Conservation Genetics (2026) 27:3
https://doi.org/10.1007/s10592-025-01730-1

RESEARCH

Genetic marker type impacts *ex situ* conservation minimum sample size estimates and their variance

Austin C. Koontz[1] · **Gavin D. Salas**[1] · **Sean M. Hoban**[1]

Received: 29 October 2024 / Accepted: 21 October 2025

Abstract

Ex situ collections in botanic gardens safeguard plant species and their genetic diversity. While past research has typically used microsatellite markers to quantify the extent of *ex situ* genetic representation in botanic gardens, next-generation sequencing approaches generating thousands of single nucleotide polymorphisms (SNPs) have become more common. Several studies have examined the impact marker choice has on measures of genetic diversity and differentiation, but no evaluation on *ex situ* conservation metrics has been made. Additionally, minimum sample size estimates (MSSEs) for representing proportions of genetic diversity are typically reported as averages, but no study has quantified the variance surrounding MSSEs. To close these gaps, we simulated microsatellite and SNP data for species with various demographic characteristics and examined the impact different markers have on MSSEs and their variance. We found that using SNPs resulted in MSSEs twice as large as estimates based on microsatellites, and differences between SNP-based and microsatellite-based MSSEs were greater when total population sizes were larger. We also found that confidence intervals surrounding MSSEs are large, but can be decreased by using higher numbers of loci. Our results indicate that *ex situ* conservation metrics are sensitive to the marker type used, the number of markers, and the total size of wild populations, and that caution is needed when interpreting MSSEs based on empirical datasets that are small relative to a species' total population size. We emphasize that the standard of conserving 95% of genetic diversity *ex situ* should be revisited, and that future minimum sample size recommendations to practitioners should incorporate uncertainty and account for the genetic marker being used.

Keywords Microsatellites · SNPs · *Ex situ* conservation · Simulations · Sampling

Introduction

One method for safeguarding species from extinction is *ex situ* conservation, where organisms are protected outside of their native range. Botanical gardens, arboreta, seed banks, zoos, and aquaria can help conserve rare or threatened species (Griffith et al. 2019; Westwood et al. 2021), thereby preventing extinction, safeguarding genetic diversity, educating the public, and sometimes producing offspring for reintroduction. Botanic gardens and arboreta (hereafter simply referred to as "botanic gardens") can intentionally design collections that represent and maintain genetic diversity from wild (in situ) populations, thereby conserving the existing evolutionary material that allows for species to adapt to threats and respond to environmental change (Guerrant et al. 2014). Because of the limited space, time, expertise, and financial resources available, such designs need to be efficient. DNA data from *ex situ* and in situ individuals can be used to quantify the level of wild genetic diversity currently represented *ex situ* (Hoban et al. 2020). This approach can answer two related questions: what proportion of wild genetic diversity is already conserved in botanic gardens; and how many individuals should gardens be sampling in order to conserve a given threshold of wild genetic diversity. A typical (if somewhat arbitrary) standard is a threshold of representing 95% of wild alleles *ex situ* (Marshall and Brown 1975). We term the number of individuals that need to be sampled in order to achieve this genetic diversity threshold the minimum sample size estimate (MSSE), which can be practically interpreted as

✉ Austin C. Koontz
akoontz@mortonarb.org

[1] Center for Tree Science, Morton Arboretum, 4100 IL-53, IL 60532 Lisle, USA

Published online: 13 November 2025

the recommended number of samples needed to establish a genetically representative *ex situ* collection.

Historically, studies comparing the genetics of garden and wild plant collections have predominantly relied on microsatellite markers (Guichoux et al. 2011; Wei and Jiang 2020); however, microsatellites are known to suffer from issues caused by their unique mutation processes, such as increased risk of homoplasy and null alleles (Putman and Carbone 2014). Additionally, the limited number of microsatellite loci sequenced per individual leads to limited genomic representation, and a subsequently limited level of statistical power for resolving fine scale genomic differences (Zhang and Hewitt 2003). Over the past decade, single nucleotide polymorphisms (SNPs) have begun replacing microsatellites as the marker of choice for population genetic studies, due to the decreasing costs associated with obtaining SNP data and the greater level of power to resolve genomic differences (Schlötterer 2004). Restriction site-associated DNA sequencing (RADseq) techniques, which allow researchers to assess almost any organism without extensive prior genomic knowledge, along with the advent of whole-genome sequencing (WGS) approaches (Fuentes-Pardo and Ruzzante 2017), have catalyzed this increased utilization of SNP markers (Andrews et al. 2016). Despite this trend, due to high heterozygosity and information content per marker, and an ability to compare to previously collected data, microsatellites have continued to be reliable for assessments of parentage, genetic structure, and summarizing genetic diversity (Hodel et al. 2016), and in many cases microsatellite genotyping remains more scalable for large numbers of samples and more affordable in supply, equipment and personnel costs (Puckett 2017; Hauser et al. 2021).

While prior research has led to a more nuanced understanding of the effect of different genetic markers in population genetics, it remains unclear how marker choice impacts measures of *ex situ* conservation, especially two important metrics: the current percent representation of in situ genetic diversity in *ex situ* collections and the MSSE (typically estimated by iterative resampling of in situ data). Additionally, past research has not evaluated either MSSE variance, which can be practically interpreted as the level of confidence associated with a recommended sampling size, nor the impact on that variance from the number of loci used for calculating genetic representation. Because botanic gardens are often tasked with balancing the preservation of genetic diversity with a range of other priorities for an *ex situ* collection (Bruns et al. 2023), reporting the confidence surrounding MSSEs can enable curators and practitioners to draft informed, reasonable sampling targets. Similarly, due to a widening array of sequencing and genotyping options becoming available for conservation applications (Narum et al. 2013), an understanding of the number of loci needed for generating accurate MSSEs can guide conservation geneticists to choose the appropriate design for future studies. These considerations are linked to the decision of what marker type to use for *ex situ* conservation analyses, since the number of loci available for each marker is one of the primary distinctions between microsatellites (ca. 15–30 loci for a given study) and SNPs (thousands to tens of thousands of loci). The only evaluation of marker type in an *ex situ* conservation context that we know of is Koontz et al. 2024; which utilized microsatellite and SNP data from two rare oak species and found that minimum sample size estimates were similar across markers for one species and slightly larger using SNP markers for the other species. This study was limited because it only used empirical data from two species of the same genus, and the sample sizes were rather small relative to each species' census size, meaning that the full in situ genetic diversity was unknown and likely underestimated.

In contrast to empirical analyses, which are typically constrained by limited time frames and costs, simulations are uniquely suited for exploring the effects of different marker types, marker numbers, and sampling strategies in a population genetic context (Hoban 2014; Lotterhos et al. 2022). This is especially true for studies of *ex situ* conservation, because simulations allow researchers to generate minimum sample size estimates based on all alleles found in an entire population, while empirical studies are subject to undercounting alleles and mistaking random error as genuine genetic polymorphism (Hoban et al. 2018; Rosenberger et al. 2021). More generally, whereas the characteristics of real world populations (genetic structure, unique mating systems, demographic history, etc.) cannot fully be accounted for in most empirical systems, simulations allow researchers to model such properties and query the importance of all included variables in an affordable, replicable, and statistically tractable way. For instance, simulation approaches have been used to evaluate bottleneck tests and other forms of demographic inference (Chikhi et al. 2010), determine the power required for detecting adaptive loci (Lotterhos and Whitlock 2015; Forester et al. 2018), and assess the effectiveness of genetic monitoring (Landguth and Cushman 2010; Hoban et al. 2014).

In this manuscript, we compare measures of *ex situ* genetic conservation using microsatellites and SNP markers generated from simulations in which we vary the following parameters: total population size, number of populations, migration rate, and genetic marker type. The aims of our study are:

- Determine if minimum sample size estimates (MSSE), and the variance surrounding those estimates, differ

when calculated using different genetic markers, and how this impact may depend on population size.
- Quantify how the variance surrounding MSSEs is impacted by the number of loci used for both marker types.
- Demonstrate MSSE variance measurements for a practical case study of a realistic simulated dataset representing population sizes and history of 14 IUCN Red List Threatened U.S. oak species.
- Synthesize these results to provide recommendations to future studies seeking to provide *ex situ* conservation guidance, and how to interpret the genetic data used to form that guidance.

Methods

Overview of simulations

To examine the impact of genetic markers on measures of *ex situ* conservation, and how this varies under different demographies, we simulated microsatellite and SNP data by running the software fastSimcoal (Excoffier and Foll 2011) in an R environment using the package strataG (Archer et al. 2017). The fastSimcoal software utilizes a coalescent, backwards-in-time approach to generate genealogies for a given set of demographic and genetic parameters (outlined below). We term each unique combination of parameter values a "scenario" (see Table S1). After exploring and optimizing parameters, we ran 5 simulation replicates of each scenario, to account for variation between simulation outputs for a single set of parameters. The output of each simulation replicate, for each scenario, was a genotype matrix (in the form of a genpop file), which was passed onto resampling and bootstrapping analyses (see Fig. 1 below). All code for creating simulations and running analyses is available on *GitHub*.

Demographic parameters

To determine whether and how demography alters the impact of genetic markers on measures of *ex situ* conservation, we analyzed 10 demographic scenarios for each marker type, chosen to reflect an idealized rare plant species with realistic ranges for each parameter value. We simulated all possible combinations of total species census size (1,200 individuals, or N1200, and 4,800 individuals, or N4800), number of populations (1, 4, and 16 populations), and migration rates between populations (defined as the probability for any

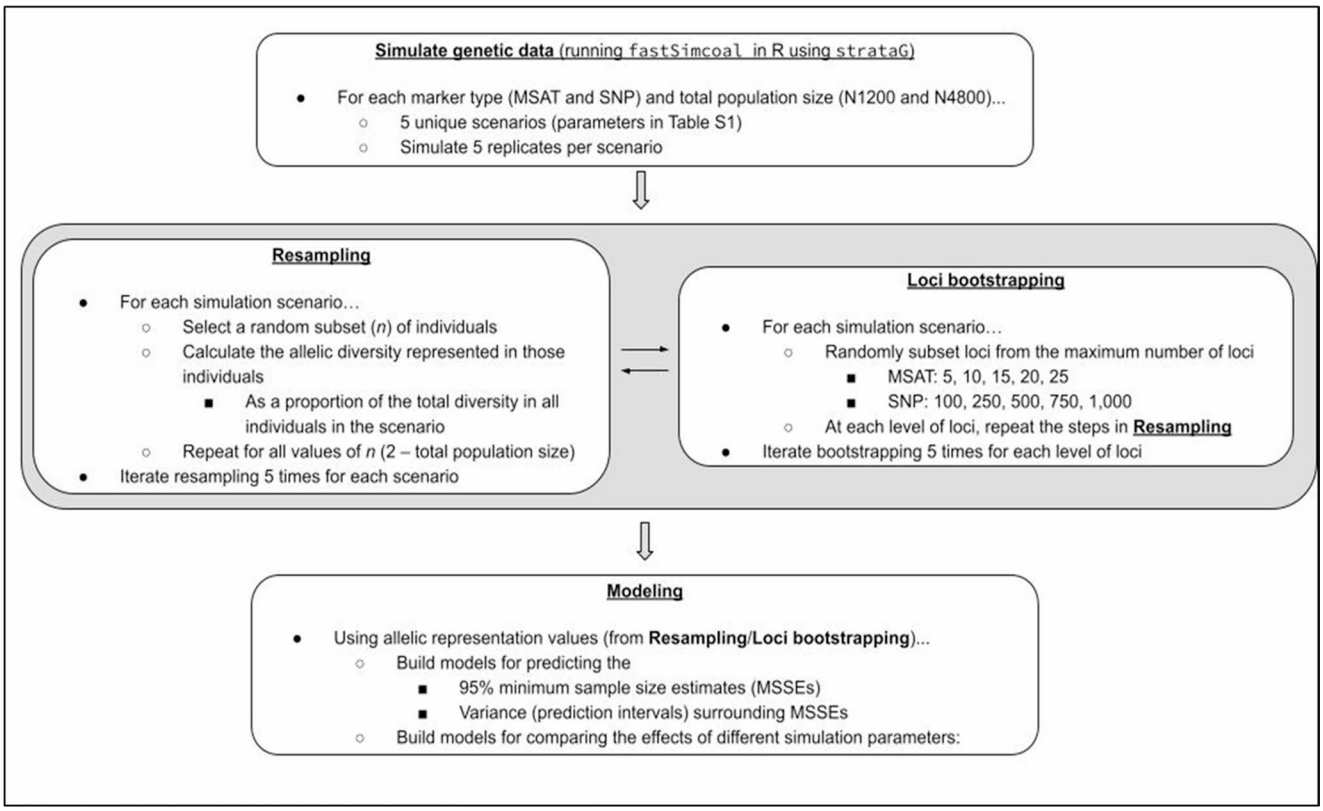

Fig. 1 Workflow of methods. For each marker type (MSAT and SNP) and total population size (N1200 and N4800), 6 scenarios were simulated, with 5 replicates per scenario. The genetic data from each replicate of each scenario was utilized in the **Resampling** and **Loci bootstrapping** steps of the workflow. Allelic representation values generated from the resampling analyses were used as data for modeling. Simulations were generated using fastSimcoal via the strataG R package

given gene to move between populations each generation–"low", or 0.001; and "high", or 0.01; see Table S1). Note that for scenarios with only 1 population, migration rates do not apply. Total species sizes were equally subdivided between populations such that each population had the same number of individuals (i.e. we did not explore scenarios in which populations had different sizes from each other).

Genetic parameters

We simulated two types of genetic data: microsatellites (MSAT; in fastSimcoal, marker type "MICROSAT") and single nucleotide polymorphisms (SNP; in fastSimcoal, marker type "DNA"). These two marker types are defined by different sets of parameters in the fastSimcoal software: the parameter space for both marker types was explored before selecting our finalized parameters to generate data for our *ex situ* conservation analyses.

Establishing realistic parameters: preliminary evaluation of genetic datasets.

We explored different values for fastSimcoal genetic parameters to ensure realistic simulation outputs by comparing the allele frequency distributions of our simulated datasets to those generated for empirical datasets of a rare oak species endemic to the southeastern United States, *Quercus acerifolia* (Maple-leaved oak; Palmer, Stoynoff & Hess; Fagaceae), genotyped using microsatellite and SNP loci generated from a *de novo* assembly (Koontz et al. 2024). We chose *Q. acerifolia* for our empirical comparisons because there's reliable population structure between its 4 native populations, allowing for a more direct comparison to our scenarios simulating 4 equally sized populations. Simulation parameter values which led to the closest match between the shapes of the simulated and empirical allele frequency distributions were selected for our final simulations. We also checked the realism of our simulations by grouping simulated alleles into frequency categories similar to (Hoban et al. 2020) (e.g. common, rare, etc.; see Table S2) and comparing the proportions of each category in simulated and empirical datasets.

In addition to comparing the allele frequency distributions between simulated and empirical datasets, we calculated various summary statistics to confirm that certain assumptions of our simulated datasets were being met. We calculated the average (across replicates) heterozygosity (H_E), fixation index (F_{ST}) and total number of alleles for each scenario. We sought to confirm that H_E values were greater in simulations using microsatellite markers (due to being multiallelic) as compared to demographically equivalent scenarios using SNPs (which are biallelic; DeFaveri et al. 2013; Zimmerman et al. 2020). Additionally, we expected F_{ST} values to be lower in scenarios with a higher migration rate, and that scenarios with more populations tended to harbor a higher number of total alleles (Wright 1984). More details on the genetic parameter optimization specific to each marker type are provided below.

Microsatellites

For microsatellites, we explored two genetic parameters: the mutation rate (defined per locus) and the range constraint, which specifies the maximum number of different alleles allowed per locus. Our initial parameter values (mutation rate: 5e-4; range constraint: 10) generated datasets with lower abundances of rare alleles as compared to empirical datasets. Increasing the range constraint to 15 and the mutation rate to 1e-3 (as in Puechmaille 2016) led to datasets with 25 microsatellite loci and allele frequency distributions more closely matching those seen for empirical datasets, and we used these parameter values for our *ex situ* conservation analyses.

SNPs

For SNP datasets, we based our comparisons on empirical SNP loci from RADseq genotyping which had no greater than 20% missing data (i.e. loci had to be present in at least 80% of the individuals in the dataset in order to be retained). We initially generated SNP datasets specifying 5 chromosomes, 4 linkage blocks per chromosome and 25 bp sequences per block. However, we found that using these values led to datasets with unrealistically low levels of SNP loci (20). Therefore, we increased these parameter values to simulate 100 chromosomes, 10 linkage blocks per chromosome, and 500 bp sequences per block. This led to SNP datasets with 1,000 loci, which is more comparable to empirical SNP datasets.

We then varied the mutation rate (defined per base) to generate SNP datasets with more realistic allele frequency distributions. We began by utilizing a rate of 1e-8 (as used in Linck and Battey 2019). We found that when simulating 1,200 total individuals in a species (N1200), utilizing this mutation rate led to skewed frequency distributions, such that the vast majority of alleles were at fixation (present in all individuals) and relatively few were rare (i.e. less than 1% frequency in a population). However, at larger total population sizes (N4800), rare alleles were more common, illustrating an expected population genetic relationship between mutation rate and population size. To simulate SNP datasets more closely matching our empirical data, we utilized a mutation rate of 1e-7, as we found that this level of mutation led to more realistic frequency distributions, especially for N1200 datasets. In addition, we simulated SNP datasets using the more standard mutation rate of 1e-8

(termed "low mutation" datasets), and passed these through our downstream analyses as well.

Ex situ conservation analyses

After simulating genetic data for both marker types and across different demographic scenarios, we assessed how 95% MSSEs–the number of individuals required to represent 95% of the total wild allelic diversity–varied depending on genetic marker. In practical terms, MSSEs can be interpreted as the recommended number of samples needed to establish a genetically representative *ex situ* collection. We also calculated the variance surrounding those estimates and determined how marker type and marker number affect that variance. To do so, we iteratively resampled our datasets to build arrays of allelic representation values, and then used this resampling data as the input into models, as described below.

Resampling

Similar to Richards et al. 2007; Hoban and Schlarbaum 2014; and similar studies, we used a repeated resampling approach to estimate the proportion of genetic diversity represented by different numbers of individuals from our simulated species. To do so, we randomly subsampled (without replacement) the entire set of individuals, from 2 to the total number of individuals, and for each subsample measured the allelic representation using both genetic marker types. Allelic representation was calculated as the proportion of alleles present in the subsample compared to all the alleles present in the entire species. Allelic representation values were used in modeling to determine the number of samples required for representing 95% of the genetic diversity in the entire species (see "Modeling of MSSE and statistical tests" below).

To account for the stochasticity of random sampling, we performed this resampling analysis five times for each of the five simulation replicates for each scenario (leading to 25 resampling analyses per simulation scenario). For each genetic marker type, we combined the resulting allelic representation values across simulation scenarios into a single data frame in R which was passed onto our models (see below). We built separate data frames for our N1200 and N4800 datasets as resampling analyses are dependent on the total number of individuals included.

Loci bootstrapping

To understand the effect the number of loci used in the resampling process has on MSSE, we bootstrapped across loci levels by randomly subsampling genetic datasets to different numbers of loci. For microsatellites we subsampled from the maximum number of loci (25) to 5, 10, 15, and 20 loci; for SNPs, we subsampled from the maximum number of loci (1,000) to 100, 250, 500, and 750 loci. Allelic representation values were calculated at each loci level for each marker type, for both N1200 and N4800 datasets (as in "Resampling", above). For each loci level, loci were subsampled five times to account for stochasticity in the random selection of loci; allelic representation values were calculated for each random draw of loci at each loci level.

Modeling of MSSE and statistical tests

To generate estimates of the MSSE for each simulated species, we constructed models in which allelic representation (as a percentage) was the explanatory variable and the number of samples corresponding to that level of allelic representation was the response variable. An exponential regression model was used to relate allelic representation to sample number as we found this led to the greatest model fit; summaries of N1200 and N4800 models (using the maximum number of loci for each marker type) can be found in Table S6, and depictions of raw data with fit models can be seen in Figures S4 (N1200) and S5 (N4800). Using the predict function in R with the prediction interval argument, models were used to predict the expected MSSE value for representing 95% of the total allelic diversity, the 95% upper and lower prediction intervals bounding each estimate, and the prediction interval width (the upper interval minus the lower interval). We compared results across microsatellite and SNP datasets to determine how marker type and number of loci impacted the MSSE and the prediction interval width. Because population genetic surveys, especially for species of conservation concern, are often tasked with balancing the number of samples sequenced with the maintenance of a sufficient level of genetic power (Fountain et al. 2016), we wanted to examine how including more loci affected the accuracy of MSSEs and whether such an analysis could provide recommendations for future *ex situ* conservation efforts. We expected that a higher number of loci for either marker type would lead to more precise estimates of MSSE, indicated by smaller prediction interval widths.

Lastly, we compared the impact that demographic parameters had on allelic representation values versus the impact from marker type. To do so, we built separate models in which allelic representation was the response variable. The explanatory variables for these models included the number of samples used in the resampling analyses, the genetic marker, and the demographic parameters varied in simulations: number of populations and migration rate. We compared these models to an identical set which did not include marker type as an explanatory variable, and used an ANOVA

to determine how much more variation was explained by the inclusion of a term for marker type. Both kinds of models were built for N1200 and N4800 datasets.

Test case: 14 red list threatened oak species

To better illustrate the value of MSSE variance in a conservation context, we calculated MSSEs and their prediction intervals for a simulated dataset generated in (Rosenberger et al. 2022) of 14 IUCN Red list Threatened oak species native to the United States (Carrero et al. 2020). Unlike the simulations used for exploring the impacts of marker type and marker number, this dataset utilized species-specific parameter values to simulate genetic data coming from species of conservation concern. Using our novel approach for quantifying MSSE uncertainty on this more realistic dataset allows us to better characterize how variance in MSSE might alter *ex situ* conservation recommendations for threatened species, particularly in the typical scenario where not all individuals of a species can be sampled.

The fastSimcoal software was used to generate genetic data from 20 microsatellite loci for each species, simulated using unique parameters for number of populations, population sizes, and migration rate (Table S7) according to demographic and geographic data sourced from the IUCN Red List and expert consultation. Then, similar to the approach described above, each species was iteratively resampled to generate allelic representation values; however, sample sizes were limited to a maximum of 500 individuals (or fewer, if the total population size of the species was less than 500). The allelic representation values from this resampling process were then used to build models to predict the expected MSSE value for representing 95% of the total allelic diversity, the 95% upper and lower intervals bounding each estimate, and the prediction interval width (as described above). More information can be found in Rosenberger et al. 2022.

Results

We first provide a general summary of the simulations, and then report our results in regards to *ex situ* conservation measures calculated using resampling analyses.

Simulation summaries

Overall we specified 36 different simulation scenarios: 12 using microsatellite markers, 12 using SNP markers with a mutation rate of 1e-7, and 12 using SNP markers with a mutation rate of 1e-8. For each set of 12 scenarios per marker category, 6 scenarios used a total number of individuals of 1,200 (N1200) and 6 scenarios used a total number of individuals of 4,800 (N4800). Specifying 5 simulation replicates per scenario generated 180 genetic datasets for analyses.

Comparisons of simulated data to *Q. acerifolia* empirical datasets showed different levels of similarity for the two marker types. Empirical and simulated allele frequency distributions for microsatellite datasets were largely similar for both N1200 and N4800, although simulated datasets showed fewer alleles at high frequencies (>0.6) than the empirical dataset. For SNPs, similarities to the empirical dataset depended on the total number of individuals and the mutation rate. For N1200 datasets, the higher mutation rate (1e-7) generated distributions more similar to SNP empirical data, while for N4800 datasets, the lower mutation rate generated more realistic frequency distributions. However, for both population sizes, simulated data exhibited a higher number of alleles at fixation and fewer alleles at high frequencies (e.g. <0.8), compared to empirical data.

Summary metrics for N1200 and N4800 simulations are provided in Tables S3 and S4, respectively; summary metrics for low mutation (1e-8) SNP simulations (N1200 and N4800) are provided in Table S5. Simulation summary statistics across marker types and total population size generally conformed to our expectations. Across marker types H_E was lower when using biallelic SNP markers as compared to multiallelic microsatellites. In all cases, F_{ST} values were lower in scenarios with higher migration rates, compared to equivalent scenarios with lower migration rates. For SNP markers, a higher number of populations tended to predict a higher number of alleles present, but this trend was not reliable in microsatellites. Finally, the H_E values in low mutation SNP datasets were an order of magnitude smaller than those in demographically equivalent SNP datasets with a mutation rate of 1e-7, and the number of alleles was around half; however, F_{ST} values between SNP datasets of different mutation rates were comparable.

Ex situ conservation analyses

Resampling

95% MSSEs and upper and lower prediction intervals for both markers and total population sizes are provided in Tables 1 and 2. Estimates are larger for SNP markers and (as expected) larger total population sizes. MSSEs based on SNP data are also more precise: for N1200 and N4800 datasets, PI widths using MSATs are over twice as large as those generated using SNP markers.

Raw allelic representation values (using the maximum amount of loci for each dataset) are shown for each total population size in Figs. 2 and 3, along with the distribution

Table 1 N1200 95% Minimum sample size estimates (MSSEs) and their prediction intervals, for different levels of loci and marker types. MSSE values are rounded to the next whole number of samples

N1200: 95% Minimum Sample Size Estimates			
MSAT		SNP	
Number of loci	*MSSE (Lower PI–Upper PI; PI Width)*	*Number of loci*	*MSSE (Lower PI–Upper PI; PI Width)*
5	373 (127.8–1,086.9; 959.1)	100	637 (463.2–874.0; 410.8)
10	355 (142.0–886.6; 744.6)	250	644 (511.7–809.2; 297.5)
15	352 (150.4–820.9; 670.5)	500	640 (540.7–756.5; 215.8)
20	355 (155.6–810.1; 654.5)	750	644 (559.9–738.8; 178.9)
25	351 (157.1–782.8; 625.7)	1,000	647 (575.0–726.9; 151.9)

Table 2 N4800 95% Minimum sample size estimates (MSSEs) and their prediction intervals, for different levels of loci and marker types. MSSE values are rounded to the next whole number of samples

N4800: 95% Minimum Sample Size Estimates			
MSAT		SNP	
Number of loci	*MSSE (Lower PI–Upper PI; PI Width)*	*Number of loci*	*MSSE (Lower PI–Upper PI; PI Width)*
5	932 (278.4–3,117.2; 2,838.8)	100	2,935 (2,483.9–3,476.6; 983.7)
10	864 (271.1–2,747.3; 2,476.2)	250	2,932 (2,625.2–3,273.9; 648.7)
15	871 (284.3–2,662.7; 2,378.4)	500	2,929 (2,717.8–3,155.8; 438.0)
20	853 (282.3–2,573.8; 2,291.5)	750	2,928 (2,730.6–3,138.2; 407.6)
25	849 (284.4–2,529.2; 2,244.8)	1,000	2,932 (2,750.8–3,123.7; 372.9)

of allelic representation values across simulation parameters for each marker type. Allelic representation curves indicate that saturation (representation of all alleles) occurs with a lower number of samples (meaning the average MSSE is smaller) when using microsatellite markers, but there's a greater distribution of representation values for any given sample size. In contrast, SNP allelic representation curves are more gradual (indicating a higher average MSSE), but the range of allelic representation estimates for any given sample size is more narrow.

Our ANOVAs indicate that genetic marker type explains a significant proportion of the variance in allelic representation for both population sizes (N1200 datasets: $F(1, 359{,}696) = 381{,}281$, $p < 2.2e\text{-}16$; N4800 datasets: $F(1, 1{,}439{,}696) = 1{,}689{,}334$, $p < 2.2e\text{-}16$).

Low mutation SNP datasets have much lower 95% minimum sample size estimates for both total population sizes, compared to SNP datasets utilizing a mutation rate of 1e-7 (Table S8). Similar to the results described above, simulation parameters for SNP low mutation datasets did not impact allelic representation values (Figures S6 and S7); however, lower mutation rates led to allelic representation values being more dispersed, for a single sample size (e.g. compare the scatter plots in Figs. 2 and S6).

Loci bootstrapping

As expected, we see prediction interval width decrease with more loci, indicating higher confidence in MSSE findings with a higher number of genetic markers (Tables 1 and 2). However, this increase in precision is moderate relative to the magnitude of PI widths: even with the maximum number of loci for each marker type, PI widths remain large, particularly for microsatellites.

Test case: MSSE variance for 14 threatened oak species

MSSEs and their prediction intervals for each of the 14 IUCN Red List Threatened U.S. oak species simulated using species-specific data can be found in Table 3. Although values vary according to the species census size and other parameters, all MSSEs are lower than N1200 microsatellite MSSEs, and prediction intervals are less than half as wide, indicating lower MSSE variance.

Discussion

Summary

Using simulations parameterized to reflect allele frequency distributions, we find that the minimum sampling size estimate (MSSE) for representing 95% of the total allelic diversity in a wild population is strongly impacted by marker choice. Additionally, we find the width of prediction intervals surrounding MSSEs to be greater when estimates are generated using microsatellite as opposed to SNP loci, and that increasing the number of loci reliably, but only marginally, decreases the MSSE prediction interval width for both marker types. In all cases, however, prediction interval widths remain large, meaning previously published MSSE recommendations (e.g. Hoban et al. 2020; Rosenberger et al. 2022) should be interpreted with caution when translating MSSE into practical sampling plans. Our analysis of simulations modeled after 14 IUCN Red List Threatened U.S. oaks further supports our main findings, that MSSE variance is large and that the number of individuals sampled impacts MSSEs and their variance. We provide likely explanations underlying the observed difference between markers, compare our findings to those of similar studies,

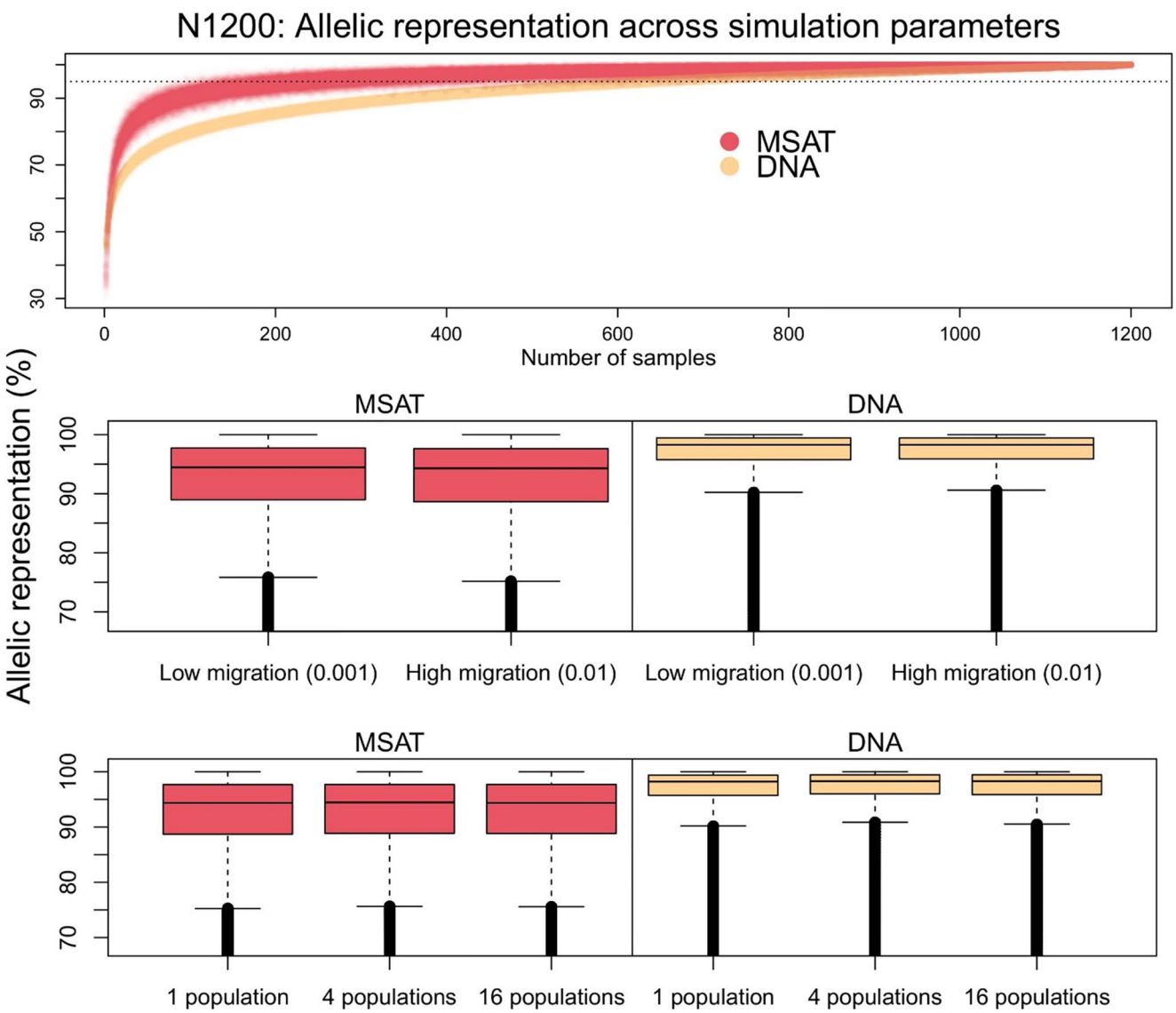

Fig. 2 Distribution of allelic representation values across all sample sizes and simulation parameters (N1200 datasets). Allelic representation percentages are on y-axes. Number of samples used for allelic representation calculations are along the x-axis for the top graph, with the 95% allelic diversity threshold marked by a dashed line; bottom graphs show the different values of simulation parameters: migration rate and number of populations. Pink points and boxes illustrate microsatellite data; yellow points and boxes illustrate SNP data. Values are for datasets with the maximum number of loci (25 microsatellites, 1,000 SNPs); boxplots show results across sample sizes. Note that the y-axis minimum value is never zero and varies among subplots

and consider the significance of these results in the broader context of *ex situ* conservation.

Impact of marker type on minimum sample size estimates

The advent of SNP markers generated using RADseq approaches alongside the continued usage of microsatellites has motivated the need to understand how different markers will impact various population genetic statistics (Fischer et al. 2017; Zimmerman et al. 2020), population clustering patterns (Lemopoulos et al. 2019; Camacho-Sanchez et al. 2020) and the usage of both marker types in landscape genomic (Hodel et al. 2017) and conservation management contexts (Hauser et al. 2021; Osborne et al. 2022). Broadly, this body of work has illustrated that general trends of population diversity and differentiation are similar between markers, but the larger number of loci implied in SNP approaches allows for a greater degree of power to resolve more subtle distinctions at the population and individual level. The analyses described here contribute to this literature by being the first to compare MSSEs using both marker types in a simulated context.

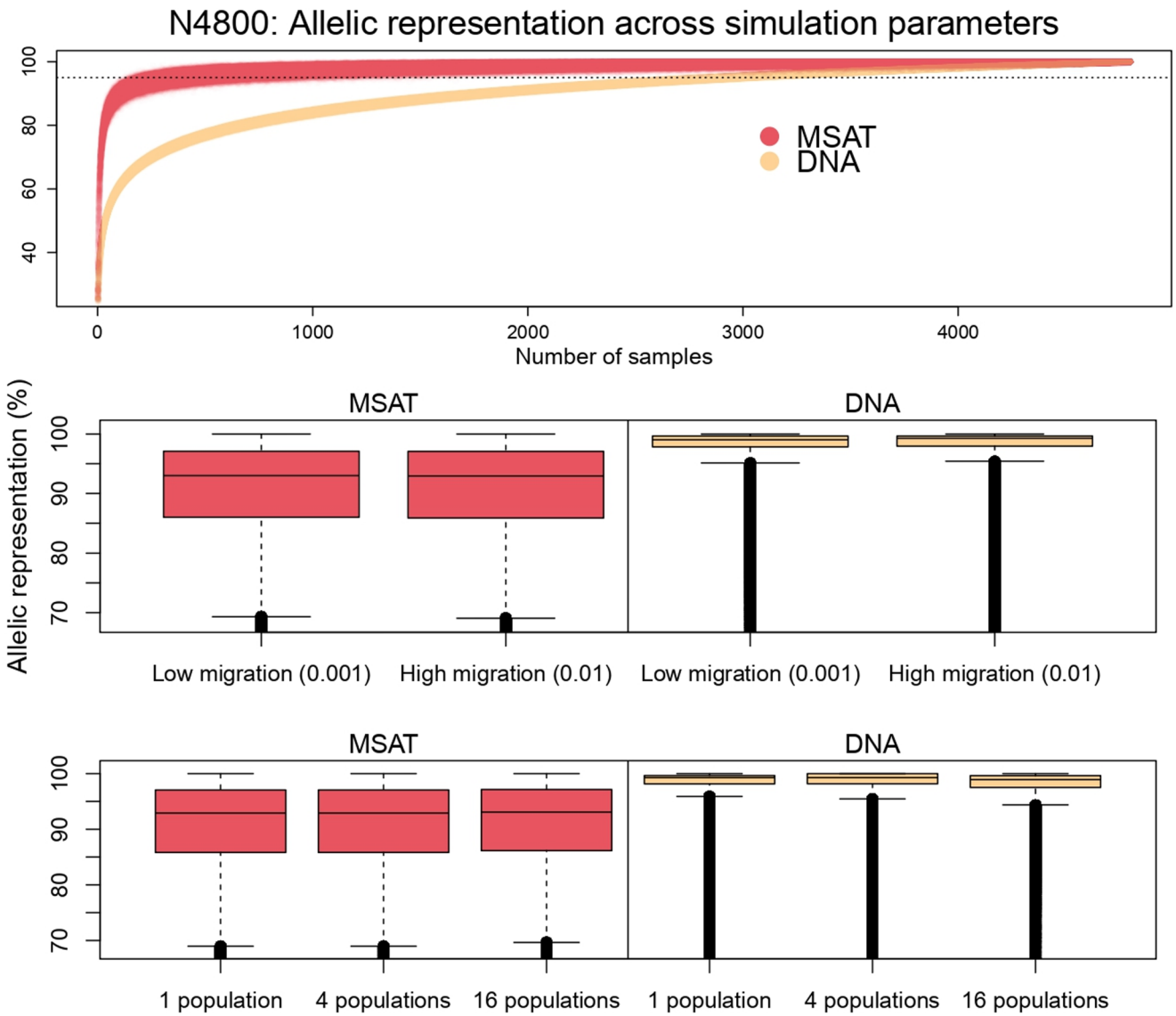

Fig. 3 Distribution of allelic representation values across all sample sizes and simulation parameters (N4800 datasets). Allelic representation percentages are on y-axes. Number of samples used for allelic representation calculations are along the x-axis for the top graph, with the 95% allelic diversity threshold marked by a dashed line; bottom graphs show the different values of simulation parameters: migration rate and number of populations. Pink points and boxes illustrate microsatellite data; yellow points and boxes illustrate SNP data. Values are for datasets with the maximum number of loci (25 microsatellites, 1,000 SNPs); boxplots show results across sample sizes. Note that the y-axis minimum value is never zero and varies among subplots

Our finding that using microsatellites leads to significantly lower MSSEs compared to those estimated using SNPs is likely explained by the mutational profile of each marker. Microsatellites tend to exhibit an increased risk of homoplasy (Selkoe and Toonen 2006; Haasl and Payseur 2011; Putman and Carbone 2014), in which individuals share an allele at a given locus due to mutation, rather than shared descent, leading to an identical state. Specifying an infinite alleles model (rather than our range constraint parameterization of 15) could possibly counteract this by allowing for a greater proportion of rare alleles and subsequently higher MSSEs, but this mutational model has typically been seen as an inaccurate depiction of microsatellite evolution (Slatkin 1995; Putman and Carbone 2014). SNPs, meanwhile, have a mutation rate orders of magnitude lower than microsatellites, which leads to many alleles at lower frequencies and therefore higher MSSEs. This difference in the site frequency spectrum (SFS) between microsatellites and SNPs, particularly in terms of rare alleles, along with much higher numbers of SNP loci (Lockwood et al. 2007), explains the difference between the resulting *ex situ* conservation recommendations using each marker type.

Table 3 95% minimum sample size estimates (MSSEs) for species-specific simulations using microsatellite markers for 14 species of IUCN red list threatened U.S. oaks. MSSE values are rounded to the next whole number of samples. Modeled species census sizes are included; species census sizes were parameterized using IUCN red list data and expert consultation (see Rosenberger et al. 2022). Allelic representation values were determined by sampling a maximum of 500 individuals (or the species census size, if lower)

IUCN Red List Threatened U.S. Oak Species: MSSEs and Prediction Intervals		
Species name	Census size	MSSE (Lower PI–Upper PI; PI Width)
Quercus acerifolia	530	212 (101.3–440.2; 338.9)
Q. arkansana	3,460	426 (207.9–871.6; 663.7)
Q. austrina	540	205 (96.6–432.3; 335.7)
Q. boyntonii	600	301 (146.0–617.5; 471.5)
Q. carmenensis	3,000	305 (132.7–699.7; 567.0)
Q. cedrosensis	8,400	439 (214.3–899.0; 684.7)
Q. engelmannii	20,000	734 (408.3–1,317.8; 909.5)
Q. georgiana	500	194 (89.6–418.1; 328.5)
Q. graciliformis	325	132 (26.2–662.9; 636.7)
Q. havardii	5,150	400 (190.8–836.4; 645.6)
Q. hinckleyi	300	122 (25.6–578.5; 552.9)
Q. oglethorpensis	1,000	251 (111.7–561.8; 450.1)
Q. pacifica	13,000	359 (163.7–786.9; 623.2)
Q. tomentella	10,500	392 (184.9–829.6; 644.7)

Variance around MSSEs

To our knowledge this study is the first to quantify the variance surrounding MSSEs, and to examine the impact that number of loci used in resampling has on those estimates. Across datasets, prediction interval widths decrease as the number of loci increases, illustrating the power that high levels of loci afford to estimates of how many individuals need to be sampled. This same trend explains why SNP MSSEs have much lower variance than those generated using microsatellites, analogous to results showing that larger numbers of SNP loci allow for a greater power in resolving fine scale genetic distinctions (Lemopoulos et al. 2019) and estimating population genetic statistics (Gargiulo et al. 2024). However, even at the maximum number of loci for each marker type, prediction intervals are wide, demonstrating that regardless of the marker used or the level of genetic sampling, the precision of MSSEs based on genetic data needs to be interpreted cautiously (as discussed further in "Caveats and implications" below).

The high variance around MSSEs across all scenarios is likely explained, in part, by the different sources of stochasticity in our analyses: stochasticity in the amount and distribution of genetic diversity within a species (which will vary for each simulation replicate) and stochasticity in the random sampling of individuals used to determine the MSSE. While both sources of stochasticity are partially accounted for by analyzing the results across replicates of our simulations and our resampling analyses, there remains a high variability in the representation of rare alleles within individuals. This, combined with an ambitious target of conserving 95% of alleles–many of which are rare–contributes to the wide prediction intervals observed here.

Comparisons for other *ex situ* conservation analyses

To contextualize past *ex situ* conservation recommendations based on microsatellites, we consider how our findings compare to previous empirical and simulated studies. Simulations offer the ability to analyze a larger number of samples than what's typically feasible to include in an empirical study, and therefore allow researchers to model the full extent of wild populations more accurately. For instance, our analysis of data from Rosenberger et al. 2022 generated a 95% MSSE for *Quercus boyntonii* of 301–comparable to the value of 330 published within the original study (which used a different approach than the predict function to determine MSSE), but much higher than a prior study based on empirical data which estimated a value of 84 (Koontz et al. 2024). Similarly, two recent empirical investigations of the species our generic simulation parameters were modeled after–*Quercus acerifolia*–reported lower MSSEs than our value of 212 reported here: Koontz et al. 2024 (91 wild samples, SNP data) reported an MSSE of 56, while Schumacher et al. 2024 (174 wild samples, microsatellite data) reported an MSSE of 93.

Other empirical studies have focused on estimating sample sizes required for accurately calculating population genetic statistics. For instance, Nazareno et al. 2017 resampled RADseq data to determine the number of individuals and loci required to accurately assess population genetic metrics of isolated populations of *Amphirrhox longifolia*, and found that, when using up to 1,500 SNPs, sample sizes as small as 8 individuals led to reliable estimates of heterozygosity and F_{ST}. An analysis of the invasive whitefly (*Bemisia tabaci*, Qu et al. 2020), found that 1,000 SNPs from as few as 3 individuals could be used to accurately estimate expected and observed heterozygosity values and F_{ST}. The discrepancy between the number of individuals required for sampling in these studies and the MSSEs reported in our analyses stems from the metrics being calculated: rather than measuring allelic representation, these empirical studies calculate fundamental population genetic statistics. Past research has demonstrated that these statistics–such as F_{ST} and heterozygosity–are not strongly impacted by rare alleles, which allows them to be calculated accurately using relatively low sample sizes (Gorman and Renzi 1979; Willing et al. 2012).

Simultaneously, the above citations demonstrate how empirical resampling approaches are restricted by the sampling extent of wild populations (35 individuals per population in Nazareno et al. 2017, p. 10 individuals per population in Qu et al. 2020). Sampling in an empirical study is almost always incomplete, due to limitations in cost and time as well as the possibility that all of a species' populations may not be represented or even known. In contrast, our simulation analyses using multiple marker types allow for a complete sampling of all individuals in all existing populations. Sample size limitations are modeled in our analysis of IUCN Red List Threatened U.S. oaks, which used a maximum sample size of 500 (or the species census size, if lower) for calculating allelic representation; this maximum sample size meant the MSSEs for some species with large population sizes (*Q. engelmannii* and *Q. cedrosensis*) could not be determined in Rosenberger et al. 2022. This methodological difference also explains why MSSEs are so much lower in oak species datasets with census sizes comparable to the total population sizes used in our generic simulations. For instance, our N1200 generic simulations had an MSSE of 502 (using 20 microsatellite loci), whereas the same number of loci used for the *Q. oglethorpensis* simulation (census size 1,000) led to an MSSE of 251. Together, results in both our generic marker datasets and oak species-specific datasets indicate that, in a real population genetic study where only a fraction of all wild individuals can be sampled, any recommended minimum sample size will be smaller than the actual threshold needed to represent a given proportion of all wild genetic diversity, and this impact on MSSEs will be driven largely by the distribution of rare alleles. Furthermore, comparisons across markers demonstrate that this impact from sample size interacts with the effect of using different markers for minimum sample size estimation, as the scale of the difference between microsatellite and SNP MSSEs increases with total population size. The *ex situ* conservation implications of this result is that any empirical MSSE based on only a partial sampling of a species' total population size will tend to be an underestimate, and MSSEs based on very small sample sizes proportional to the species' total population size are likely inaccurate (such as those in Hoban et al. 2020) and should be interpreted cautiously by botanic gardens and conservation practitioners.

Caveats and implications

An important caveat to the results provided here is that our simulation parameters for marker type comparisons led to equally sized populations across all scenarios, which is a major simplification of the complex and idiosyncratic demographics of real world species. Our analyses of simulations modeling 14 Red List Threatened U.S. oaks, as well as results of prior simulation studies (Rosenberger et al. 2021), suggest that, regardless of marker type, MSSEs and their variance will tend to scale with sample size and total population size. Meanwhile, the impacts on MSSEs from more complex demographic histories than those modeled here (e.g. bottlenecks or other significant shifts in population size, population extinction or fragmentation, more complicated migration patterns) would likely be multifaceted, depending on the resulting allele frequency distributions. A related limitation of the analyses here is that our sampling was entirely random and non-spatial, taking neither relatedness nor proximity of sampled individuals into account. Such an agnostic approach goes against sampling guidance for maximizing genetic diversity, which recommends sampling widely and avoiding the collection of closely related individuals (Hoban and Strand 2015). A spatially informed sampling strategy could serve to decrease both MSSEs and their variance, and future analyses could improve upon the results of this paper by simulating individuals in a spatially explicit context and determining if following the best sampling practices impacts *ex situ* conservation recommendations. Finally, while we visually compared empirical and simulated allele frequency distributions to parameterize simulations, a more thorough approach could employ statistical analyses (e.g. a Kolmogorov-Smirnov test).

If SNPs represent the type of genetic polymorphism that conservationists seek to protect *ex situ*, then one of the foremost implications of our analyses would be that recommended minimum sample sizes from past estimates based on microsatellites likely need to be substantially increased; simply put, larger sample sizes are needed if the conservation goal is SNP based variation. However, such an interpretation assumes that *ex situ* representation of every detected polymorphism in every sampled genome of a species is necessary or valuable for that species' long-term survival–a point which relates to the wider discussion of what kind of genetic diversity should be prioritized for conservation. For instance, while it has been maintained that both neutral and functional genomic diversity are important for species conservation (García-Dorado and Caballero 2021; Kardos et al. 2021), some botanic gardens may instead be primarily concerned with protecting known adaptive variation in a species. Such a focus may imply sampling fewer individuals, but would also require a much more extensive knowledge of the species' genomics and the fitness landscape of wild populations (Fragata et al. 2019). Additionally, such a conservation genetic program centered on adaptive variation would likely not utilize microsatellites due to their tendency to be found in non-coding regions of the genome. Meanwhile, if the intent for *ex situ* collections is to represent the maximal amount of all genetic variation found within a species, then even using biallelic SNPs to quantify that representation

may lead to an underestimation of the true number of individuals that need to be sampled. Because nucleotides are not inherited singly but as haplotype blocks (sets of loci linked together due to physical proximity on chromosomes), a collection seeking to conserve all genetic variation would need to consider unique combinations of SNPs rather than individual loci, which prior work has shown will lead to many more character states to represent *ex situ* and much higher MSSE recommendations (Reeves and Richards 2017).

Ultimately, genetically informed sampling recommendations for botanic gardens will depend on the *ex situ* conservation goal (Rossetto et al. 2021). In the absence of genetic data or any principle determining how genetic diversity will be incorporated into conservation planning, however, our results imply that sample sizes need to be increased from those indicated using microsatellite markers for meeting predefined thresholds of genetic diversity. More broadly, the impacts that marker type, marker number, and total population size have on MSSE values in our study speak to the necessity of contextualizing any sampling size recommendations derived from genetic data when communicating with practitioners. In addition to explaining the marker type used to reach a recommended MSSE, including the confidence surrounding that estimate will help geneticists communicate the extent of uncertainty in the number of individuals needed to reach a proportion of genetic diversity, which can help practitioners and managers gauge the likelihood of meeting different conservation genetic targets. However, given the caveats implied by marker type, rare allele frequency, and the size of empirical datasets, an updated paradigm that revisits the ex situ conservation genetics standard of calculating MSSEs based on conserving 95% of alleles (established 50 years ago in Marshall and Brown 1975) is likely called for (see also Lockwood et al. 2007). A straightforward, updated benchmark or heuristic to guide botanic garden collections for the "genomics era" is beyond the scope of this study, but a new framework should acknowledge the complexities and ambiguities inherent in the *ex situ* representation of genetic diversity, including uncertainty surrounding which loci are neutral or selective and whether alleles or haplotypes are the unit of selection. Simultaneously, conservationists and practitioners striving to maximize genetic diversity can continue relying on the well-established recommendations derived from existing research: sample as many populations (Hoban and Schlarbaum 2014) and maternal lines (Schumacher et al. 2024) as possible, sample in a spatially-distributed manner (Hoban and Strand 2015), sample more when pollen dispersal is limited (Rosenberger and Hoban 2024), and sample proportionally according to population size (Rosenberger et al. 2021).

Finally, our results illustrate the advantages of integrating empirical data and simulation-based approaches for assessing population genetic diversity. Empirical research constrained by limited time and resources cannot typically replicate the species-wide sampling simulated in this study (although see Diaz-Martin et al. 2023) which allows for MSSE calculations across all existing individuals, rather than just sampled individuals, and MSSEs based on empirical data are therefore typically underestimates (as noted in Hoban et al. 2020). Simultaneously, simulations often fail to fully represent the demographic and genetic complexity of real species (Hoban et al. 2012; Hoban 2014; Lotterhos et al. 2022) which can only be ascertained by empirical studies. Methods which utilize simulations to model aspects of species biology that are otherwise intractable due to real world limitations, while also relying on empirical data to help tailor recommendations based on unique aspects of a species or population, offer a promising opportunity for future conservation efforts seeking to harness the complete power of genetic data. Geneticists using this variety of approaches can complement the skills and expertise of practitioners, horticulturists, and managers to ensure that botanic gardens and arboreta continue to act as invaluable institutions for the conservation of threatened plants around the world (Westwood et al. 2021).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10592-025-01730-1.

Acknowledgements The authors would like to thank the reviewers for their insights and constructive feedback on this research, as well as Patrick Griffith, Lauren Eserman, Chris Richards, Abby Meyer, Colin Khoury, Jeremie Fant, the members of the Morton Arboretum Global Tree Conservation Program, and the Global Conservation Consortia for valuable discussions on this topic. The authors would also like to thank Kaylee Rosenberger, for her code contributions and guidance.

Author contributions CRediT authorship contribution statementConceptualization: Austin C. Koontz, Sean M. Hoban; Data curation: Austin C. Koontz, Gavin D. Salas; Formal analysis: Austin C. Koontz, Gavin D. Salas; Funding acquisition: Sean M. Hoban; Investigation: Austin C. Koontz, Gavin D. Salas, Sean M. Hoban; Methodology: Austin C. Koontz, Gavin D. Salas, Sean M. Hoban; Project administration: Austin C. Koontz, Sean M. Hoban; Resources: Sean M. Hoban; Software: Austin C. Koontz, Gavin D. Salas; Validation: Austin C. Koontz, Gavin D. Salas, Sean M. Hoban; Visualization: Austin C. Koontz, Sean M. Hoban; Writing–original draft: Austin C. Koontz, Gavin D. Salas, Sean M. Hoban; Writing–review and editing: Austin C. Koontz, Gavin D. Salas, Sean M. Hoban.

Funding This research was funded by IMLS grants MG-251613-OMS-22 and MG-245575-OMS-20, NSF Awards 1759759 and 2216650, and the Morton Arboretum Center for Tree Science.

Competing Interests.

The authors have no relevant financial or non-financial interests to disclose.

Data availability The code used to generate and analyze all marker comparison datasets described above are available in the [GitHub repository](https:/github.com/HobanLab/Morton_SSRvSNP_Simulations). The raw data files used for the marker comparison analyses are available from the corresponding author on reasonable request; the raw data files for the 14 IUCN Red List Threatened oak analyses are included on the [GitHub repository](https:/github.com/HobanLab/Morton_SSRvSNP_Simulations/blob/main/SimulationOutputs/Rosenberger2022_IUCN14Oaks.Rdata).

Declarations

Competing interests The authors declare no competing interests.

References

Andrews KR, Good JM, Miller MR et al (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet 17:81–92

Archer FI, Adams PE, Schneiders BB (2017) StrataG: an r package for manipulating, summarizing and analysing population genetic data. Mol Ecol Resour 17:5–11

Bruns E, Westwood M, Griffith MP et al (2023) Quantifying endangerment value: A promising tool to support curation decisions. https://doi.org/10.24823/sibbaldia.2022.1970. Sibbaldia

Camacho-Sanchez M, Velo-Antón G, Hanson JO et al (2020) Comparative assessment of range-wide patterns of genetic diversity and structure with SNPs and microsatellites: a case study with Iberian amphibians. Ecol Evol 10:10353–10363

Carrero C, Jerome D, Beckman E et al (2020) The red list of Oaks 2020. The Morton Arboretum

Chikhi L, Sousa VC, Luisi P et al (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics 186:983–995

DeFaveri J, Viitaniemi H, Leder E, Merilä J (2013) Characterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. Mol Ecol Resour 13(3):377–392

Diaz-Martin Z, Fant J, Havens K et al (2023) Current management practices do not adequately safeguard endangered plant species in conservation collections. Biol Conserv 280:109955

Excoffier L, Foll M (2011) Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27:1332–1334

Fischer MC, Rellstab C, Leuzinger M et al (2017) Estimating genomic diversity and population differentiation - an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. BMC Genomics 18:69

Forester BR, Lasky JR, Wagner HH, Urban DL (2018) Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. Mol Ecol 27:2215–2233

Fountain ED, Pauli JN, Reid BN et al (2016) Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol Ecol Resour 16:966–978

Fragata I, Blanckaert A, Dias Louro MA et al (2019) Evolution in the light of fitness landscape theory. Trends Ecol Evol 34:69–82

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: advantages, limitations and practical recommendations. Mol Ecol 26:5369–5406

García-Dorado A, Caballero A (2021) Neutral genetic diversity as a useful tool for conservation biology. Conserv Genet 22:541–545

Gargiulo R, Decroocq V, González-Martínez SC et al (2024) Estimation of contemporary effective population size in plant populations: limitations of genomic datasets. Evol Appl 17:e13691

Gorman GC, Renzi J (1979) Genetic distance and heterozygosity estimates in electrophoretic studies: effects of sample size. Copeia 1979:242

Griffith MP, Beckman E, Calicrate T et al (2019) Toward the metacollection: safeguarding plant diversity and coordinating conservation collections. Botanic Gardens Conservation International-US

Guerrant EO Jr, Havens K, Vitt P (2014) Sampling for effective ex situ plant conservation. Int J Plant Sci 175:11–20

Guichoux E, Lagache L, Wagner S et al (2011) Two highly validated multiplexes (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus* spp). Mol Ecol Resour 11:578–585

Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity 106:158–171

Hauser SS, Athrey G, Leberg PL (2021) Waste not, want not: microsatellites remain an economical and informative technology for conservation genetics. Ecol Evol 11:15800–15814

Hoban S (2014) An overview of the utility of population simulation software in molecular ecology. Mol Ecol 23:2383–2401

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

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

Hoban S, Bertorelle G, Gaggiotti OE (2012) Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet 13:110–122

Hoban S, Arntzen JA, Bruford MW et al (2014) Comparative evaluation of potential indicators and temporal sampling protocols for monitoring genetic erosion. Evol Appl 7:984–998

Hoban S, Kallow S, Trivedi C (2018) Implementing a new approach to effective conservation of genetic diversity, with Ash (*Fraxinus excelsior*) in the UK as a case study. Biol Conserv 225:10–21

Hoban S, Callicrate T, Clark J et al (2020) Taxonomic similarity does not predict necessary sample size for ex situ conservation: a comparison among five genera. Proc Biol Sci 287:20200102

Hodel RGJ, Segovia-Salcedo MC, Landis JB et al (2016) The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. Appl Plant Sci. https://doi.org/10.3732/apps.1600025

Hodel RGJ, Chen S, Payton AC et al (2017) Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-seq and investigating loci filtering. Sci Rep 7:17598

Kardos M, Armstrong EE, Fitzpatrick SW et al (2021) The crucial role of genome-wide genetic variation in conservation. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.2104642118

Koontz AC, Schumacher EK, Spence ES, Hoban SM (2024) Ex situ conservation of two rare oak species using microsatellite and SNP markers. Evol Appl 17:e13650

Landguth EL, Cushman SA (2010) Cdpop: a spatially explicit cost distance population genetics program. Mol Ecol Resour 10:156–161

Lemopoulos A, Prokkola JM, Uusi-Heikkilä S et al (2019) Comparing RADseq and microsatellites for estimating genetic diversity and relatedness - Implications for brown trout conservation. Ecol Evol 9:2106–2120

Linck E, Battey CJ (2019) Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Resour 19:639–647

Lockwood DR, Richards CM, Volk GM (2007) Probabilistic models for collecting genetic diversity: comparisons, caveats, and limitations. Crop Sci 47:861–866

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol Ecol 24:1031–1046

Lotterhos KE, Fitzpatrick MC, Blackmon H (2022) Simulation tests of methods in evolution, ecology, and systematics: pitfalls, progress, and principles. Annu Rev Ecol Evol Syst 53:113–136

Marshall DR, Brown AHD (1975) Optimum sampling strategies in genetic conservation. In: Frankel OH, Hawkes JG (eds) Crop Genetic Resources for Today and Tomorrow. Cambridge University Press, 53–80

Narum SR, Buerkle CA, Davey JW et al (2013) Genotyping-by-sequencing in ecological and conservation genomics. Mol Ecol 22:2841–2847

Nazareno AG, Bemmels JB, Dick CW, Lohmann LG (2017) Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. Mol Ecol Resour 17:1136–1147

Osborne MJ, Caeiro-Dias G, Turner TF (2022) Transitioning from microsatellites to SNP-based microhaplotypes in genetic monitoring programmes: lessons from paired data spanning 20 years. Mol Ecol. https://doi.org/10.1111/mec.16760

Puckett EE (2017) Variability in total project and per sample genotyping costs under varying study designs including with microsatellites or SNPs to answer conservation genetic questions. Conserv Genet Resour 9:289–304

Puechmaille SJ (2016) The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. Mol Ecol Resour 16:608–627

Putman AI, Carbone I (2014) Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecol Evol 4:4399–4428

Qu W-M, Liang N, Wu Z-K et al (2020) Minimum sample sizes for invasion genomics: empirical investigation in an invasive whitefly. Ecol Evol 10:38–49

Reeves PA, Richards CM (2017) Capturing haplotypes in germplasm core collections using bioinformatics. Genet Resour Crop Evol 64:1821–1828

Richards CM, Antolin MF, Reilley A et al (2007) Capturing genetic diversity of wild populations for ex situ conservation: Texas wild rice (*Zizania texana*) as a model. Genet Resour Crop Evol 54:837–848

Rosenberger KJ, Hoban S (2024) Simulating pollen flow and field sampling constraints helps revise seed sampling recommendations for conserving genetic diversity. Appl Plant Sci. https://doi.org/10.1002/aps3.11561

Rosenberger KJ, Schumacher E, Brown A, Hoban S (2021) Proportional sampling strategy often captures more genetic diversity when population sizes vary. Biol Conserv 261:109261

Rosenberger KJ, Schumacher E, Brown A, Hoban S (2022) Species-tailored sampling guidelines remain an efficient method to conserve genetic diversity ex situ: a study on threatened oaks. Biol Conserv 275:109755

Rossetto M, Yap J-YS, Lemmon J et al (2021) A conservation genomics workflow to guide practical management actions. Glob Ecol Conserv 26:e01492

Schlötterer C (2004) The evolution of molecular markers–just a matter of fashion? Nat Rev Genet 5:63–69

Schumacher E, Wu Y, Byrne A et al (2024) Examining previously neglected aspects of ex situ gene conservation in two IUCN threatened plant species: rare alleles, redundancy, ecogeographic representativeness, and relatedness. Int J Plant Sci 185:198–209

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett 9:615–629

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Wei X, Jiang M (2020) Meta-analysis of genetic representativeness of plant populations under ex situ conservation in contrast to wild source populations. Conserv Biol. https://doi.org/10.1111/cobi.13617

Westwood M, Cavender N, Meyer A, Smith P (2021) Botanic garden solutions to the plant extinction crisis. Plants People Planet 3:22–32

Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. PLoS One 7:e42649

Wright S (1984) Evolution and the genetics of population: variability within and among natural populations. University of Chicago Press

Zhang D-X, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol 12:563–584

Zimmerman SJ, Aldridge CL, Oyler-McCance SJ (2020) An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. BMC Genomics 21:382