# Environmental and geographic data optimize *ex situ* collections and the preservation of adaptive evolutionary potential

Lionel N. Di Santo[1,2] and Jill A. Hamilton[1]

[1] North Dakota State University, Department of Biological Sciences, 1340 Bolley Drive, Stevens Hall, Fargo, ND, USA

[2] lionel.disanto@ndsu.edu

**Keywords:** variance partitioning, isolation-by-distance, isolation-by-environment, simulations, neutral and functional genetic diversity, plant conservation.

## Abstract

Maintenance of biodiversity, through seed banks and botanical gardens where the wealth of species' genetic variation may be preserved *ex situ*, is a major goal of conservation. However, challenges can persist in optimizing *ex situ* collections where trade-offs exist between expense, effort, and conserving species evolutionary potential, particularly when genetic data is not available. Within this context, we evaluate the genetic consequences of guiding population preservation using geographic (isolation-by-distance, IBD) and environmental (isolation-by-environment, IBE) data for *ex situ* collections where provenance data is available. We use 19 genetic and genomic datasets from 15 plant species to (i) assess the proportion of population genetic differentiation explained by geographic and environmental factors, and (ii) simulate *ex situ* collections prioritizing source populations based on pairwise geographic or environmental distances. Specifically, we test the impact prioritizing sampling based on environmental and

geographic distances may have on capturing neutral, functional or putatively adaptive genetic diversity and differentiation. We find that collectively IBD and IBE explain a substantial proportion of genetic differences among functional (median 45%) and adaptive (median 71%) loci, but not for neutral loci (median 21.5%). Simulated *ex situ* collections reveal that inclusion of IBD and IBE increases both allelic diversity and genetic differentiation captured among populations, particularly for loci that may be important for adaptation. Thus, prioritizing population collections using environmental and geographic distance data can impact genetic variation captured *ex situ*. This provides value for the vast majority of plant species for which we have no genetic data, informing conservation of genetic variation needed to maintain evolutionary potential within collections.

## Introduction

Genetic variation is fundamentally a prerequisite for adaptive evolution (Carlson et al. 2014). Consequently, to maintain species' evolutionary potential, conservation often focuses on the preservation and maintenance of genetic variation. *Ex situ* collections provide one approach to preserve genetic diversity outside species' native ranges. This includes extensive efforts to collect, preserve, and maintain variation across the range of different crop species, wild relatives, and rare or threatened species (Li et al. 2002; Westengen et al. 2013; Naredo et al. 2017). The Global Strategy for Plant Conservation (GSPC) aims to have at least 75% of endangered plant species preserved *ex situ* by 2020 and available for use in recovery or restoration (Target 8; https://plants2020.net/). While significant progress has been made, major gaps remain in the maintenance of genetic variation within collections (Sharrock et al. 2018). Consequently, *ex situ* programs designed to maintain genetic diversity are yet needed.

Traditionally, *ex situ* methods rely on either probabilistic equations (Brown & Marshall 1995; Lawrence et al. 1995), or stochastic resampling using pre-existing genetic datasets to optimize sampling efforts (Caujapé-Castells & Pedrola-Monfort 2004; Gapare et al. 2008). However, these approaches have limitations as they either require the availability of genetic data (population resampling strategy) or make ungeneralizable assumptions of within species population structure (probability-based strategy; Lockwood et al. 2007). More recently, simulation-based strategies have been developed and tested to guide sampling practices (Hoban & Schlarbaum 2014; Hoban 2019). Simulation-based approaches do not require previously published genetic datasets but enable realistic simulations of population structure using available estimates of population size and genetic connectivity. To overcome challenges associated with *a priori* data requirements, the use of surrogate data, such as environmental or spatial data, to estimate neutral and nonneutral genetic variation has received considerable attention (Guerrant Jr et al. 2013; Whitlock et al. 2016; Hanson et al. 2017). Empirical work has focused mainly on testing these data surrogates in preserving genetic diversity *in situ* or in wild populations (Whitlock et al. 2016; Hanson et al. 2017). However, using environmental and geographic data to optimize *ex situ* sampling could have substantial value to conservation.

Evolutionary processes have predictable impacts on the distribution of standing genetic variation, which may be used to guide *ex situ* collections. IBD or "isolation-by-distance" (Wright 1943) arises when gene flow between geographically distant populations is not enough to counteract the accumulation of genetic differences via genetic drift or following successive founder events during colonization (Slatkin 1993; Ledig 2000). In this way, IBD is a proxy for the relationship between pairwise population geographic and genetic distances associated with spatial structure and serial colonization across a landscape. Likewise, IBE or "isolation-by-

environment" (Wang & Summers 2010) describes the accumulation of genetic differences between environmentally distinct populations. IBE predicts that environmental differences are correlated with genetic differences, as selection differs across environments (Keller et al. 2000; Lowry et al. 2008; McBride & Singer 2010), providing a proxy for the relationship between genetic and environmental distance (Dobzhansky 1937; Wang & Bradburd 2014). The influence of geographic and environmental variation in structuring patterns of genetic variation, either independently or collectively, has received extensive support across taxa (summarize in Sexton et al. 2014). Given these observations, spatial and environmental data may provide valuable proxies in designing *ex situ* conservation collections that optimize the preservation of neutral and nonneutral evolutionary processes.

The impact of IBD and IBE on population genetic structure is expected to differ for neutral and adaptive genetic variation (Table 1). This includes the prediction that IBD will have a greater influence at neutral loci relative to IBE. IBD reflects past and current demographic history, as well as the interplay between drift and gene flow in structuring genetic variation, whereas IBE is influenced by natural selection, largely reflecting adaptive genetic variation. Cumulatively, we predict that IBD and IBE will explain the greatest proportion of genetic differences among populations for nonneutral loci. Finally, for those genetic markers underlying functional genetic diversity, including polymorphisms within genes or expressed sequences, we predict patterns of IBE and IBD will be intermediate as they may reflect a combination of adaptive and neutrally evolving loci.

The explosion of genetic and genomic datasets publicly available provides a timely opportunity to compare the contribution of IBD and IBE to genetic structure. In the present study, we compare the influence of genetic marker type on IBD and IBE. We classify single-

sequence repeats (SSRs) and genome-wide single-nucleotide polymorphisms (SNPs) as neutral genetic variation (neutral class), SNPs identified previously as candidate loci for selection using statistical or empirical methods as underlying adaptive genetic diversity (adaptive class), and genetic markers within known genes or expressed sequences (genic SNPs or expressed sequence tag SSRs) as a functional class. We distinguish functional polymorphisms from neutral and adaptive classes as these markers estimate quantitative genetic variation and likely represent a combination of neutral and adaptive processes.

To optimize sampling of genetic variation and differentiation *ex situ*, we have re-analyzed existing genetic and genomic datasets to (i) quantify the impact of IBD and IBE have on population genetic structure across neutral, functional and putatively adaptive genetic datasets, and (ii) to evaluate whether inclusion of IBD and IBE during population sampling influences genetic diversity captured at neutral, functional, and adaptive loci using simulated *ex situ* collections. We use variation partitioning to disentangle the effect of IBD, IBE, their intersection, and union on population genetic structure and then simulate *ex situ* collections using geographic and environmental distance metrics to optimize genetic variation and differentiation conserved. This study advances our understanding of the role non-genetic factors play in the distribution of genetic variation across natural populations, providing new parameters to optimize *ex situ* sampling designs where genomic data may be limited or non-existent.

## Methods

### Source of genetic and geographic data

We searched the Dryad Digital Repository (https://datadryad.org/) to identify genetic or genomics datasets for plant species using three discrete search categories: “Population structure

plant", "SSR population structure" and "SNP population structure". Following this, for inclusion in our study, a dataset or a subset of a dataset had to meet the following criteria:

1. Populations were collected range-wide or were sampled across an isolated fraction of a species' distribution.
2. Geographic coordinates (latitude, longitude) were available for each population sampled.
3. Genetic data, categorized as SSRs (single-sequence repeats), EST-SSRs (expressed sequence tag SSRs) or SNPs (single-nucleotide polymorphism), were available.

Range-wide sampling or sampling of populations spanning a large isolated fraction of a species' distribution were required to ensure the majority of a species' ecological niche space was captured. In addition, sampling a broad range of environmental and geographic distances can reduce the likelihood of covariance between environmental and geographic factors (Wang & Bradburd 2014). Using publicly available databases, population-specific latitude and longitude were used to model climatic variation associated with geographic provenance. These data were used in variation partitioning analyses and to calculate pairwise population environmental and geographic distances for each species. To calculate genetic distances, we included studies using SSRs, SNPs or EST-SSRs. SNP genotyping varied across studies, therefore we divided SNP datasets into two categories: SNPs assessed genome-wide (SNPs) and SNPs assessed within genes (Gen-SNPs). If specific SNPs were identified as being under selection based on previous work, we included a fifth category, SEL-SNPs. Finally, genetic markers were broadly classified as either putatively neutral (neutral class: SSRs, SNPs), underlying functional variation (functional class: EST-SSRs, Gen-SNPs) or putatively adaptive (adaptive class: SEL-SNPs).

Overall, we gathered 17 genetic or genomic datasets, in addition to two genomic datasets received directly from Holliday et al. (2010) (Table 2; Appendix S1). To meet the above criteria,

datasets associated with seven of the 15 studied species were sub-sampled and individual geographic coordinates for one study were averaged to create population-scale coordinates (Table 2; Appendix S2).

**Environmental data**

We used latitude, longitude and elevation associated with population provenance to extract annual, seasonal, and monthly climate variables using ClimateNA (North America), ClimateSA (South America), ClimateEU (Europe) or ClimateAP (Asia Pacific) (https://sites.ualberta.ca/~ahamann/data.html) (Appendix S3). Where elevation was not provided, GPS Visualizer (http://www.gpsvisualizer.com/elevation) was used to assign population elevation values. In total, 80 environmental variables were assigned to each population; including 79 climate-related variables and elevation. For each of the species, all environmental variables associated with population origin were filtered, standardized, and transformed to summarize environmental differences among populations. First, dataset-specific environmental variables exhibiting no population-level variation were excluded from analyses. Environmental variables were then standardized and used to conduct a principal component analysis (PCA). PCA was used to reduce the overall number of environmental variables by summarizing environmental differences across two major axes of differentiation, which together explain more than 70% of environmental variation observed between populations (Appendix S4). These two major PC axes were considered as predictor variables for variation partitioning and used to calculate population pairwise environmental distances in simulations.

**Variation partitioning analysis**

To quantify the contribution of IBD and IBE to genetic divergence within each of the 19 datasets, we conducted a variation partitioning analysis in R (R core Team 2018) using the "vegan" package (Oksanen et al. 2007). We used standard estimates of population genetic differentiation re-calculated for all population pairs within each dataset as our response variable. To account for variation in genetic markers, we used Nei's $F_{ST}$ (Nei 1987), as this metric can provide comparable estimates of population genetic differentiation for both biallelic (e.g. SNPs) and multi-allelic (e.g. SSRs) loci. For each dataset, population divergence was partitioned between two sets of predictor variables; including the geographic coordinates (latitude, longitude) and the two major environmental PC axes (PC1, PC2) associated with each population within a dataset. Following variation partitioning, we conducted a partial distance-based redundancy analysis (dbrda) on each dataset to test the significance of (i) variance explained by each set of predictor variables alone (IBD, IBE; Table 2), and (ii) the variance explained by the union of predictor variables (IBD∪IBE; Table 2). We did not evaluate the significance of the variance explained by the intersection of geographically structured environmental variables (IBD∩IBE; Table 2), as this variance fraction is not testable using dbrda.

**Quantifying the correlation between genetic, environmental and geographic distances**

Geographic and environmental distance between population pairs was measured as the Euclidean distance between populations' geographic coordinates (latitude, longitude) or. between populations' two major environmental PC axes (PC1, PC2), respectively. To visualize and evaluate the covariance structure between genetic, environmental and geographic distance matrices, we graphed and estimated the correlation between all distance metrics (Table 2;

Appendix S5). Correlation coefficients were estimated using the nonparametric mantel test implemented in the R package "adegenet" (Jombart 2008) for each dataset separately.

**Simulating an *ex situ* collection: an idealized framework**

We simulated an idealized *ex situ* conservation collection for each dataset using a customized R script relying on R packages "adegenet" (Jombart 2008), "hierfstat" (Goudet 2005) and "data.table" (Dowle & Srinivasan 2019). This simulation measured the amount of genetic differentiation and the proportion of allelic diversity captured in *ex situ* collections that prioritize population sampling based on environmental and geographic distances. We simulated *ex situ* collections using four different population sampling strategies. This included random sampling, as well as sampling prioritized based on distances between populations' two major environmental PC axes (Euclidean environmental distance), sampling based on distances between populations' geographic coordinates (Euclidean geographic distance) or both (Fig. 1a).

*Ex situ* collections were simulated using between two and the total number of populations available for each dataset (*Np*, Fig. 1a). Randomized sampling sampled populations without replacement from the pool of available populations. Environmentally or geographically prioritized simulations sampled population pairs with the greatest pairwise distances in decreasing order. Collections simulated using the combination of environmental and geographic distances sampled population pairs that exhibited the greatest sum of environmental and geographic distances following standardization, prioritized in decreasing order. All individuals within each population were sampled as part of the idealized simulation.

To compare genetic diversity captured across simulated collections, we estimated two genetic parameters: Nei's $F_{ST}$ and allelic diversity captured ($A_c/A_d$). These indices were chosen as they

quantify different aspects of population genetic diversity. Nei's $F_{ST}$ provides an estimate of genetic differentiation across sampled populations and $A_c/A_d$ provides an estimate of the number of alleles captured in collections ($A_c$) relative to the total number of alleles present within a dataset ($A_d$). All genetic parameters were estimated in R using the "hierfstat" package.

Population sampling and associated genetic summary statistics were simulated 500 times for each dataset to account for the variance introduced through randomly sampling across populations. Summary statistics were estimated based on average values across all 500 simulations. No replication was used for environmental and/or geographic distance-based population sampling, as neither provenance of source populations nor genetic summary statistics would have changed with repeated iterations.

For these idealized simulations, all individuals were sampled within each target population (equivalent to protecting the entire population), regardless of collection strategy, assuming 100% of the standing genetic variation was captured. However, monetary or logistical constraints usually impact the number of individuals that could be sampled within a target population. Given this, we predict that genetic diversity captured within source populations will vary. To assess whether insights gained from idealized simulations were maintained under more realistic conditions, we conducted additional simulations, introducing differences in the amount of genetic diversity captured between populations (hereafter referred to as realistic simulations).

**Simulating an *ex situ* collection: The realistic framework**

To simulate a realistic *ex situ* collection, a subset of individuals was sampled within each population. This provides the opportunity to evaluate the impact varying genetic diversity captured within populations may have on total genetic diversity and differentiation captured

across populations collected. We assume that *ex situ* collections aim to preserve as much genetic variation as possible within each population. Within this framework, we postulated that at least 80% of within-population allelic diversity would be captured *ex situ*. Therefore, for each dataset, we assessed the number of individuals ($N_{80\%}$) that when sampled capture between 80%-100% of allelic diversity across populations.

An additional simulation was used to determine the value of $N_{80\%}$ for each dataset (Fig. 1b). For every population, *N* individuals (ranging from one up to the size of the smallest population within the assessed dataset) were randomly sampled without replacement. Following this, the number of alleles captured for *N* individuals ($A_s$) divided by the total number of alleles in the population ($A_p$) was quantified for each population. Sampling of individuals and quantification of allelic diversity captured was replicated 500 times for each population and value of *N* to calculate confidence intervals around $A_s/A_p$ ratios. The number of individuals required to capture 80% or more (As/Ap ≥0.8) of allelic diversity in every population ($N_{80\%}$) was visually assessed for each dataset independently (Appendix S6) and used to parametrize realistic simulations (Fig. 1a). *Ex situ* collections were simulated 500 times using the realistic scenario to estimate genetic summary statistics regardless of the population sampling strategy used (Fig. 1a). For these simulations, $N_{80\%}$ were often much lower than the existing size of most populations and performing repeated iterations accounted for the variation in genetic summary statistics introduced by small values of $N_{80\%}$.

Maintaining the range of $A_s/A_p$ ratios across datasets is crucial as unbalanced variance may confound the influence of prioritization strategies in downstream analyses. Four of the 19 datasets (*H. argophyllus* (Gen-SNPs), *M. lacinatus* (SSRs), *R. oldhamii* (EST-SSRs) and *S. leprosula* (EST-SSRs)) were discarded from realistic simulations as $N_{80\%}$ values were not

253 reached for these datasets (Appendix S6). These same datasets were also removed from idealized
254 simulations to ensure that differences in summary statistics between idealized and realistic
255 simulations originated solely from variation in allelic diversity captured across populations
256 introduced in the latter. See Appendix S7 for a complete list of parameters tested and used for
257 simulations.
258
259 **Analysis of simulated data**
260 We tested whether prioritizing source population collection using environmental and/or
261 geographic distance data influences genetic variation and differentiation captured *ex situ*. For
262 every number of populations sampled (*Np*), genetic summary statistics simulated using random
263 sampling were subtracted from values based on prioritization strategies using environmental
264 distances, geographic distances, or both. Summary statistics were averaged for each dataset
265 following repeated iterations, grouped by distance-based strategies, genetic marker class, and
266 simulation framework (idealized or realistic) (Fig. 2). Differences in genetic summary statistics
267 are provided based on the proportion of populations sampled as the number of populations
268 sampled for analysis varied across studies. For each dataset, we selected four numbers of
269 populations sampled (*Np*) representing between 30-40%, 50-60%, 70-80%, and 90-100% of
270 populations present in a dataset (Appendix S7).
271 Finally, we fitted a linear model between proportions of populations sampled and differences
272 in genetic summary statistics for every combination of genetic marker class, distance-based
273 prioritization strategy, and simulation framework (Fig. 2). A negative relationship indicates that
274 a given distance-informed sampling generally increases the genetic summary statistics relative to
275 random sampling while a positive relationship would suggest the opposite. In addition, it is

276 important to note that a significant relationship (positive or negative) will always be approaching
277 zero as the proportion of populations sampled increases. This is because with additional
278 populations sourced, the probability that identical populations are sampled randomly or via
279 distance-based strategies increases and will reach one when all populations are sampled. As the
280 number of shared populations between sampling strategies increases, the difference in genetic
281 summary statistics decreases.
282
283 **Results**
284 **Relative contributions of IBD and IBE to population genetic differentiation**
285 Variation partitioning revealed that IBD explained significantly more among-population
286 genetic differences (13%) than IBE alone (5.5%) or IBD∩IBE (3%) for neutral genetic datasets
287 (Table 3). This contrasts with functional and adaptive datasets, where a significant proportion of
288 among-population genetic differences was explained by geographically structured environmental
289 variables relative to environmental or geographic factors alone (Table 3). Overall, 31% and 42%
290 of population genetic differences were explained by IBD∩IBE for functional and adaptive
291 datasets, respectively, while only a small proportion was explained by IBD (functional: 10%,
292 adaptive: 16%) and IBE alone (functional: 2.5%, adaptive:1%).
293 While significant differences in the proportion of genetic differentiation explained were
294 observed across genetic marker classes for IBD∩IBE and IBD∪IBE, no significant differences
295 were observed in the individual contribution of IBD and IBE (Table 3). IBD∪IBE explained the
296 greatest proportion of genetic differences for adaptive genetic markers (71%), followed by
297 functional (45%) and neutral (21.5%) genetic markers, respectively. Interestingly, IBD∩IBE
298 explained substantial among-population genetic differences for both functional and adaptive

datasets but explained limited variation for neutral datasets (Table 3). The contribution of IBD∩IBE to population genetic differentiation for adaptive and functional datasets likely reflect high correlations observed between environmental and geographic distance matrices (Table 2; Appendix S5). Therefore, the relative contribution of geography and environment should be interpreted with caution for these genetic marker classes, as population genetic differentiation could not be partitioned solely by IBD or IBE.

## Genetic diversity and differentiation captured in simulated *ex situ* collections

*Genetic differentiation (Nei's $F_{ST}$)*

Significant negative relationships were observed between proportions of populations sampled and changes in genetic differences ($F_{ST}$) captured for collections simulated using both adaptive and functional datasets, but not neutral genetic datasets (Fig. 2a). This suggests that using environmental and/or geographic distance to prioritize population sampling may potentially increase adaptive and functional genetic differences but does not consistently impact neutral genetic variation. Simulations revealed that using all three distance-based population sampling strategies increased genetic differentiation captured among adaptive loci in *ex situ* collections (Fig. 2a). This contrasts with the results obtained for functional datasets, where sampling prioritizing source populations using environmental distance, or the combination of both environmental and geographic distances increased genetic differences captured.

For both adaptive and functional genetic makers classes, simulations based on realistic and idealized within-population sampling scenarios led to similar slopes, regardless of the distance-based population sampling strategy used (Fig. 2a; Appendix S8). This indicates that the ability of

distance-based population sampling strategies to increase $F_{ST}$ among functional and adaptive loci was not impacted by the within-population sampling scenarios simulated.

*Proportion of allelic diversity captured ($A_c/A_d$)*

Both realistic and idealized *ex situ* collection simulations using functional and adaptive genetic datasets indicated allelic diversity captured ($A_c/A_d$) is likely sensitive to within-population sampling. Prioritizing population sampling using environmental distances increased allelic diversity captured at functional loci under realistic within-population sampling conditions, but had no impact using idealized within-population sampling scenario (Fig. 2b). This contrasts with results obtained for adaptive datasets, where the opposite pattern was observed. Prioritizing population sampling using environmental or the combination of environmental and geographic distances increased $A_c/A_d$ under idealized within-population sampling conditions (Fig. 2b).

For neutral genetic datasets no consistent change in allelic diversity was observed in response to varying proportions of population sampled, regardless of population prioritization strategy tested or within-population sampling scenario simulated (Fig. 2b). Together, these results suggest that incorporating environmental and/or geographic distances to prioritize collections may increase allelic diversity captured at functional and adaptive loci, but not at neutral loci. Nonetheless, simulations also indicate that increasing allelic diversity captured in *ex situ* collections is dependent on within-population sampling scenarios and may thus only be achieved under specific sampling conditions.

## Discussion

Optimizing efforts to conserve genetic variation relies upon an understanding for how non-genetic factors, geographic and environmental variation, contribute to population genetic structure. Here, we leverage population provenance and environmental data to optimize genetic differences captured in simulated conservation collections. Environmental and geographic factors explain some portion of the genetic differences observed among populations, although the extent differs by genetic marker class. The proportion of genetic differentiation explained by IBD∪IBE was significantly higher for adaptive and functional datasets relative to neutral datasets. This suggests that geographic and environmental data may provide a useful guide when designing *ex situ* population sampling, particularly where the goal is to conserve adaptive and functional genetic variation. We simulated *ex situ* sampling and found that, as predicted, strategies that included environmental and/or geographic distance data to prioritize population sampling increased genetic differences and diversity captured at both functional and adaptive loci. Overall, we suggest that inclusion of IBD and IBE in guiding *ex situ* sampling can ensure adaptive and functional genetic variation are conserved, crucial for long-term preservation and maintenance of species' evolutionary potential.

Consistent with previous plant studies, our results demonstrate that genetic differentiation across neutral, functional, and adaptive loci can, at least partly, be explained by environmental and geographic factors (Bjørnstad et al. 1995; Nadeau et al. 2016; Xia et al. 2018) (Table 2). Interestingly, limited genetic differentiation was explained by IBD or IBE alone across all three genetic marker classes. For functional and adaptive datasets, this is likely due to the fact that substantial genetic structure is explained by their intersection (Table 3). Indeed, IBD∩IBE reflects covariance between geographic and environmental factors that cannot be teased apart.

Additional empirical work minimizing this covariance would be required to completely disentangle these factors (Wang & Bradburd 2014). Nonetheless, when combined, environmental and geographic factors explained a substantial proportion of population genetic differentiation for both functional and adaptive datasets (IBD∪IBE; Table 3). This suggests that geographic and environmental differences contribute largely to genetic divergence at nonneutral loci (Huang et al. 2016; Xia et al. 2018). Consequently, the inclusion of IBD∪IBE may provide a means to capture adaptive and functional genetic variation *ex situ*. For neutral datasets, geographic and environmental factors, either individually (IBD, IBD) or cumulatively (IBD∪IBE), explained very small proportions of among-population genetic differences (Table 3). This indicates that stochastic processes, such as genetic drift or founding events likely influence neutral genetic structure. Random fixation or loss of alleles through genetic drift (Stern & Orgogozo 2009) and accelerated allele fixation within populations following demographic changes, including bottlenecks or founder events (Maruyama & Fuerst 1985; Gavrilets & Hastings 1996), may lead to population structure that is not explained by environment or spatial data. Overall, our findings indicate that environmental and geographic distance metrics can be used to target genetic differences which likely reflect adaptive or functional genetic variation over neutral genetic variation.

*Ex situ* strategies relying on existing genetic datasets (Caujapé-Castells & Pedrola-Monfort 2004; Gapare et al. 2008) or genetic simulations (Hoban & Schlarbaum 2014; Hoban 2019) have previously optimized variation captured in collections. These approaches require substantial *a priori* information and target neutral genetic variation. Where knowledge of population location is available, pairwise geographic and environmental distances may be leveraged to extend previous sampling to conserve adaptive and functional genetic variation. Our simulations

388 demonstrate that *ex situ* collections prioritized using environmental or the combination of
389 environmental and geographic distances increase both Nei's $F_{ST}$ and $A_c/A_d$ captured for adaptive
390 and functional datasets relative to random sampling (Fig. 2). This indicates that divergent
391 selection and adaptation to local environments contribute to genetic differentiation at nonneutral
392 loci (Hancock et al. 2011; Wang et al. 2016), likely influenced by IBE. IBE-based prioritization
393 strategies suggest that part of the additional genetic differences captured in collections consist of
394 spatially and/or environmentally restricted alleles (Fig. 2b). However, simulations also revealed
395 that increasing allelic diversity captured in collections using distance-based prioritization
396 strategies depends on within-population sampling conditions (realistic or idealized). These
397 results have important applications to applied conservation efforts. First, a realistic sampling
398 scenario was sufficient to increase genetic differentiation captured at adaptive and functional loci
399 (Fig. 2a). This suggests that inclusion of IBD and IBE in population prioritization would likely
400 increase among-population genetic differences captured at these loci by sampling only a subset
401 of individuals within populations. However, only an idealized sampling scenario increased allelic
402 diversity captured at adaptive loci (Fig. 2b). This indicates that extensive within-population
403 sampling may be needed to increase adaptive allelic diversity conserved in collections. Overall,
404 simulations demonstrate that prioritizing population sampling using IBD and/or IBE can increase
405 genetic differences and diversity captured at both functional and adaptive loci without the need
406 for prior genetic data, providing a means to target genetic variation that may be needed to
407 maintain adaptive potential within collections.

408 Despite the fact conservation has long valued environmental and geographic data (Brown &
409 Marshall 1995; Guerrant et al. 2004; Guerrant Jr et al. 2013), use of these data for conservation
410 planning have only emerged during the past decade (Vinceti et al. 2013; Whitlock et al. 2016;

411 Hanson et al. 2017). Consistent with previous work, we observe inconsistent benefits of
412 leveraging geography for the preservation of neutral genetic diversity (Fig. 2). This could be due
413 to the fact that gene flow between populations may be disturbed by landscape characteristics
414 (Dudaniec et al. 2016), or some species may exhibit greater gene flow between geographically
415 distant populations (O'Connell et al. 2007). Our results do provide additional empirical support
416 for inclusion of environmental and geographic data in conservation planning, to target and
417 increase adaptive genetic diversity conserved (Hanson et al. 2017) (Fig. 2). In addition, this study
418 is the first to provide evidence that IBD- and/or IBE-based population prioritization strategies
419 may increase genetic differentiation and diversity captured at functional loci. This indicates that
420 using environmental and/or geographic surrogates may not only preserve current adaptive
421 genetic diversity but may also secure genetic variation crucial for future adaptations. Finally,
422 where other studies use amplified fragment length polymorphisms (AFLPs; Whitlock et al. 2016;
423 Hanson et al. 2017), we focus on SSRs and SNPs datasets. The concordance across studies
424 suggests a broad applicability for environmental and geographic data to act as surrogates to
425 optimize the conservation of genetic variation.

426 Although simulations are a powerful inferential tool, they can include a number of
427 assumptions. Here, we assumed that maternal plants used in realistic and idealized simulations
428 were collected for storage *ex situ*. However, the progeny of these plants more accurately reflects
429 those likely to be included in collections (FAO, 2010). Future studies will need to consider
430 empirical or simulated progeny data to evaluate whether environmental and/or geographic
431 distance-based prioritization captures genetic variation across generations. In this study, we
432 evaluated the overall impact of population sampling strategies on genetic variation and
433 differentiation captured in *ex situ* collections. Nonetheless, simulations revealed important

434 variation in genetic summary statistics across datasets within genetic marker classes (Fig. 2).
435 This variation is likely introduced by differences in species' life history traits including mode of
436 reproduction and breeding system (Loveless & Hamrick 1984). Despite this variance, our data
437 suggest that inclusion of IBD and IBE in *ex situ* guidelines may still be valuable to optimizing
438 functional and adaptive genetic variation captured. Future work assessing the influence trait
439 combinations may have on predicting genetic variation captured in collections will complement
440 the present research, providing sampling guidelines for species exhibiting specific life history
441 characteristics. Finally, we grouped different genetic markers into genetic diversity classes to test
442 the effect of prioritizing population sampling using environmental and/or geographic data at a
443 broader scale. However, allelic distributions and mutation models largely differ between these
444 genetic markers. Thus, future work should evaluate marker-specific patterns associated with
445 IBD- and IBE-based prioritization strategies.

446 Anthropogenic changes have had substantial impacts on global biodiversity, resulting in a
447 global call for the preservation of biodiversity. This research expands existing *ex situ* population
448 sampling strategies, leveraging geographic provenance and environmental distance to increase
449 functional and adaptive genetic differences conserved in collections. Incorporating an
450 understanding of evolutionary and ecological processes influencing population structure
451 alongside new and existing datasets will be critical to enhancing current conservation practice.

452

453 **Supporting Information**

454 Reference and availability information associated with every genetic and genomic dataset
455 (Appendix S1), modifications applied to genetic and genomic datasets (Appendix S2), raw set of
456 climatic variables used in simulations and variation partitioning analyses (Appendix S3),

457 proportion of variance explained by the two major environmental principal components for each
458 dataset (Appendix S4), covariance between environmental, geographic and genetic distances
459 (Appendix S5), proportion of allelic diversity captured within populations using $N_{80\%}$ or the size
460 of the smallest population within datasets (Appendix S6), a list of tested and used parameters for
461 realistic and idealized simulations (Appendix S7), and regression statistics associated with
462 realistic and idealized simulations (Appendix S8) are available online.
463

464 **Acknowledgments**

465 The authors thank the Hamilton Lab, T. L. Parchman, and S. M. Hoban for their valuable
466 comments on early versions of this manuscript. This work was supported by a new faculty award
467 from the office of the North Dakota Experimental Program to Stimulate Competitive Research
468 (ND-EPSCoR NSF-IIA-1355466) and funding from the NDSU Environmental and Conservation
469 Sciences Program to J.A.H.
470

471 **Literature Cited**

472 Bjørnstad ON, Iversen A, Hansen M. 1995. The spatial structure of the gene pool of a viviparous
473 population of Poa alpina environmental controls and spatial constraints. Nordic Journal of
474 Botany **15**:347–354. Wiley Online Library.
475 Brown AHD, Marshall DR. 1995. A basic sampling strategy: theory and practice. Collecting
476 plant genetic diversity: technical guidelines. CAB International, Wallingford:75–91.
477 Carlson SM, Cunningham CJ, Westley PAH. 2014. Evolutionary rescue in a changing world.
478 Trends in Ecology and Evolution **29**:521–530.
479 Caujapé-Castells J, Pedrola-Monfort J. 2004. Designing ex-situ conservation strategies through

480 the assessment of neutral genetic markers: application to the endangered Androcymbium
481 gramineum. Conservation Genetics **5**:131–144. Springer.
482 Dobzhansky TG. 1937. Genetics and the origin of species. New York City. NY Columbia
483 University Press.
484 Dowle M, Srinivasan A. 2019. data. table: Extension of 'data. frame'. R package version 1.12.6.
485 https://CRAN.R-project.org/package=data.table.
486 Dudaniec RY, Worthington Wilmer J, Hanson JO, Warren M, Bell S, Rhodes JR. 2016. Dealing
487 with uncertainty in landscape genetic resistance models: a case of three co□occurring
488 marsupials. Molecular ecology **25**:470–486. Wiley Online Library.
489 FAO. 2010. The second report on the state of the world's plant genetic resources for food and
490 agriculture. Food & Agriculture Org.
491 Gapare WJ, Yanchuk AD, Aitken SN. 2008. Optimal sampling strategies for capture of genetic
492 diversity differ between core and peripheral populations of Picea sitchensis (Bong.) Carr.
493 Conservation Genetics **9**:411–418. Springer.
494 Gavrilets S, Hastings A. 1996. Founder effect speciation: a theoretical reassessment. The
495 American Naturalist **147**:466–491. University of Chicago Press.
496 Goudet J. 2005. Hierfstat, a package for R to compute and test hierarchical F□statistics.
497 Molecular Ecology Resources **5**:184–186. Wiley Online Library.
498 Guerrant EO, Havens K, Maunder M. 2004. Ex situ plant conservation: supporting species
499 survival in the wild. Island Press.
500 Guerrant Jr EO, Havens K, Vitt P. 2013. Sampling for effective ex situ plant conservation.
501 International Journal of Plant Sciences **175**:11–20. University of Chicago Press Chicago, IL.
502 Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian C,

503 Roux F, Bergelson J. 2011. Adaptation to Climate Across the Arabidopsis
504 thaliana Genome. Science **334**:83 LP – 86. Available from
505 http://science.sciencemag.org/content/334/6052/83.abstract.
506 Hanson JO, Rhodes JR, Riginos C, Fuller RA. 2017. Environmental and geographic variables are
507 effective surrogates for genetic variation in conservation planning. Proceedings of the
508 National Academy of Sciences **114**:12755–12760. National Acad Sciences.
509 Hoban S. 2019. New guidance for ex situ gene conservation: Sampling realistic population
510 systems and accounting for collection attrition. Biological Conservation **235**:199–208.
511 Elsevier.
512 Hoban S, Schlarbaum S. 2014. Optimal sampling of seeds from plant populations for ex-situ
513 conservation of genetic biodiversity, considering realistic population structure. Biological
514 Conservation **177**:90–99.
515 Huang C-L, Chen J-H, Chang C-T, Chung J-D, Liao P-C, Wang J-C, Hwang S-Y. 2016.
516 Disentangling the effects of isolation-by-distance and isolation-by-environment on genetic
517 differentiation among Rhododendron lineages in the subgenus Tsutsusi. Tree genetics &
518 genomes **12**:53. Springer.
519 Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers.
520 Bioinformatics **24**:1403–1405. Oxford Univ Press.
521 Keller M, Kollmann J, Edwards PJ. 2000. Genetic introgression from distant provenances
522 reduces fitness in local weed populations. Journal of applied ecology **37**:647–659. Wiley
523 Online Library.
524 Lawrence MJ, Marshall DF, Davies P. 1995. Genetics of genetic conservation. I. Sample size
525 when collecting germplasm. Euphytica **84**:89–99. Available from

526 https://doi.org/10.1007/BF01677945.

527 Ledig FT. 2000. Founder effects and the genetic structure of Coulter pine. Journal of Heredity

528 **91**:307–315. Oxford University Press.

529 Li Q, Xu Z, He T. 2002. Ex situ genetic conservation of endangered Vatica guangxiensis

530 (Dipterocarpaceae) in China. Biological Conservation **106**:151–156. Elsevier.

531 Lockwood DR, Richards CM, Volk GM. 2007. Probabilistic models for collecting genetic

532 diversity: comparisons, caveats, and limitations. Crop Science **47**:861–866.

533 Loveless MD, Hamrick JL. 1984. Ecological determinants of genetic structure in plant

534 populations. Annual review of ecology and systematics **15**:65–95. Annual Reviews 4139 El

535 Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

536 Lowry DB, Rockwood RC, Willis JH. 2008. Ecological reproductive isolation of coast and

537 inland races of Mimulus guttatus. Evolution: International Journal of Organic Evolution

538 **62**:2196–2214. Wiley Online Library.

539 Maruyama T, Fuerst PA. 1985. Population bottlenecks and nonequilibrium models in population

540 genetics. II. Number of alleles in a small population that was formed by a recent bottleneck.

541 Genetics **111**:675–689. Genetics Soc America.

542 McBride CS, Singer MC. 2010. Field studies reveal strong postmating isolation between

543 ecologically divergent butterfly populations. PLoS biology **8**. Public Library of Science.

544 Nadeau S, Meirmans PG, Aitken SN, Ritland K, Isabel N. 2016. The challenge of separating

545 signatures of local adaptation from those of isolation by distance and colonization history:

546 The case of two white pines. Ecology and evolution **6**:8649–8664. Wiley Online Library.

547 Naredo MEB, Mercado SMQ, Banaticla-Hilario MCN, Berdos ML, Rodriguez MA, McNally

548 KL, Hamilton RS. 2017. Genetic diversity patterns in ex situ collections of Oryza officinalis

549 Wall. ex G. Watt revealed by morphological and microsatellite markers. Genetic Resources
550 and Crop Evolution **64**:733–744. Springer.
551 Nei M. 1987. Molecular evolutionary genetics. Columbia university press.
552 O'Connell LM, Mosseler A, Rajora OP. 2007. Extensive long-distance pollen dispersal in a
553 fragmented landscape maintains genetic diversity in white spruce. Journal of Heredity
554 **98**:640–645. Oxford University Press.
555 Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M. 2007. The
556 vegan package. Community ecology package **10**:631–637.
557 Sexton JP, Hangartner SB, Hoffmann AA. 2014. Genetic isolation by environment or distance:
558 which pattern of gene flow is most common? Evolution **68**:1–15. Wiley Online Library.
559 Sharrock S, Hoft R, Dias BF de S. 2018. An overview of recent progress in the implementation
560 of the Global Strategy for Plant Conservation-a global perspective. Rodriguésia **69**:1489–
561 1511. SciELO Brasil.
562 Slatkin M. 1993. Isolation by distance in equilibrium and non□equilibrium populations.
563 Evolution **47**:264–279. Wiley Online Library.
564 Stern DL, Orgogozo V. 2009. Is genetic evolution predictable? Science **323**:746–751. American
565 Association for the Advancement of Science.
566 Vinceti B, Loo J, Gaisberger H, van Zonneveld MJ, Schueler S, Konrad H, Kadu CAC, Geburek
567 T. 2013. Conservation priorities for Prunus africana defined with the aid of spatial analysis
568 of genetic data and climatic variables. PloS one **8**. Public Library of Science.
569 Wang IJ, Bradburd GS. 2014. Isolation by environment. Molecular ecology **23**:5649–5662.
570 Wiley Online Library.
571 Wang IJ, Summers K. 2010. Genetic structure is correlated with phenotypic divergence rather

than geographic isolation in the highly polymorphic strawberry poison dart frog. Molecular Ecology **19**:447–458. Wiley Online Library.

Wang T, Wang Z, Xia F, Su Y. 2016. Local adaptation to temperature and precipitation in naturally fragmented populations of Cephalotaxus oliveri, an endangered conifer endemic to China. Scientific reports **6**. Nature Publishing Group.

Westengen OT, Jeppson S, Guarino L. 2013. Global ex-situ crop diversity conservation and the Svalbard Global Seed Vault: Assessing the current status. PloS one **8**:e64146. Public Library of Science.

Whitlock R, Hipperson H, Thompson DBA, Butlin RK, Burke T. 2016. Consequences of in-situ strategies for the conservation of plant genetic diversity. Biological Conservation **203**:134–142. Elsevier.

Wright S. 1943. Isolation by distance. Genetics **28**:114. Genetics Society of America.

Xia H, Wang B, Zhao W, Pan J, Mao J, Wang X. 2018. Combining mitochondrial and nuclear genome analyses to dissect the effects of colonization, environment, and geography on population structure in Pinus tabuliformis. Evolutionary Applications **11**:1931–1945. Wiley Online Library.

589 **Tables**

590 **Table 1** Evolutionary processes [a] contributing to genetic structure across neutral and adaptive
591 genetic markers and their predicted weight [b] on expected patterns of among-population genetic
592 differentiation (Random, IBD and IBE).

| **Neutral genetic markers** | Random | IBD | IBE |
|---|---|---|---|
| Stochastic processes (e.g. genetic drift, inbreeding) | ++ | - | - |
| Demographic history (e.g. founder events) | ++ | + | - |
| Genetic drift combined with gene flow | - | +++ | - |
| Natural selection | - | - | + |
| **Adaptive genetic markers** | Random | IBD | IBE |
| Stochastic processes (e.g. genetic drift, inbreeding) | - (+) | - | - |
| Demographic history (e.g. founder events) | - | - | - |
| Genetic drift combined with gene flow | - | + | - |
| Natural selection | - | - | +++ |

593 [a] Here we distinguish between genetic drift alone as a stochastic evolutionary force and genetic
594 drift combined with gene flow as a process leading to a pattern of IBD.
595 [b] -: no, +: small, ++: intermediate and +++: important influence of the evolutionary forces on the
596 specified pattern.

**Table 2** Proportion of genetic differentiation explained by environmental and geographic variables [a], obtained using variation partitioning analyses, and correlation coefficients estimated between pairwise geographic and environmental Euclidean distances for all 19 genetic and genomic datasets downloaded from Dryad (see Appendix S1).

| Study system | | Data | | Results | | | | |
|---|---|---|---|---|---|---|---|---|
| Species | Distribution | Number of Populations | Genetic Marker [d] | IBD (Adj. $R^2$) | IBE (Adj. $R^2$) | IBD∩IBE (Adj. $R^2$) | IBD∪IBE (Adj. $R^2$) | Corr. (r) |
| *Betula maximowicziana* | Japan | 48 | EST-SSRs | 0.02 | 0.02 | 0.42 | 0.46 [e] | 0.48 [e] |
| *Centaurea solstitialis* [b] | Eurasia | 25 | SNPs | 0.14 [e] | 0.33 [e] | 0 | 0.47 [e] | -0.02 |
| *Helianthus annuus* | North America | 15 | SNPs | 0.1 [e] | 0.08 [f] | 0.02 | 0.2 [e] | 0.93 [e] |
| *Helianthus argophyllus* [b] | Texas | 51 | Gen-SNPs | 0.02 | 0.04 [e] | 0.32 | 0.38 [e] | 0.9 [e] |
| *Mimulus guttatus* [b] | United Kingdom | 14 | SNPs | 0.14 | 0.09 | 0 | 0.23 | 0.56 [e] |
| *Mimulus lacinatus* [b] | California | 23 | SSRs | 0.01 | 0.03 | 0.04 | 0.08 [f] | 0.35 [e] |
| *Narcissus papyraceus* [b] | Spain and Morocco | 26 | SSRs | 0.12 [f] | 0.03 | 0.02 | 0.17 [f] | 0.08 |
| *Nothofagus alpina* | Chile | 12 | SSRs | 0 | 0 | 0.18 | 0.18 | 0.49 [e] |
| *Nothofagus glauca* | Chile | 8 | SSRs | 0.75 [e] | 0.05 | 0.06 | 0.86 [e] | 0.2 |
| *Nothofagus obliqua* | Chile | 20 | SSRs | 0.17 [e] | 0.06 | 0.39 | 0.62 [e] | 0.31 [e] |
| *Picea sitchensis* [b] | North America | 10 | Gen-SNPs | 0.07 | 0 | 0.37 | 0.44 | 0.44 [e] |
| | | 10 | SEL-SNPs | 0.15 | 0 | 0.56 | 0.71 [f] | 0.44 [e] |
| *Populus balsamifera* [b] | North America | 31 | Gen-SNPs | 0.35 [e] | 0.01 | 0.3 | 0.66 [e] | 0.42 [e] |
| | | 31 | SEL-SNPs | 0.32 [e] | 0.01 | 0.42 | 0.75 [e] | 0.42 [e] |
| *Populus* | Sweden | 12 | Gen-SNPs | 0.02 | 0 | 0.02 | 0.04 | 0.71 [e] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *tremula* [c] | | | [control set] | | | | | |
| | | 12 | Gen-SNPs [defense set] | 0.15 | 0.05 | 0.33 | 0.53 [e] | 0.71 [e] |
| | | 12 | SEL-SNPs | 0.16 | 0.07 | 0.25 | 0.48 [e] | 0.71 [e] |
| *Rhododendron oldhamii* | Taiwan | 18 | EST-SSRs | 0.13 [e] | 0.05 | 0.24 | 0.42 [e] | 0.29 [e] |
| *Shorea leprosula* | South-East Asia | 24 | EST-SSRs | 0.24 [e] | 0.03 | 0.25 | 0.52 [e] | 0.27 [e] |

[a] Proportion of population genetic differentiation explained by pure geographic factors (IBD), pure environmental factors (IBE), the shared variation between environmental and geographic factors (IBD∩IBE), and both environmental and geographic factors combined (IBD∪IBE).

[b] Subsampled genetic or genomic datasets; [c] Adjusted geographical coordinates

[d] SSR (single-sequence repeat, neutral class), EST-SSR (expressed sequence tag single-sequence repeat, functional class), SNPs (genome-wide single-nucleotide polymorphism, neutral class), Gen-SNPs (genic single-nucleotide polymorphism, functional class) and SEL-SNPs (single-nucleotide polymorphism identified as potentially under selection, adaptive class).

[e, f] Fractions of variation explained and correlation coefficients are significant ($\alpha$=0.05 [f], $\alpha$=0.1 [g]).

608 **Table 3** Median proportion and 95% CI * of population genetic differences explained by IBD, IBE, IBD∩IBE, and IBD∪IBE given by

609 genetic marker classes.

| Genetic Marker Class | IBD<br>Median Adj. $R^2$ (95% CI) | IBE<br>Median Adj. $R^2$ (95% CI) | IBD∩IBE<br>Median Adj. $R^2$ (95% CI) | IBD∪IBE<br>Median Adj. $R^2$ (95% CI) |
|---|---|---|---|---|
| Neutral | 0.13 (0.09, 0.25) | 0.055 (0.02, 0.08) | 0.03 (-0.12, 0.06) | 0.215 (-0.19, 0.26) |
| Functional | 0.1 (-0.04, 0.18) | 0.025 (0, 0.045) | 0.31 (0.25, 0.38) | 0.45 (0.37, 0.52) |
| Adaptive | 0.16 (0, 0.17) | 0.01 (-0.05, 0.02) | 0.42 (0.28, 0.59) | 0.71 (0.67, 0.94) |

610 * 95% CI were obtained by bootstrapping. We considered two medians to be significantly different ($\alpha$=0.05) if their confidence
611 intervals did not overlap.

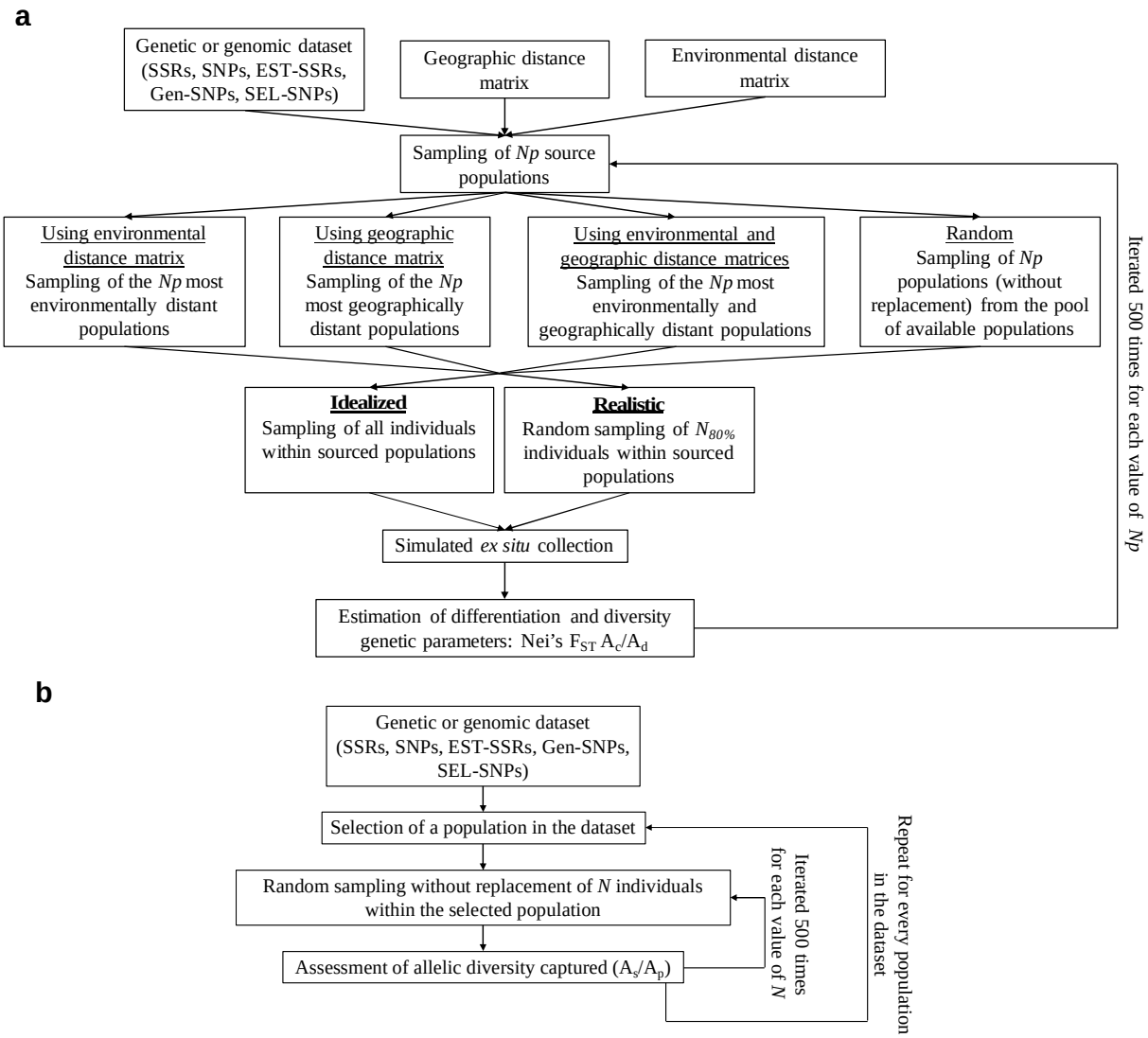


**Figure 1** (a) Simulation framework used to estimate genetic variation and differentiation parameters in *ex situ* collections simulated under two different within-population sampling scenarios (realistic and idealized) and four distinct population prioritization strategies (random, based on environmental distance, based on geographic distance, and based on both environmental and geographic distance combined). (b) Simulation framework used to estimate the number of individuals required to capture between 80-100% of allelic diversity in every population of a dataset ($N_{80\%}$, see Figure 1a). Simulations using both frameworks were conducted on each dataset independently. Computation proceeds from top to bottom.

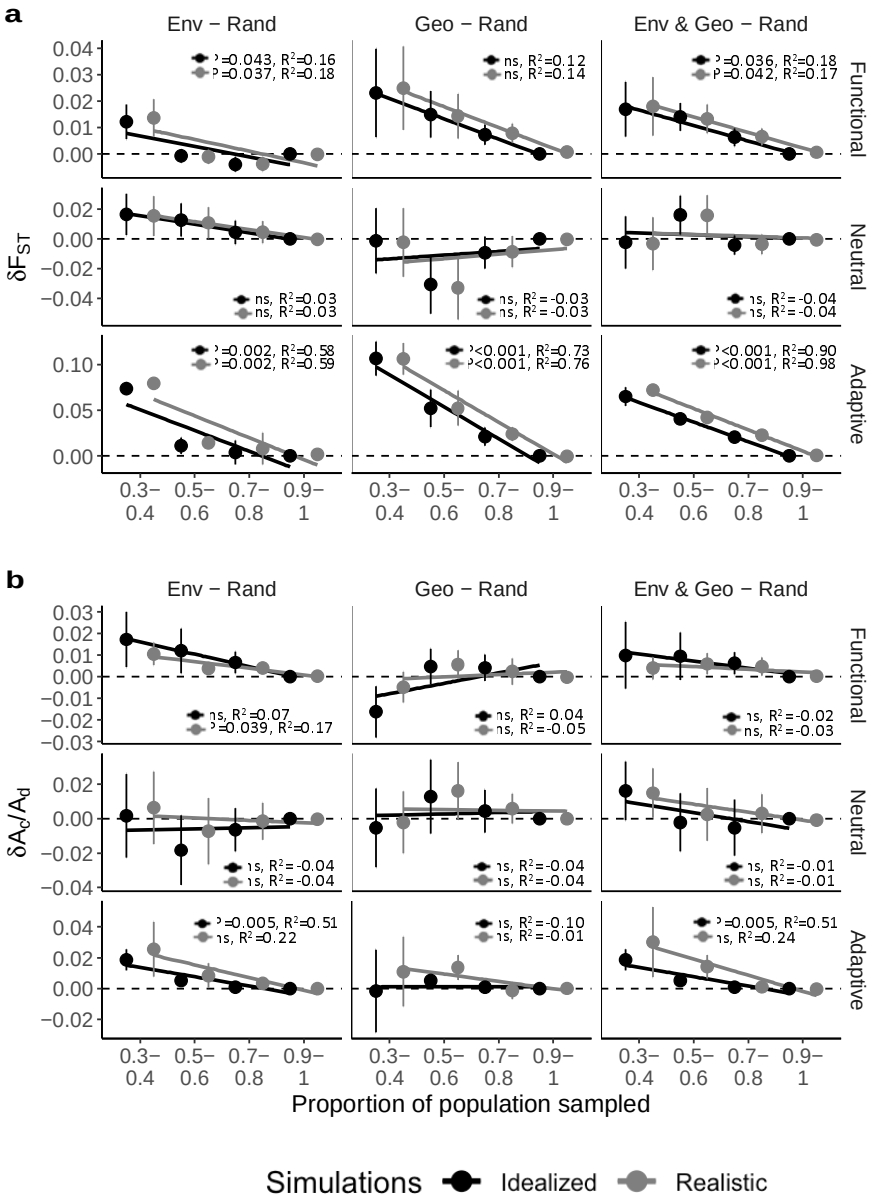


**Figure 2** Average differences and SE across datasets in genetic summary statistics (y-axis) estimated from *ex situ* collections simulated using distance-informed (environmental: Env, geographic: Geo, environmental and geographic: Env & Geo) and random (Rand) population sampling strategies (columns) separated by genetic marker classes (rows). Differences in genetic summary statistics were estimated for various proportions of populations sampled (x-axis). (a) Populations genetic differentiation (Nei's $F_{ST}$). (b) Allelic diversity captured in simulated *ex situ* collections ($A_c/A_d$). ns: non-significant.