing the proportion of alleles conserved from a given sampling strategy across different pollen dispersal types (Aim 1 in Figure 1) and for more realistic vs. ideal sampling strategies (Aims 2 and 3 in Figure 1), across a large range of maternal plants and seeds available (see “Statistical analyses,” below).

Simulating a model species

We ran simulations in the software SIMCOAL 2.0 (version 2.1.2; Laval and Excoffier, 2004) to simulate population genetic data sets for a generic species, as previously done in similar studies (Hoban et al., 2014; Hoban and Strand, 2015; Hoban, 2019; Rosenberger et al., 2022), because it allows hypotheses to be tested in a controlled environment. We wrote parameter files (.par files) for two sets of simulations representing two hypothetical species: one with a single population of 2500 individuals and another with two populations each of 2500 individuals (5000 individuals total). In the simulation with two populations, we modeled a migration rate of 0.001. Populations were held at a constant size. In both simulations, we modeled 20 unlinked microsatellite loci with mutation rates of 0.0025 per generation and an allele limit of 20 (constrains the number of alleles per locus). The SIMCOAL 2.0 output files represent the individual genotypes of the population. We ran 50 replicates for both simulations to account for the stochasticity of the simulation output.

Simulating pollen flow and creating new populations of seeds

We simulated three different pollen dispersal types to create a realistic population of seed to sample from and to determine the impact of different pollen donor types on

the diversity conserved within a sample. For each of the following scenarios, the output from SIMCOAL 2.0 represents the parental or adult individuals; thus, all of the scenarios have the same data set as a starting point prior to creating the seed set based on the different pollen dispersal types.

First, we simulated a scenario of random, population-wide mating that is similar to the assumptions of previous simulation studies. In this scenario (referred to here as “panmictic”), all individuals in the population have equal probability of pollinating a given maternal plant. Next, we simulated a limited-dispersal scenario that may better represent pollen dispersal in reality, in which the closest plants spatially to a given maternal plant are most likely to pollinate that plant. Here, the number of potential donors for a given maternal plant is restricted to a maximum of 10 potential donors—one with a 60% chance to donate pollen, one with a 20% chance to donate, three with a 5% chance to donate, and five with a 1% chance (hence, we refer to this scenario as “limited”). Lastly, we modeled a highly limited dispersal scenario representing a simplified version of reality, where there is only one potential pollen donor for a given maternal plant (referred to as “highly limited”). This single pollen donor in the highly limited situation is almost always another plant (non-selfing), although a small amount of selfing occurs in the system (about 0.000016%). A visual depiction of our pollen donor scenarios is provided in Figure 1 (Aim 1).

In each scenario, a new set of seeds is created by randomly selecting maternal plants to sample seeds from and creating a list of potential pollen donors for those plants. To make a seed, a pollen donor is first chosen from the list based on the probabilities defined. Alleles from the maternal plant and the selected pollen donor are randomly selected to create the seed's genotype (no mutations occur during the formation of seed). This is repeated for every seed sampled from a maternal plant. After all maternal plants have been sampled, the function returns a matrix containing the genotypes of the new seed population. The code to create new populations of seeds can be found at our GitHub repository (see Data Availability Statement).

Sampling seeds

We sampled from this seed set according to two defined sampling scenarios: an ideal strategy (where an even, equal number of seeds is sampled per plant) and a realistic strategy (where an uneven, unequal number of seeds is sampled from each plant) to determine if more realistic sampling results in significantly lower genetic diversity conservation (see Aim 3 in Figure 1 for a visual representation).

For the ideal strategy, we sampled seeds from a range of maternal plants (1, 2, 5, 10, 25, 50, and 100) and sampled 1–500 seeds per plant. We sampled a range of seeds from each maternal plant to determine the relative balance of sampling seeds and maternal lineages (see Aim 2 in Figure 1). In total, we defined 935 ideal sampling scenarios, which are applied for each pollen donor type.

For the realistic strategy, we sampled an unequal number of seeds per plant, such that the majority of the total sample size was sampled from a single maternal plant. For example, we defined a scenario of sampling 200 total seeds from 100 maternal plants, in which we sampled 50% of the total sample size from one individual (100 seeds), 1% of the total sample size from another individual (two seeds), and 0.5% of the total sample size from each of 98 other individuals (one seed each, 98 seeds combined). Here, we defined scenarios that varied slightly from the ideal strategy, sampling 2, 5, 10, 25, 50, and 100 maternal plants (because sampling an unequal number of seeds per plant cannot occur with only one plant sampled), and sampled 5–400 seeds total. Based on these percentages and by avoiding sampling a fraction of a seed, fewer realistic scenarios were defined than the ideal scenarios. We defined a total of 217 realistic sampling scenarios, which were applied for each pollen donor type. Sampling scenarios are shown in detail in Tables S1 and S2 in Appendix S1 (see Supporting Information with this article) for the ideal and realistic sampling strategies, respectively, with a subset of scenarios shown in Table 1.

Calculations

We calculated the genetic diversity conserved in a sample in terms of the proportion of alleles conserved, by dividing the

TABLE 1 Subset of the 935 ideal sampling scenarios we tested. This table represents only a small subset, meant to demonstrate the possible combinations of variables. In these scenarios, we sample a varying number of seeds from a given number of maternal plants. The seed populations are created based on three different pollen donor types: panmictic, highly limited, and limited. A complete listing of scenarios is provided in Tables S1 and S2 in Appendix S1 for all ideal and realistic sampling scenarios, respectively.

No. of maternal plants sampled	No. of seeds sampled per maternal plant	Pollen donor types	Total seeds sampled	Possible combinations of parameters
100	1, 2, 3 ... 5	Panmictic, limited, highly limited	100, 200, 300 ... 500	5
25	1, 2, 3, 4, 5, 6, 7 ... 20	Panmictic, limited, highly limited	25, 50, 75 ... 500	20
10	1, 2, 3 ... 50	Panmictic, limited, highly limited	10, 20, 30 ... 500	50
2	1, 2, 3 ... 250	Panmictic, limited, highly limited	2, 4, 6, 8 ... 500	250
1	1, 2, 3 ... 500	Panmictic, limited, highly limited	1, 2, 3, 4 ... 500	500

number of alleles captured in the sample by the total number of alleles present in the system. We recognize that seed collections may have other goals, such as focusing only on rare alleles, conserving multiple allele copies, or on relatedness, effective population size, or other metrics of “success”; however, here we focus on all alleles. We calculated this over 50 simulation replicates, 935 ideal and 217 realistic sampling scenarios, and three different pollen donor types, resulting in 172,800 total calculations.

Statistical analyses

To determine the impacts of different pollen donor types on the genetic diversity conserved across a range of total seed sample sizes and maternal plants sampled, we created a linear regression model to compare the proportion of alleles represented in seed samples from our different sampling scenarios. We used the function *lm()* in base R (version 4.3.0; R Core Team, 2023) to predict the proportion of total alleles conserved from the total seeds sampled, number of maternal plants sampled, and pollen donor type, and all pairwise interactions between these factors. There was a distinct nonlinear trend in the data when 1–10 maternal plants were sampled, but when 25–100 maternal plants were sampled, the relationship was linear. Therefore, we ran separate linear models on these subsets of the data. Specifically, we log-transformed the numeric response variables (total seeds sampled and number of maternal plants) in scenarios where 1–10 maternal plants were sampled and did not use a transformation when 25–100 maternal plants were sampled, resulting in a piecewise regression. We assessed the significance of each factor and each interaction, with an alpha threshold of 0.05. We ran the models for ideal and realistic sampling strategies separately; ideal vs. realistic sampling was not a predictor in the model. The equation of the linear regression is provided in Appendix S1.

To determine whether ideal sampling outperforms realistic sampling, we ran a series of pairwise analyses using the function *t.test()* in R to determine significant differences in the proportion of alleles conserved for key scenarios. Specifically, we compared the mean proportion of alleles conserved for ideal and realistic sampling scenarios for each donor type when 200 total seeds were sampled, for different numbers of maternal plants, with a series of *t*-tests (18 *t*-tests total, representing the different combinations of number of maternal plants and pollination types). To correct for multiple comparisons, we used the Benjamin–Hochberg procedure to adjust *P* values. We report both raw and adjusted *P* values.

RESULTS

Overview

We created new seed sets from each model species (one and two population simulations) and sampled seed based

on the defined sampling strategies. We repeated each sampling scenario for each pollen donor type. This process resulted in a total of 172,800 seed sets between the ideal and realistic scenarios. In simulations with one population, the total number of alleles simulated across replicates ranged from 235 to 288 (due to the stochasticity of the simulations), which for our 20 loci is approximately 12 to 14 alleles per locus, a reasonable number for a microsatellite study. We focus here on the results of the single population simulations and provide results for the two-population scenarios in Figures S4–S6 and Tables S6–S8 in Appendix S1.

Increasing the number of maternal plants sampled

Figure 3 shows the genetic diversity conservation for all ideal sampling scenarios (ideal vs. realistic seed sampling will be discussed below under “Aim 3”) and for each donor type, along with the fitted linear regression model (model coefficients and *P* values are provided in Table S3 in Appendix S1). When more unique maternal plants were sampled, the proportion of genetic diversity conserved increased for all donor types (curves approach 100% diversity conservation as maternal plants sampled increase).

Aim 1: Impact of pollen donor types

Across nearly all scenarios, the “panmictic” pollen donor type conserved the most genetic diversity, followed by “limited” and “highly limited,” visualized by the differences in the curves for each donor type in Figure 3. The difference in diversity conservation between pollen donor types is largest when few maternal plants are sampled and when many seeds are sampled from those plants. In scenarios with 10 or fewer maternal plants sampled, the pollen donor type had a particularly strong impact on the diversity conserved for a given sampling scenario. For instance, when two maternal plants are sampled and 200 total seeds are collected, “panmictic” scenarios conserve 86.3% of the total alleles as predicted by the linear model, “limited” pollination conserves 65.9%, and “highly limited” pollination conserves just 39.3%. Thus, the same sampling effort in scenarios with a single pollen donor conserves less than half the diversity compared to scenarios where pollen dispersal is population-wide. In extreme cases of a single pollen donor (“highly limited”) and a single maternal plant, conserved genetic diversity does not exceed 25%. In scenarios with many maternal plants sampled (25 or more), the difference in the proportion of alleles conserved between pollen donor types became small, and at 100 maternal plants, is negligible. Nearly all factors were significant in the regression, with the exception of the interaction of the

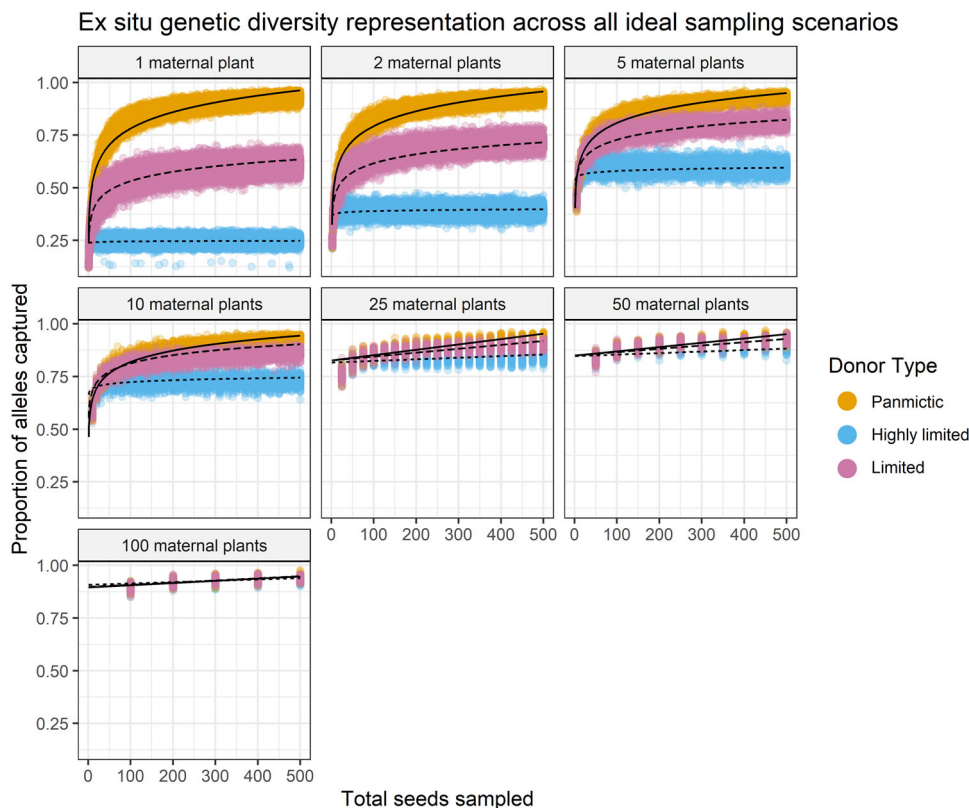


FIGURE 3 Genetic diversity conserved for all ideal sampling scenarios and donor types. Genetic diversity conservation is shown on the y-axis as the proportion of all wild alleles conserved by sampling, and the x-axis represents the total number of seeds sampled, ranging from 1 to 500. Each facet on the plot represents the number of unique maternal plants sampled. Colors represent each pollen donor type, and each point on the plot is a single sampling scenario.

“limited” and “panmictic” donor types, and the interaction of maternal plants and the “limited” donor type when 25–100 maternal plants were sampled (Table S4 in Appendix S1). When 200 total seeds are collected from 100 maternal plants, all pollen donor types conserve around 92% of the total alleles, as predicted by the linear model.

Aim 2: Sampling multiple seeds from a maternal plant

In general, sampling additional seeds from each maternal plant in the population slightly increased genetic diversity conservation for some scenarios, although not as strongly as sampling more unique maternal plants. Specifically, as additional seeds were sampled from each plant in the population, the conserved genetic diversity increased for both the “limited” and “panmictic” pollen donor types, although with diminishing returns (see Figure 3, where the curve along the x-axis increases slightly before leveling off). Sampling additional seeds per plant for the “highly limited” pollen donor scenarios did not result in much more genetic diversity conserved (“highly limited” scenarios have a slope close to 0), although this interaction was significant in the regression (Table S4 in Appendix S1). Thus, the amount of

possible increase in genetic diversity conservation from resampling an individual depends on the number of pollen donors.

Aim 3: Ideal vs. realistic sampling

The relationships between the number of maternal plants sampled, pollen donor types, total sample size, and proportion of alleles conserved were often similar between the ideal and realistic sampling scenarios (see Figure S1 in Appendix S1 for a comparison to Figure 3). In other words, sampling an unequal number of seeds per plant appeared to have a small impact on the genetic diversity conserved by a given sampling strategy. In Figure 4, we compare the diversity conserved between ideal and realistic sampling strategies for each pollen donor type in scenarios when 200 total seeds are sampled, because this is when most scenarios can be directly compared. Figure 4 indicates comparisons between ideal and realistic sampling for each pollen donor type, and Table 2 shows the *P* values for each *t*-test and as the actual difference in the proportion of alleles conserved for ideal and realistic sampling. The ideal sampling scenario performs significantly better than realistic sampling when 50–100 maternal plants are sampled and in some other scenarios for “limited” and “highly limited” pollen donor

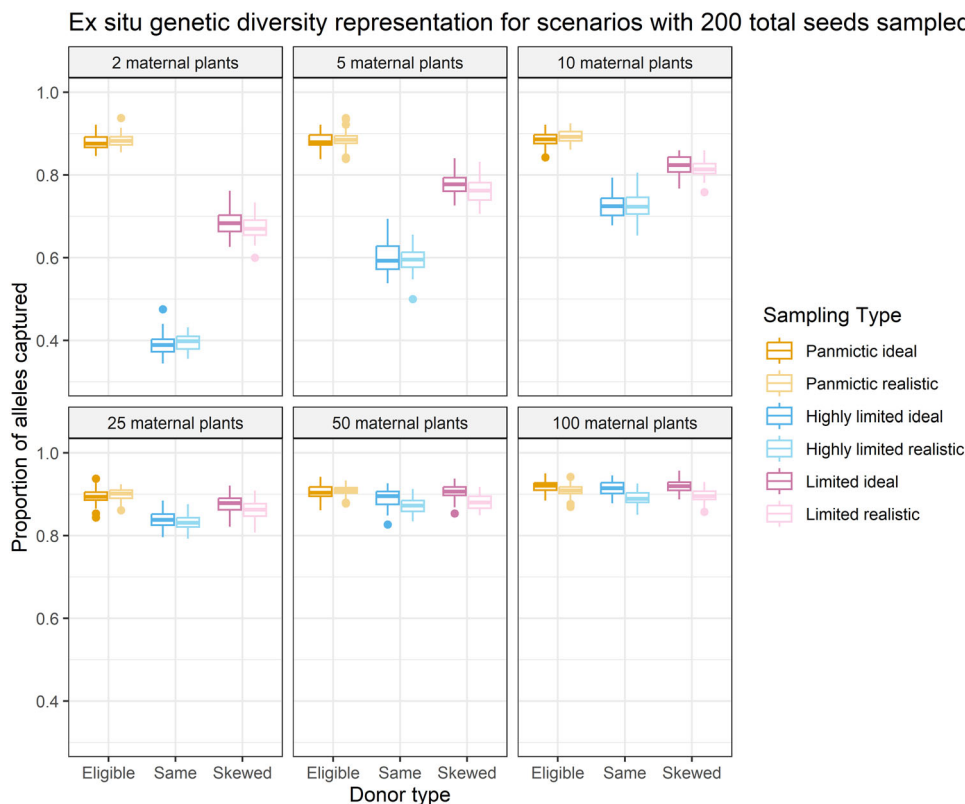


FIGURE 4 Genetic diversity conserved between ideal and realistic sampling scenarios for all pollen donor types when 200 total seeds are sampled across the population. The colors represent the type of sampling strategy, ranging from lighter (realistic scenarios) to darker shades (ideal scenarios). The y-axis represents the proportion of wild alleles conserved by sampling, and the x-axis shows the pollen donor types. The plot is faceted based on the number of maternal plants sampled.

types, although there is often less than a 2% difference (see Tables S1 and S2 in Appendix S1 for detailed notes on this particular sampling scenario).

DISCUSSION

Overview

In this paper, we aimed to determine how pollination biology, sampling more maternal plants, and realistic on-the-ground sampling will impact the genetic diversity captured in a sample of seed collected for conservation purposes, such as for conservation planting or storing in a seed bank. Recent empirical and simulation work has shown that current widely used rule-of-thumb sampling strategies may not effectively conserve genetic diversity of species ex situ; this can be due to, for example, species' biological traits or population structure, or to overrepresentation of maternal lines in sampling (Griffith et al., 2017; Hoban et al., 2020). Past simulation studies (e.g., Hoban and Schlarbaum, 2014; Hoban, 2019) have lacked several key elements of realism for this issue. Here, we implemented more complex pollination biology and realistic sampling strategies on a simulated population genetic data set to

determine how these factors influence the wild genetic diversity that can be represented in a sample.

Key results

We confirmed that representing more unique maternal lineages is the most efficient way of adequately representing diversity ex situ (as described in Hoban and Strand, 2015; Hoban et al., 2018; Griffith et al., 2019). That is, the number of maternal individuals to sample is one of the most important factors to consider when designing a sampling strategy. In a recent study, Bragg et al. (2021) also confirmed this result by resampling from an empirically derived data set, in contrast to the data set we created in silico. Although this finding has been established in the field, we confirm this is true in all cases of varying pollination biology and realistic sampling employed here.

We determined that genetic diversity is more difficult to represent ex situ for species with limited pollen dispersal than for species with widespread pollen dispersal. For species with limited or highly limited dispersal, it becomes increasingly important to sample more unique maternal plants to adequately conserve genetic diversity ex situ. In scenarios with limited pollen donors, assuming the

TABLE 2 Significance levels from *t*-tests comparing the mean proportion of alleles conserved in ideal vs. realistic sampling scenarios for each donor type, when 200 seeds were sampled total. Boldfaced text indicates significant differences.

No. of maternal plants sampled	Donor type	Proportion of alleles conserved (ideal)	Proportion of alleles conserved (realistic)	Difference (ideal – realistic)	<i>P</i> values	Adjusted <i>P</i> values
2	Highly limited	0.390	0.395	–0.005	0.291	0.349
2	Limited	0.684	0.671	0.013	0.026	0.051
2	Panmictic	0.878	0.883	–0.005	0.130	0.191
5	Highly limited	0.600	0.596	0.004	0.434	0.459
5	Limited	0.777	0.761	0.016	<0.001	0.002
5	Panmictic	0.882	0.886	–0.004	0.160	0.206
10	Highly limited	0.723	0.726	–0.004	0.500	0.500
10	Limited	0.824	0.816	0.008	0.060	0.098
10	Panmictic	0.886	0.894	–0.008	0.019	0.043
25	Highly limited	0.838	0.832	0.006	0.138	0.191
25	Limited	0.877	0.862	0.015	<0.001	0.001
25	Panmictic	0.894	0.900	–0.006	0.058	0.098
50	Highly limited	0.891	0.872	0.019	<0.001	<0.001
50	Limited	0.905	0.881	0.024	<0.001	<0.001
50	Panmictic	0.906	0.909	–0.003	0.342	0.384
100	Highly limited	0.914	0.890	0.024	<0.001	<0.001
100	Limited	0.919	0.896	0.024	<0.001	<0.001
100	Panmictic	0.919	0.910	0.010	0.003	0.007

conservation target is representing more alleles, sampling additional seeds per individual does not result in significantly more novel genetic diversity conserved. The diversity available from a sample of seeds from a single maternal plant is limited to the number of pollen donors, i.e., less genetic diversity is available with fewer pollen donors. For example, in Figure 3, for scenarios with only one potential pollen donor (“highly limited” scenarios), sampling additional seeds from a maternal plant does not result in more genetic diversity conserved after fewer than 50 seeds are sampled (slope for these curves = 0). We also note that such seed sets are more likely to produce inbred future populations because there are fewer parental contributions. Thus, it is necessary to sample more unique maternal individuals for pollen-limited species than for widely dispersing species to conserve sufficient genetic diversity. This is an important observation for seed collectors in the field, who sometimes may find only a single fruiting plant in a given year. Unfortunately, if the species’ pollen dispersal is limited, collecting hundreds of seeds from this plant may not significantly increase the genetic diversity conserved. Knowledge about the extent of pollen dispersal may not be available for a target species, as pollen dispersal is influenced by many factors, including abundance, density, ecosystem management, the type of pollinators and their abundance, and phenological overlap (Ghazoul, 2005). Usually, pollen is

not panmictic (Degen et al., 2004; Wagenius and Lyon, 2010; Deacon and Cavender-Bares, 2015; Xiang et al., 2022), with many species being more similar to our “limited” scenario; therefore, we advise seed collectors to (a) consider these biological factors of their target species, and (b) in the absence of knowledge, assume that the species is limited or highly limited and sample many unique maternal lineages per population.

Lastly, we determined that sampling an unequal number of seeds per plant (which is more realistic than sampling an idealized, equal number of seeds per plant, as assumed in previous studies) has a small impact on the genetic diversity conserved by sampling. We did observe significant differences in the mean proportion of alleles conserved for some ideal vs. realistic scenarios, occurring mostly in “limited” pollen donor scenarios and when more maternal plants were sampled; however, the actual difference was only 1–2.5% or less (Table 2). In fact, the difference between ideal and realistic sampling strategies is negligible when compared to the variation between pollen donor types and sampling more unique maternal plants. This result indicates that in practice, seed collectors can sample an unequal number of seeds per plant (up to 50% of the total seed collection size from a single maternal individual in some cases), as long as many unique maternal plants are sampled. This result was surprising, as we expected realistic sampling

to perform significantly worse, as described above under Aim 3. Nonetheless, this is encouraging for seed samplers in the field.

Related work

Our work builds on research conducted over the past several decades, developing informed sampling guidelines to conserve genetic diversity in *ex situ* collections, while addressing some limitations of previous work. Similar to Hoban and Strand (2015), we examined the effects of short- and long-distance dispersal (“limited” and “panmictic” scenarios in our study, respectively) and sampling more seeds per maternal plant. To build on their work, we included the additional case of a single pollen donor (“highly limited” scenario in our study), representing a species with extremely limited dispersal. We also increased the complexity of sampling by investigating ideal vs. realistic sampling scenarios. We implemented a wide range of realistic sampling scenarios in which an unequal number of seeds were sampled per plant in the population (for example, 50 seeds from one plant, and one seed from 49 other plants). Lastly, to build on previous work, we examined each scenario in detail and determined when the pollination biology significantly impacts the diversity conserved in a sample (i.e., very little after 50 maternal plants are sampled; see Figure 3). In related work, Bragg et al. (2021) examined the spatial arrangement of establishing the plants in the *ex situ* collection, to avoid pollination among close relatives (inbreeding) and conserve genetic diversity in future generations.

Our results indicate that the commonly used 50-sample guideline (originating from Marshall and Brown, 1975) is robust to different pollen dispersal types if the samples come from 50 unique maternal plants. As shown in Figure 3, when 50 maternal plants are sampled, there are negligible differences in the diversity conserved between pollen donor types. However, we note that the 50-sample guideline should be followed with caution, as we did not exceed 95% genetic diversity (a common threshold for adequate genetic diversity) in any of our sampling scenarios when all alleles were considered. This reinforces the finding that this generic guideline may not adequately conserve genetic diversity for all species, depending on collection goals.

Caveats and future work

Simulations are a useful tool for uncovering complex biological processes; however, they have limitations. Although we build on the complexity of pollen movement within a population, we make other simplifying biological assumptions. First, we do not explicitly model a spatial distribution in the simulation, i.e., individuals are not arranged in a fixed position throughout space. Instead,

maternal plants and pollen donors are chosen from throughout the population. In reality, genetic diversity is clustered within a population according to a particular structure. However, the limited pollen pool defined here implements a spatial distribution implicitly by restricting the pollen movement to a subset of the population. Future work could model the population with an explicit spatial distribution.

In addition, our simulated species is assumed to be hermaphroditic (i.e., each individual can produce both pollen and seeds), but some species are dioecious. To make the results generalizable, a future study could determine if the same results will be found for both hermaphroditic and dioecious species. In a dioecious species, the total pollen pool is proportionally smaller compared to hermaphroditic species, which may reduce the amount of genetic diversity in seeds collected. Furthermore, because individuals are randomly selected from the population as maternal plants and pollen donors, there is a small chance of selfing if a maternal plant is also chosen as the pollen donor. In reality, not all plant species are self-compatible, while others have very high rates of selfing.

Here, we modeled microsatellite loci, which may not represent genetic diversity that has adaptive value for a species. We also did not consider species with a historical bottleneck resulting in low genetic diversity (see Rosenberger et al., 2022). Finally, we did not implement mutations when we created the population of seeds. Future work can help better link plant biology, genomics, and effectiveness of conservation outcomes, and we emphasize that there remain many opportunities for more realistic simulations to refine seed sampling guidance. Sometimes, as with our observations on a small impact of “realistic” sampling, simulations will help reveal unexpected results; moreover, simulations are often the only way to obtain such knowledge (Peck, 2004; Landguth et al., 2010; Hoban et al., 2012).

Lastly, our recommendations are based on the assumption that the goals of this conservation collection are for planting as a living collection or for storage in a seed bank. In this study, we do not consider the use of seed collected for other purposes, such as scientific study or conserving multiple allele copies. We note that if conserving multiple allele copies is important for a collection, sampling will need to be increased, particularly for pollen-limited systems (Hoban et al., 2018; Schumacher et al., 2023).

Conclusions

We conclude that simulating a more complex level of biological and logistical reality can help improve guidance for seed samplers. In particular, the degree to which a maternal plant receives highly limited, moderately limited, or unlimited/panmictic paternal pollen pools greatly impacts the genetic diversity in a conservation seed collection, especially when relatively few maternal

plants and large numbers of seeds are collected. The situation of few maternal plants and large numbers of seeds is commonly encountered by seed collectors, particularly when sampling rare species. Our results suggest that the minimum sampling guidelines for species with limited pollen dispersal should likely be even higher than previously suggested. Conservation seed collectors must understand and account for pollination biology. On the other hand, and surprisingly, logistical realities in allocating sampling effort among plants had less impact on genetic diversity conservation. We still recommend that collectors sample as many maternal plants as possible, but our results demonstrate that collectors do not need to worry about sampling precisely the same amount of seed from every plant. There are many future avenues for simulations to generate more precise guidance for seed collectors.

AUTHOR CONTRIBUTIONS

S.H. and K.J.R. conceived the study design and wrote the manuscript. K.J.R. wrote the majority of the simulation and analysis code, with some input from S.H. K.J.R. created the figures. Both authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank the Hoban lab for comments, mentoring, and training, especially Emily Schumacher, Austin Koontz, and Ash Hamilton. We acknowledge funding from the National Science Foundation (NSF; REU grant 1851961 and ABI grant 1759759), The Morton Arboretum Center for Tree Science Research Experience Extension Fellowship, and the Institute of Museum and Library Services (grants MA-30-18-0273-18 and MG-251613-OMS-22). K.J.R. is also supported by an NSF Graduate Research Fellowship.

DATA AVAILABILITY STATEMENT

Code for recreating all simulated data and all data analysis can be found at the GitHub repository: https://github.com/HobanLab/Pollen_dispersal_sims.

ORCID

Kaylee J. Rosenberger  <http://orcid.org/0000-0003-1890-8933>

REFERENCES

- Bacles, C., and R. Ennos. 2008. Paternity analysis of pollen-mediated gene flow for *Fraxinus excelsior* L. in a chronically fragmented landscape. *Heredity* 101: 368–380. <https://doi.org/10.1038/hdy.2008.66>
- Bragg, J. G., J. Y. S. Yap, T. Wilson, E. Lee, and M. Rossetto. 2021. Conserving the genetic diversity of condemned populations: Optimizing collections and translocation. *Evolutionary Applications* 14(5): 1225–1238.
- Deacon, N. J., and J. Cavender-Bares. 2015. Limited pollen dispersal contributes to population genetic structure but not local adaptation in *Quercus oleoides* forests of Costa Rica. *PLoS ONE* 10(9): e0138783.
- Degen, B., E. Bandou, and H. Caron. 2004. Limited pollen dispersal and biparental inbreeding in *Symphonia globulifera* in French Guiana. *Heredity* 93: 585–591.
- Excoffier, L., and M. Foll. 2011. fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27(9): 1332–1334.
- Exposito-Alonso, M., T. Booker, L. Czech, L. Gillespie, S. Hateley, C. Kyriazis, P. Lang, et al. 2022. Genetic diversity loss in the Anthropocene. *Science* 377: 1431–1435. <https://doi.org/10.1126/science.abn5642>
- Ghazoul, J. 2005. Pollen and seed dispersal among dispersed plants. *Biological Reviews* 80(3): 413–443.
- Griffith, M. P., M. Calonje, A. Meerow, F. Tut, A. Kramer, A. Hird, T. Magellan, and C. Husby. 2015. Can a botanic garden cypad collection capture the genetic diversity in a wild population? *International Journal of Plant Sciences* 176(1): 1–10. <https://doi.org/10.1086/678466>
- Griffith, M. P., M. Calonje, A. W. Meerow, J. Francisco-Ortega, L. Knowles, R. Aguilar, F. Tut, et al. 2017. Will the same ex situ protocols give similar results for closely related species? *Biodiversity and Conservation* 26: 2951–2966. <https://doi.org/10.1007/s10531-017-1400-2>
- Griffith, M. P., E. Beckman, T. Calicrate, J. R. Clark, T. Clase, S. Deans, M. Dosmann, et al. 2019. Toward the metacollection: Safeguarding plant diversity and coordinating conservation collections. Botanic Gardens Conservation International, Richmond, Surrey, United Kingdom. Available at: <https://www.bgci.org/wp/wp-content/uploads/2019/09/Toward-the-Metacollection-Coordinating-conservation-collections-to-safeguard-plant-diversity.pdf> [accessed 5 December 2023].
- Guerrant Jr., E. O., K. Havens, and P. Vitt. 2014. Sampling for effective ex situ plant conservation. *International Journal of Plant Sciences* 175(1): 11–20.
- Hoban, S. 2014. An overview of the utility of population simulation software in molecular ecology. *Molecular Ecology* 23(10): 2383–2401.
- Hoban, S. 2019. New guidance for ex situ gene conservation: Sampling realistic population systems and accounting for collection attrition. *Biological Conservation* 235: 199–208.
- Hoban, S., and S. Schlarbaum. 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., and A. Strand. 2015. Ex situ seed collections will benefit from considering spatial sampling design and species' reproductive biology. *Biological Conservation* 187: 182–191.
- Hoban, S., G. Bertorelle, and O. E. Gaggiotti. 2012. Computer simulations: Tools for population and evolutionary genetics. *Nature Reviews Genetics* 13(2): 110–122. <https://doi.org/10.1038/nrg3130>
- Hoban, S., J. Arntzen, M. Bruford, J. Godoy, A. Hoelzel, G. Segelbacher, C. Vila, and G. Bertorelle. 2014. Comparative evaluation of potential indicators and temporal sampling protocols for monitoring genetic erosion. *Evolutionary Applications* 7(9): 984–998.
- Hoban, S., S. Kallow, and C. Trivedi. 2018. Implementing a new approach to effective conservation of genetic diversity, with ash (*Fraxinus excelsior*) in the UK as a case study. *Biological Conservation* 225: 10–21.
- Hoban, S., T. Calicrate, J. Clark, S. Deans, M. Dosmann, J. Fant, O. Gailing, et al. 2020. Taxonomic similarity does not predict necessary sample size for ex situ conservation: A comparison among five genera. *Proceedings of the Royal Society B: Biological Sciences* 287(1926): 20200102.
- Landguth, E. L., S. A. Cushman, M. K. Schwartz, K. S. McKelvey, M. Murphy, and G. Luikart. 2010. Quantifying the lag time to detect genetic barriers in landscape genetics. *Molecular Ecology* 19(19): 4179–4191.
- Laval, G., and L. Excoffier. 2004. SIMCOAL 2.0: A program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* 20: 2485–2487.
- Lu-Irving, P., J. G. Bragg, M. Rossetto, K. King, M. O'Brien, and M. M. van der Merwe. 2023. Capturing genetic diversity in seed collections: An empirical study of two congeners with contrasting

- mating systems. *Plants* 12(3): 522. <https://doi.org/10.3390/plants12030522>
- Marshall, D. R., and A. H. D. Brown. 1975. Optimum sampling strategies in conservation. International Biological Programme 2: Crop genetic resources for today and tomorrow. Cambridge University Press, Cambridge, United Kingdom.
- Menges, E. S., E. O. Guerrant Jr., and S. Hamz . 2004. Effects of seed collection on the extinction risk of perennial plants. In E. O. Guerrant Jr., K. Havens, and M. Maunder [eds.], *Ex situ plant conservation: Supporting species survival in the wild*, 305–324. Island Press, Washington, D.C., USA.
- Peck, S. L. 2004. Simulation as experiment: A philosophical reassessment for biological modeling. *Trends in Ecology and Evolution* 19(10): 530–534.
- R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website: <https://www.R-project.org/> [accessed 5 December 2023].
- Rosenberger, K., E. Schumacher, A. Brown, and S. Hoban. 2022. Species-tailored sampling guidelines remain an efficient method to conserve genetic diversity ex situ: A study on threatened oaks. *Biological Conservation* 275: 109755.
- Schumacher, E. K., Y. Wu, A. Byrne, S. Gray, L. Ladd, P. M. Griffith, and S. Hoban. 2023. Examining previously neglected aspects of ex situ gene conservation in two IUCN Threatened plant species: Rare alleles, redundancy, ecogeographic representativeness, and relatedness. *International Journal of Plant Sciences*. <https://doi.org/10.1086/728186>
- Sork, V., F. Davis, P. Smouse, R. Dyer, J. Fernandez-M., and B. Kuhn. 2002. Pollen movement in declining populations of California Valley oak, *Quercus lobata*: Where have all the fathers gone? *Molecular Ecology* 11(9): 1657–1668.
- Wagenius, S., and S. P. Lyon. 2010. Reproduction of *Echinacea angustifolia* in fragmented prairie is pollen-limited but not pollinator-limited. *Ecology* 91(3): 733–742.
- Westwood, M., N. Cavender, A. Meyer, and P. Smith. 2021. Botanic garden solutions to the plant extinction crisis. *Plants People Planet* 3(1): 22–32.
- Xiang, W. Q., P. L. Malabrigo, L. Tang, and M. X. Ren. 2022. Limited-distance pollen dispersal and low paternal diversity in a bird-pollinated self-incompatible tree. *Frontiers in Plant Science* 13: 806217. <https://doi.org/10.3389/fpls.2022.806217>
- Yuan, X., D. J. Miller, J. Zhang, D. Herrington, and Y. Wang. 2012. An overview of population genetic data simulation. *Journal of Computational Biology* 19(1): 42–52.
- Zumwalde, B., B. Fredlock, E. Beckman Bruns, D. Duckett, R. McCauley, E. Suzuki Spence, and S. Hoban. 2022. Assessing ex situ genetic and ecogeographic conservation in a threatened but widespread oak after range-wide collecting effort. *Evolutionary Applications* 15(6): 1002–1017.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Supporting information for “Simulating pollen flow and field sampling constraints helps revise seed sampling recommendations for conserving genetic diversity.”

How to cite this article: Rosenberger, K. J., and S. Hoban. 2024. Simulating pollen flow and field sampling constraints helps revise seed sampling recommendations for conserving genetic diversity. *Applications in Plant Sciences* e11561. <https://doi.org/10.1002/aps3.11561>