I. Introduction

Seeds (and other plant material) are collected from wild plant populations to form *ex situ* collections in botanic gardens and arboreta. Collections in botanic gardens have many purposes including aesthetics, public education and awareness, research, and conservation. In fact, due to these collections and resources, botanic gardens are well-positioned to help prevent plant extinctions (Westwood et al., 2020). Long term goals of *ex situ* collections often include reintroduction or supplementation of wild populations, especially for rare or vulnerable species. Due to climate change and habitat loss, many plant species may need conservation or restoration intervention in the future. For plant species to adapt to rapid climate and environmental change, high genetic diversity is required. Genetic diversity allows populations to recover after extreme events and to adapt to diseases and pests (Reusch et al., 2005; Reynolds et al., 2012; Morikawa and Palumbi, 2019). In addition, high genetic diversity is a sign of a generally healthy population, as genetic diversity is correlated to fitness (Spielman et al., 2004; Griffith et al., 2014). Lastly, for keystone species like trees, high genetic diversity supports ecosystem processes, community structure, and nature's contributions to people (Stange et al., 2020; Des Roches et al., 2021). Thus, high genetic diversity is required when creating collections that may be used to supplement wild populations or to reintroduce populations in the future. Genetic diversity can ensure successful conservation and restoration of species.

One method of obtaining a genetically diverse collection is through careful design of the sample size (number of seed/cuttings) taken and strategy used (Guerrant et al., 2004). The genetic diversity represented in *ex situ* collections is normally a reduced subset of that in the wild, so the sample size and strategy can have a great effect on the diversity represented in the collection (Guerrant et al., 2004). In the past, generalized 'rules of thumb' have been used to sample seed from the wild (Marshall and Brown, 1975). One widely used guideline advises sampling material from 50 individuals per population (Brown and Marshall, 1995). However, recent work has shown that this is not the most effective or efficient method of sampling a species. In some cases, this guideline may underestimate the sample size required, resulting in a collection that doesn't capture the diversity of the species. In fact, research has shown that most collections in botanic gardens underrepresent the diversity of wild populations (Hoban et al.,

2020). Typically, a sufficient amount of genetic diversity is defined as collecting 95% of the alleles present in the wild population. The more individuals that are conserved results in a greater amount of genetic diversity that is conserved. This amount of genetic diversity allows for more successful reintroductions and conservation of the species (Godefroid et al., 2011).

Previous work has found that when species traits are accounted for in the seed sampling strategy, a higher level of species genetic diversity can be captured more efficiently (Hoban and Strand, 2015). Even between closely related species (which often have similar traits, such as seed and pollen dispersal, life form, breeding system, successional status), optimal sample size and sampling strategy can vary widely, as described in a recent study by Hoban et al. (2020) and previously by Griffith et al. (2017). However, in this study, it is noted that the genus *Quercus* (oaks) could be an exception to this. The authors found that one minimum sample size could be recommended to fit all oak species in that study; about 80 individuals (Hoban et al., 2020). However, it is not known if this finding was due to chance, or because oaks have a similar enough biology that their genetic diversity can truly be captured in one strategy.

In this project, we focused on further investigating this recommendation, and expanding the study to several IUCN Red List endangered oaks. Oaks have high ecological importance and are keystone species for many ecosystems. Furthermore, oaks are described as ‘exceptional’ species, meaning they cannot be seed banked using traditional methods. Instead, oak diversity must be maintained through living collections. Living collections (whole specimen collections) require extensive space and resources; therefore, creating collections that efficiently represent the diversity of wild populations is important. Due to climate change, habitat loss, and agriculture, many species, including oaks, are at great risk for extinction (Carrero et al., 2020). Thus, creating and maintaining genetically diverse collections in botanic gardens and arboreta is essential for the future survival and restoration of these rare, endangered species, and this can be achieved through proper sampling techniques.

Here, we used simulations and computational resampling techniques in R to develop sampling guidelines to efficiently capture the diversity of 14 endangered oak species in the US. Simulations have been used extensively to test and develop sampling strategies for species (Hoban and Schlarbaum, 2014; Hoban and Strand, 2015; Hoban, 2019). Simulation studies also

require less time and resources to complete than empirical studies, so they are highly practical- we do not have genetic data from large numbers of oak species. The main goals of this project were:

I. Determine the minimum sample sizes to efficiently capture the diversity of all 14 species which will directly inform ongoing conservation efforts for these species
II. Answer the question, "Can one minimum sampling strategy fit all oaks?"

II. Methods

A. Simulation overview

For this project, we used genetic simulations to assess the diversity captured by a given sample size. We used the backwards-in-time simulation software Simcoal 2 for all species simulations (Version 2.1.2, Excoffier and Laval, 2005). This software has been used extensively to simulate genetic diversity. First, we created parameter files, which control the simulations, that realistically represent each species. Parameter files include values that represent species traits, such as number of populations, population sizes, migration rates, and historical events. Parameter files were then provided as input for the simulation software. We ran 1000 replicate simulations for every species to account for the stochasticity of the simulation output. The output files of the simulation represent individual genotypes.

B. Study species

We simulated 14 species of oaks in the United States that the IUCN Red List identifies as threatened (see Figure 1 for photos and Figure 2 for the geographic range of several species). This study encompasses almost all threatened oaks within the United States; however, we had to omit a few species with unclear taxonomy, as this would be difficult to simulate. For example, we did not simulate *Quercus ajoensis* even though it is vulnerable because, "it is part of a poorly understood species complex which requires taxonomic revision" (Kenny et al., 2020). To write realistic parameter values for each species, we used population and geographic range data from the IUCN Red List. For most species, we simplified the data obtained from the Red List by aggregating localities, such as for *Quercus oglethorpensis,* which had a large number of listed localities that we combined into 5 larger populations. We also simplified data by reducing the population sizes by some factor for some species (due to computational and time constraints); for

example, we reduced estimated population sizes for *Quercus engelmannii* by a factor of 10 (from an estimate of 200,000 individuals to a simulation of 20,000 individuals). We wrote unique parameter values for each species; a subset of the parameter values used to represent several species is shown in Table 1. To verify whether these parameter values realistically represented each species, we compared simulated Fst values to empirical data where it was available. Species with smaller population sizes and lower migration rates typically have higher Fst values (more genetically differentiated populations). If these values were not comparable, we altered the migration rates in the parameter files and re-ran simulations.

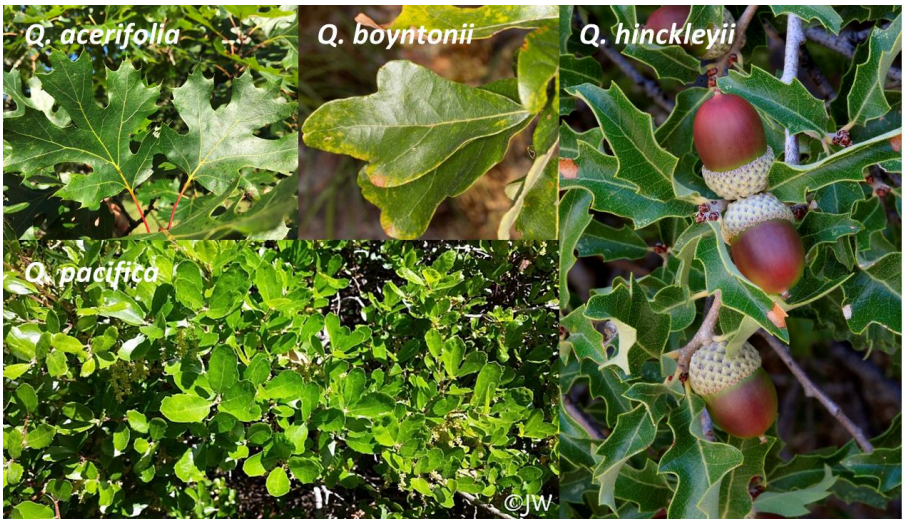
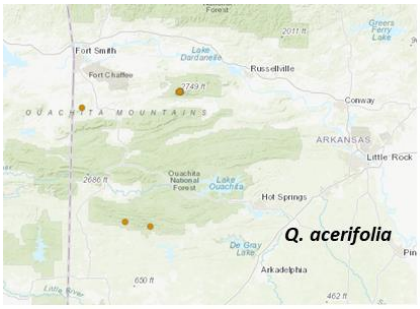
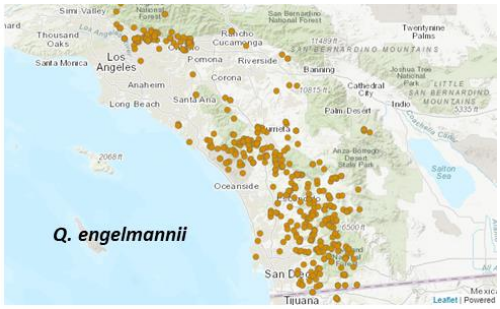


Figure 1: Photos of several species used for simulation.





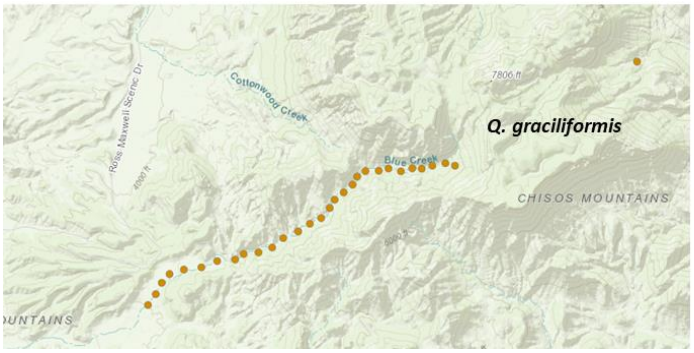


Figure 2: Geographic ranges of Q. acerifolia, Q. engelmannii, and Q. graciliformis (obtained from https://iucnredlist.org)

Table 1: A subset of the parameter values used for several species simulations, to demonstrate the substantial range in terms of number of populations, total number of individuals and differentiation across threatened oaks.

	Q. arkansana	*Q. boyntonii*	*Q. engelmannii*	*Q. georgiana*	*Q. graciliformis*	*Q. pacifica*
Number of populations	10	8	4	5	3	3
Total size	3460	1035	20000	500	325	13000
Migration rates	0.002	0.001	0.002	0.002	0.005	0.001

C. Sampling

To simulate 'sampling,' we wrote code in R to randomly select some number of individuals from each species' population. This simulates sampling individuals from the wild to form an ex situ collection. We tested a broad range of sample sizes, sampling from 1 individual to 500 individuals for every replicate of every species. For species with total population sizes smaller than 500, we stopped sampling once the total population size was reached.

D. Analyses

To assess and compare the genetic diversity captured for each species, we calculated the proportion of alleles captured by each sample size (from 1 to 500 individuals), and averaged this across all 1000 replicates. The proportion of alleles captured represents genetic conservation success; a higher proportion of alleles captured in the sample means a more genetically diverse collection can be created. Then, we calculated the minimum sample size required to capture 95% of the total alleles for each species, a common threshold for describing 'sufficient' genetic diversity.

III. Results

The realism of each species simulation was measured by Nei's pairwise Fst. We compared values to empirical data from previous work that was available (6 species). Otherwise, we used the logic that species with smaller population sizes and lower migration rates should be

more genetically differentiated and have higher Fst values. Table 2 shows compared mean Fst values of several simulated species and measured Fst values from empirical studies.

Table 2: Comparison of mean Fst values from several simulated species to measured, empirical Fst values.

	Q. acerifolia	*Q. boyntonii*	*Q. georgiana*	*Q. oglethorpensis*	*Q. pacifica*
Simulated Fst mean (range)	0.130 (0.049-0.214)	0.061 (0.009-0.165)	0.185 (0.106-0.275)	0.164 (0.06-0.278)	0.013 (0.009-0.016)
Measured Fst range	0.073-0.125	0.015-0.043	0.039-0.062	0.052-0.074	0.016-0.0272

The results of sampling for all 14 species is shown in Figure 3. To reiterate, we sampled from 1 to 500 individuals for each simulation replicate for each species, then averaged the proportion of alleles captured across replicates to get a smooth curve. Of course, as the sample size increases, the proportion of alleles captured (i.e., the genetic conservation success) also increases. However, past a certain threshold, there are diminishing returns. That is, the proportion of alleles captured drastically increases as the sample size increases to about 100 individuals (see the steepness of the curves), but past that point, the proportion of alleles captured increases less dramatically as the sample size increases (curves begin to level off). Eight of the 14 species lie in the middle of the plot (between about 100 and 175); however, there is still a great amount of variation between all species of oaks. In fact, the range of the minimum sample size required to capture 95% of the total alleles spans from 29 individuals for *Q. pacifica*, to 346 individuals for *Q. arkansana* (Table 3).

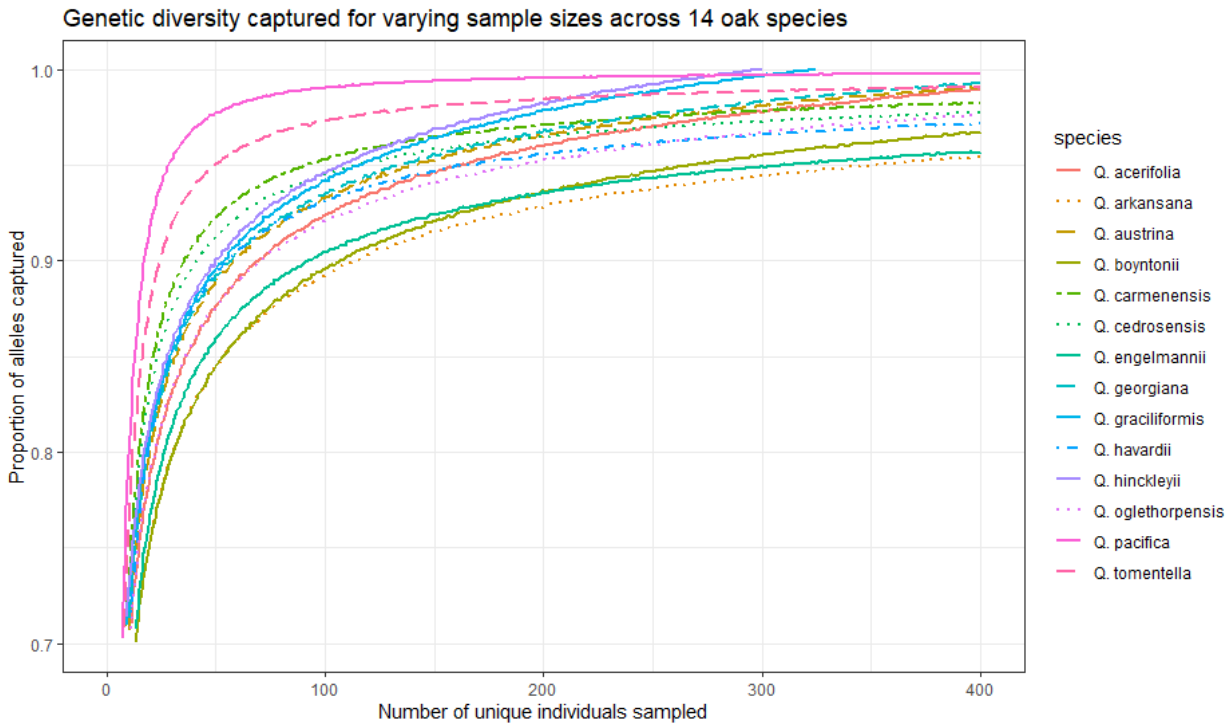


Figure 3: Results for sampling for 14 species of threatened US oaks. The x-axis shows the number of individuals sampled, from 1 to 500. The y-axis shows the proportion of alleles captured, which represents the genetic conservation success. Each species of oak we simulated is represented by a different color line. Note that some species have total population sizes less than 500, so we stopped sampling when the total population was sampled. The proportion of alleles captured across replicates for a given sample size were averaged.

Table 3: Minimum sample size required to capture 95% of the total alleles present in wild populations for several simulated oak species.

Species	Minimum sample size
Q. acerifolia	162
Q. arkansana	346
Q. austrina	140
Q. boyntonii	267
Q. carmenensis	92

Q. cedrosensis	116
Q. engelmannii	310
Q. georgiana	135
Q. graciliformis	116
Q. havardii	163
Q. hinckleyii	108
Q. oglethorpensis	189
Q. pacifica	29
Q. tomentella	50

IV. Discussion

A. Overview

An important step in preventing the extinction of threatened and rare plant species is to create genetically diverse collections. Due to climate change, habitat loss, and other related factors, many plant species are becoming vulnerable to extinction. Genetically diverse collections protect species from extinction (Spielman et al., 2004) and allow organisms to respond to environmental change, aiding in future conservation and restoration efforts for species. In this study, we focused on the genus *Quercus* (oaks), due to their high ecological importance, recently completed Red List assessments (which include geographic information) (Carrero et al., 2020), and intriguing results from Hoban et al. about this genus (2020). For this study, we aimed to recommend minimum sample sizes to capture the diversity of 14 IUCN Red List threatened oak species, and to answer the question: "Can one sampling guideline be recommended to capture the diversity of all threatened US oaks?"

We found several patterns that were consistent among all species of oaks. As the sample size increases, the proportion of alleles captured also increases for all species. However, past a sample size of about 200 individuals for most species, the curves flatten, indicating that there is little benefit to increasing the sample size beyond that size. For example, a sample size of 100 individuals captures nearly 100% of the diversity for *Q. pacifica*. Though there were some

consistent trends across species, we were unable to determine a sample size that could capture the diversity of *all* oaks.

B. Can one recommendation fit all threatened oaks?

Prior work by Hoban et al. determined that one sample size (80 individuals) could effectively capture the diversity of oaks (2020). Here, we aimed to expand on this finding and to compare the results using a simulation study. We found that one minimum sample size could not be recommended to fit all species of oaks we studied. The minimum sample size required to capture 95% of the alleles of a species ranges from approximately 30 to over 340 individuals. However, it should be noted that many species do lie in the middle of this range, with a sample size of around 160 individuals effectively capturing the species diversity for 8 species. Interestingly, this minimum size is higher than found in Hoban et al., 2020 (80 vs. 160). In Hoban et al., *Q. boyntonnii, Q. georgiana*, and *Q. oglethorpensis* were the three species of oaks that were studied, and the minimum sample size recommended for each species was 82, 74, and 76 individuals respectively. For the same species in our study, we obtained much higher minimum sample sizes of 267, 135, and 189 individuals respectively--however, it does follow the same pattern with *Q. boyntonnii* requiring the largest sample size, and *Q. georgiana* requiring the smallest sample size. As Hoban 2019 explained, simulations may contain more alleles than real empirical datasets and thus may lead to higher recommended sample sizes.

There are a few differences between our study and the Hoban et al. study that could explain the discrepancy in results. In the study by Hoban et al., only 3 species of oaks were studied. For this project, we simulated 14 species of oaks, so there is likely more variation between these species in terms of number of populations, distribution, mode of reproduction, population history, and more. In fact, Hoban et al. noted that all 3 oak species in their study were rare, wind-pollinated habitat specialists (2020). All of our species are described by the IUCN Red List as threatened; however, some species, like *Quercus engelmannii,* have very large populations, and others have relatively large geographic distributions. Furthermore, we used simulations in our study while Hoban et al. (2020) based their results on empirical data. We aimed to make our simulations as realistic as possible, but simulations will always represent idealized and simplified versions of the real world and real processes. We note that we have

simplified our species and made estimations about their distributions, especially in cases when little information is known (eg., *Quercus austrina*).

The results of our study reiterate the hypothesis that species-tailored guidelines provide the most effective and efficient method of capturing the genetic diversity of a species. Hoban et al. also state that for most species, taxonomic similarity is not predictive of the minimum sample size required (2020). Thus, wherever possible, sampling guidelines should be tailored at the species level in order to best capture the genetic diversity of the species.

C. Caveats

As previously mentioned, simulations represent simplified versions of real processes. These tools allow us to study processes that would be difficult or unfeasible to study empirically (e.g., by allowing for large numbers of replicates of simulations). For this study and other simulation studies, it is assumed that simulations adequately represent real processes. However, we did make some simplifications in our simulations, such as decreasing the population sizes by a factor of 10 for some species and by assuming equal, constant migration across all populations of the species. Of course, in reality, these processes are much more complex and can change over time.

Furthermore, our sampling strategy was a simplified version of how seeds are sampled in reality. We assumed that 1 seed is sampled from 1 maternal plant, and that sampling occurs at all populations randomly. It is generally recommended that collectors sample from as many unique plants as possible (Hoban, 2019; Hoban et al., 2020). In reality, it may not be possible for collectors to reach all populations, or few individuals within a population may be reproducing during the time sampling occurs. Therefore, multiple seeds are typically harvested from every maternal plant, which could result in decreased genetic diversity capture of the sample compared to sampling all unique plants, as in our simulations.

D. Conclusion

Using information about the species of interest can lead to a more informed sample strategy resulting in a more genetically diverse collection. Here, we determined that one sample size does not fit all IUCN Red List threatened oaks. However, we did determine minimum

sample sizes required to sufficiently capture the diversity of each species individually. These sampling guidelines can directly contribute to conservation efforts for these threatened species of oaks.

References

Brown, A. H. D. and Marshall, D. R. (1995). A basic sampling strategy: theory and practice. Collecting plant genetic diversity: technical guidelines. *CAB International, Wallingford*, pp. 75-91.

Carrero, C., Jerome, D., Beckman, E., Byrne, A., Coombes, A. J., Deng, M., Gonzalez Rodriguiez, A., Van Sam, H., Khoo, E., Nguyen, N., Robiansyah, I., Rodriguez Correa, H., Sang, J., Song, Y., Strijk, J., Sugau, J., Sun, W., Valencia-Avalos, S., and Westwood, M. (2020). The Red List of Oaks 2020. *The Morton Arboretum, BGCI.*

Des Roches, S., Pendleton, L. H., Shapiro, B., and Palkovacs, E. P. (2021). Conserving intraspecific variation for nature's contributions to people. *Nature Ecology and Evolution.*

Excoffier, L. (2003). SIMCOAL: A general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. http://cmpg.unibe.ch/software/simcoal/

Godefroid, S., Piazza, C., Rossi, G., Buord, S., Stevens, A-D., Aguraiuja, R., Cowell, C., Weekley, C. W., Vogg, G., Iriondo, J. M., Johnson, I., Dixon, B., Gordon, D., Magnanon, S., Valentin, B., Bjureke, K., Koopman, R., Vicens, M., Virevaire, M., and Vanderborght, T. (2011). How successful are plant species reintroductions? *Biological Conservation*, 144(2), pp. 672-682.

Griffith, M. P., Calonje, M., Meerow, A. W., Francisco-Oretega, J., Knowles, L., Aguilar, R., Tut, F., Sanchez, V., Meyer, A., Noblick, L. R., Magellan, T. M. (2017). Will the same ex situ protocols give similar results for closely related species? *Biodiversity and Conservation,* 26, pp. 2951-2966.

Griffith, M. P., Calonje, M., Meerow, A. W., Tut, F., Kramer, A. T., Hird, A., Magellan, T. M., and Husby, C.E. (2014). Can a botanic garden Cycad collection capture the genetic diversity in a wild population? *International Journal of Plant Sciences,* 176(1), pp. 1-10.

Guerrant, E. O., Fiedler, P., Havens, K., and Maunder, M. (2004). Revised genetic sampling guidelines for conservation collections of rare and endangered plants: supporting species survival in the wild. *Island Press,* pp. 419-441.

Hoban, S. (2019). New guidance for ex situ gene conservation: Sampling realistic population systems and accounting for collection attrition. *Biological Conservation,* 235, pp.199-208.

Hoban, S., Callicrate, T., Clark, J., Deans, S., Dosmann, M., Fant, J., Gailing, O., Havens, K., Hipp, A., Kadav, P., Kramer, A., Lobdell, M., Magellan, T., Meerow, A., Meyer, A., Pooler, M., Sanchez, V., Spence, E., Thompson, P., Toppila, R., Walsh, S., Westwood, M., Wood, J.,

and Griffith, M. (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).

Hoban, S. and Schlarbaum, S. (2014). Optimal sampling of seeds from plant populations for ex-situ conservation of genetic biodiversity, considering realistic population structure. *Biological Conservation,* 177, pp. 90-99.

Hoban, S. and Strand, A. (2015). Ex situ seed collections will benefit from considering spatial sampling design and species' reproductive biology. *Biological Conservation,* 187, pp.182-191.

Kenny, L., Wenzell, K. & Jerome, D. (2020). Quercus ajoensis (amended version of 2017 assessment). *The IUCN Red List of Threatened Species 2020*: https://dx.doi.org/10.2305/IUCN.UK.2020-2.RLTS.T194050A171680879.en.

Marshall, D. R. and Brown, A. H. D. (1975). Optimum sampling strategies in conservation. International Biological Programme 2: Crop genetic resources for today and tomorrow. *Cambridge University Press.*

Morikawa, M. K. and Palumbi, S. R. (2019). Using naturally occurring climate resilient corals to construct bleaching-resistant nurseries. *PNAS,* 116(21), pp. 10586-10591.

Reusch, T. B. H., Ehlers, A., Hammerli, A., and Worm, B. (2005). Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *PNAS,* 102(8), pp. 2826-2831.

Reynolds, L. K., McGlathery, K. J., and Waycott, M. (2012). Genetic Diversity Enhances Restoration Success by Augmenting Ecosystem Services. *PLOS One*, 7(6), pp. E38397.

Spielman, D., Brook, B., and Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *PNAS,* 101(42), pp. 15261-15264.

Stange, M., Barrett, R. D. H., and Hendry, A. P. (2020). The importance of genomic variation for biodiversity, ecosystems and people. *Nature Reviews Genetics,* 22, pp. 89-105.

Westwood, M., Cavender, C., Meyer, A., and Smith, P. (2020). Botanic Garden Solutions to the plant extinction crisis. *Plants, People, Planet,* 3(1), pp. 22-32.