frontiers in Genetics

ORIGINAL RESEARCH
published: 18 September 2020
doi: 10.3389/fgene.2020.00870

Evaluation of the Minimum Sampling Design for Population Genomic and Microsatellite Studies: An Analysis Based on Wild Maize

Jonás A. Aguirre-Liguori[1,2†], Javier A. Luna-Sánchez[1†], Jaime Gasca-Pineda[1] and Luis E. Eguiarte[1*]*

[1] Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2] Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA, United States

Massive parallel sequencing (MPS) is revolutionizing the field of molecular ecology by allowing us to understand better the evolutionary history of populations and species, and to detect genomic regions that could be under selection. However, the economic and computational resources needed generate a tradeoff between the amount of loci that can be obtained and the number of populations or individuals that can be sequenced. In this work, we analyzed and compared two simulated genomic datasets fitting a hierarchical structure, two extensive empirical genomic datasets, and a dataset comprising microsatellite information. For all datasets, we generated different subsampling designs by changing the number of loci, individuals, populations, and individuals per population to test for deviations in classic population genetics parameters (H_S, F_{IS}, F_{ST}). For the empirical datasets we also analyzed the effect of sampling design on landscape genetic tests (isolation by distance and environment, central abundance hypothesis). We also tested the effect of sampling a different number of populations in the detection of outlier SNPs. We found that the microsatellite dataset is very sensitive to the number of individuals sampled when obtaining summary statistics. F_{IS} was particularly sensitive to a low sampling of individuals in the simulated, genomic, and microsatellite datasets. For the empirical and simulated genomic datasets, we found that as long as many populations are sampled, few individuals and loci are needed. For the empirical datasets, we found that increasing the number of populations sampled was important in obtaining precise landscape genetic estimates. Finally, we corroborated that outlier tests are sensitive to the number of populations sampled. We conclude by proposing different sampling designs depending on the objectives.

Keywords: genomics of populations, landscape genomics, local adaptation, massive parallel sequencing, Mexican wild maize, sampling design

OPEN ACCESS

Edited by:
Genlou Sun,
Saint Mary's University, Canada

Reviewed by:
Zheng-Feng Wang,
South China Botanical Garden, Chinese Academy of Sciences, China
Pablo Orozco-terWengel,
Cardiff University, United Kingdom

****Correspondence:***
Jonás A. Aguirre-Liguori
jonas_aguirre@hotmail.com
Luis E. Eguiarte
fruns@unam.mx

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 May 2020*
Accepted: *16 July 2020*
Published: *18 September 2020*

Citation:
Aguirre-Liguori JA, Luna-Sánchez JA, Gasca-Pineda J and Eguiarte LE (2020) Evaluation of the Minimum Sampling Design for Population Genomic and Microsatellite Studies: An Analysis Based on Wild Maize. Front. Genet. 11:870. doi: 10.3389/fgene.2020.00870

INTRODUCTION

Massive parallel sequencing (MPS) has revolutionized the fields of molecular ecology, population genetics, and landscape genetics (Metzker, 2010; Stapley et al., 2010; Ekblom and Galindo, 2011). By increasing the number of polymorphic sites, it is now possible to estimate, with higher resolution, the genetic diversity, genetic structure, and demographic history of populations (Allendorf et al., 2010; Schoville et al., 2012; Excoffier et al., 2013; van Meier et al., 2017;

Aguirre-Liguori et al., 2019a), and the environmental and geographic mechanisms that determine the connectivity between populations (Bradburd et al., 2013). MPS also allows for identifying genomic regions that could be under selection (Foll and Gaggiotti, 2008; Coop et al., 2010; Stapley et al., 2010; De Villemereuil and Gaggiotti, 2015).

MPS is powerful in detecting patterns of local adaptation and understanding how the environment structures genetic diversity; nevertheless, its potential capacity depends on sampling a large geographic area, and encompassing an adequate environmental and genomic representation of the species (Schoville et al., 2012; De Mita et al., 2013; Tiffin and Ross-Ibarra, 2014). Unfortunately, for many research groups MPS is still expensive, or in some other cases, such as rare or endangered species, obtaining a large number of populations or individuals, and/or enough DNA for genomic studies can be challenging. In addition, the bio-informatic processing required for large samples can be limiting, making it difficult to obtain adequate genomic representation for enough individuals and populations. A solution has been to prioritize sequencing power to compensate for fewer individuals or populations (Schiffels and Wang, 2020). However, in the context of local adaptation, sampling populations in different parts of the distribution or different environments can affect the adequate estimation of genetic parameters (Meirmans, 2015). For instance limited sampling can make it difficult identifying center-edge patterns (Eckert et al., 2008), or the detection of outlier regions that might be under selection (De Mita et al., 2013). Thus, it is crucial to determine the potential biases associated with sampling (number of individuals, loci, and populations) and to define the tradeoff between the sampling effort and the number of polymorphic regions obtained with MPS that are needed to obtain robust estimates (Pruett and Winker, 2008; Willing et al., 2012; De Mita et al., 2013; Fumagalli, 2013).

So far, different studies have evaluated the errors and biases generated in estimates of genetic parameters when a different number of populations, the number of polymorphic sites, and the number of individuals are used (**Table 1** summarizes a list of studies that have evaluated sampling designs on population genetics studies). In summary, these studies have shown that parameters of mean genetic diversity (F_{ST}, F_{IS}, Hs) are not affected by sampling a different number of loci, the number of individuals or the number of populations; however, the variance decreases as the number of populations, individuals, and loci increases (see summaries and references in **Table 1**). In contrast, these studies have shown that patterns of isolation by distance and isolation by environment (IBE) across reduced areas are sensitive to the number of populations sampled and the sampling design (linear, aggregated, random sampling).

Sampling design has been studied widely. Nevertheless, the majority of the studies mentioned above (**Table 1**) were conducted mainly considering microsatellites, and thus focused on fewer loci and higher mutation rates than MPS data. In addition, studies centered on MPS markers were mostly based on bio-informatic simulations (**Table 1**). Among these, three studies have evaluated the effect of sampling design on estimates of genetic parameters using empirical and genomic data. Puckett and Eggert (2016) compared data for 15 microsatellite and 1,000 SNPs in *Ursus americanus* and found that the SNP dataset was more precise than the microsatellites in assigning the provenance of 96 individuals sampled across 34 populations. Nazareno et al. (2017) analyzed different sampling tests of *Amphirrhox longifolia* (Violaceae), ∼4,000 SNPs and 70 individuals distributed in two populations. They found that sampling over eight individuals per population and 1,000 SNPs did not increase the accuracy in the estimation of summary statistics. Flesch et al. (2018) analyzed four populations of rocky mountain bighorn sheep, 14,000 SNPs, and 120 individuals in total, finding that an accurate estimation of genetic parameters was achieved after sampling 25 individuals per population.

While the studies of Puckett and Eggert (2016), Nazareno et al. (2017), and Flesch et al. (2018) are without doubt informative and relevant, they were performed in most cases in relatively few populations (34, 2, and 4 populations, respectively) and were based in a relatively small number of individuals (96, 70, 120, respectively) or SNPs (1,000, ∼4,000, and ∼14,000, respectively). More importantly, these studies did not test the effect of sampling design on the detection of outlier SNPs using empirical data.

In this study we aimed at testing the effect of sampling design to assess the potential biases and errors in estimates of population genomics parameters, in patterns of isolation, and in the detection of outlier SNPs while using empirical datasets. For this, we compared two simulated data sets, two large genomic datasets (33,454 SNPs, 646 individuals, and 49 populations obtained with the MaizeSNP50 Genotyping BeadChip; and 9,735 SNPs, and individuals pooled from 47 populations obtained with the DArTseqTM data), and one microsatellite dataset (22 microsatellite loci, 527 individuals, and 29 populations) of Mexican wild maize populations (*Zea mays* ssp. *mexicana* and *Zea mays* ssp. *parviglumis*) to explore the effects of sampling design in the estimation of population genomics parameters (Hs, F_{IS}, F_{ST}), landscape genetics (tests of isolation by distance and environment), test of centrality (association between genetic diversity and the distance from the center of the geographic or niche distribution; Eckert et al., 2008; Lira-Noriega and Manthey, 2014; Aguirre-Liguori et al., 2017), and estimation of candidate SNPs (outlier SNP detection tests).

In particular, we compared the effect of (1) using MPS vs. microsatellites markers; (2) using individual data with known ascertainment bias (MaizeSNP50 Genotyping BeadChip) vs. pooled non-ascertained biased data (DArTseqTM data); (3) varying the number of sampled loci (genomic datasets: 100, 1,000, 5,000, 15,000; microsatellite datasets: 5, 10, 15); (4) varying the number of sampled individuals per population (3, 6, and 9 individuals); (5) changing the number of sampled populations (5, 10, 20, 30, 40 populations); and (6) testing the effect of the number of sampled populations in the detection of outlier SNPs.

MATERIALS AND METHODS

Studied Taxon

Mexican wild maize, or teosintes, are divided into two main subspecies, the lowland subspecies *Zea mays* ssp. *parviglumis* (hereafter *parviglumis*) and the highland subspecies *Zea mays*

TABLE 1 | Summary of 19 studies that have evaluated sampling designs using different markers (Microsatellites, AFLPs, SNPs); empirical vs. simulated data; and varying the number of loci, individuals, and populations.

References	Dataset	Type of sampling	No. of populations	No. of individuals	No. of Loci	Principal conclusions
Miyamoto et al. (2008)	Microsatellite	Empirical	1	480	4	> 30 individuals increases the precision in *Hs* Between 200 and 300 individuals increase the precision of allelic richness estimates.
Pruett and Winker (2008)	Microsatellite	Empirical	1	200	8	Precision in summary statistics is increased when > 20 individuals are genotyped.
González-Ramos et al. (2015)	Microsatellite	Empirical	2	64	15	Above 6 polymorphic markers are enough to adequately define the genetic structure between populations.
Peterman et al. (2016)	Microsatellite	Empirical	5	80	15	Increasing the number of loci does not change the mean summary statistics, but increases the precision across replicates. IBD patterns are sensitive to fewer loci genotyped.
Sánchez-Montes et al. (2017)	Microsatellite	Empirical	17–21 (different species)	547, 652, and 516	18, 16, and 15	> 20 individuals and between 50 and 80 individuals per population are needed to estimate H_S with precision, and allelic richness, respectively.
Rico (2017)	Microsatellite	Simulation	17 and 34 (different species)	5,000 and 3,000	20	Spatial sampling design (random, systemic, cluster) affect IBD patterns. Increasing loci, over individuals, increases the accuracy of IBD estimates.
Schwartz and McKelvey (2009)	Microsatellite	Simulation	1	10,000	15	Different sampling designs generate different F_{ST} estimates, and different *Structure* outputs.
Landguth et al. (2012)	Microsatellite	Simulation	1	1,000	25	Increasing the number of polymorphic loci increases the precision of patterns of isolation by resistance (IBR).
Oyler-McCance et al. (2013)	Microsatellite	Simulation	1	1,000	25	Increasing the number of polymorphic loci, individuals, and number of alleles increases the precision and the accurate estimation of patterns of (IBR).
Landguth and Schwartz (2014)	Microsatellite	Simulation	64	64	20	Increasing the number of populations (even if fewer individuals are sampled) increases the possibility of finding correct patterns of IBD.
Smith and Wang (2014)	Microsatellite	Simulation	3	100	100	Reducing the number of samples do not affect *Hs*, F_{ST} estimates, but reduces the power to detect accurate allelic richness.
Hale et al. (2012)	Microsatellite	Mixed (Simulation and empirical)	4	100	9, 5, 7, and 8	For four different species, sampling between 25 and 30 individuals are enough to estimate accurately H_S and F_{ST}.
Dubois et al. (2017)	Microsatellite	Mixed (Simulation and empirical)	4	4 different taxa: 726, 408, 372, 384	16	Sex proportions do not affect summary statistics estimates. >20 individuals increase the precision of summary statistics. Empirical and simulated data show different patterns of deviation.
Sinclair and Hobbs (2009)	AFLPs	Empirical	6	159	59 and 117	>30 individuals per population needed to estimate accurately F_{ST}.
Willing et al. (2012)	SNPs	Simulation	2	1,000	21,000	Fewer individuals are needed to accurately estimate F_{ST} for MPS datasets.

(Continued)

TABLE 1 | Continued

References	Dataset	Type of sampling	No. of populations	No. of individuals	No. of Loci	Principal conclusions
Fumagalli (2013)	SNPs	Simulation	1	1,000	20,000	Low individual sampling, with a high genome coverage underestimates the number of segregating sites, H_S estimates and genetic structure.
Nazareno et al. (2017)	SNPs	Empirical	2	70	3,500	Fewer individuals (8) but with a large number of SNPs (>1,000) increase the precision of H_S and F_{ST}.
Flesch et al. (2018)	SNPs	Empirical	4	120	14,000	>25 individuals (with 10,000 SNPs) are needed to estimate accurate kinship indexes (10,000 SNPs), identifying as identical by descent alleles and F_{ST} values.
Puckett and Eggert (2016)	Mixed (SNPs and Microsatellite)	Empirical	34	Microsatellites dataset: 506 SNP dataset: 96	Microsatellite dataset: 15 SNP dataset: 1,000	1,000 SNPs are more precise than microsatellites for assigning birth areas, even if fewer individuals are sampled.

ssp. *mexicana* (hereafter *mexicana*) (Aguirre-Liguori et al., 2016). Demographic studies suggest that *mexicana* was originated from *parviglumis* between 20,000 and 60,000 years ago and that divergence occurred in the presence of gene flow (Aguirre-Liguori et al., 2019a). Consequently, the Mexican wild teosintes fit a model of hierarchical gene flow, with higher gene flow occurring within subspecies. Given the close relatedness of teosintes to maize, different genomic resources are available (Hufford et al., 2012; Aguirre-Liguori et al., 2016) and several studies have analyzed their population genomics (van Heerwaarden et al., 2011; Hufford et al., 2013; Pyhäjärvi et al., 2013; Aguirre-Liguori et al., 2017, 2019a,b; Fustier et al., 2017, 2019; Moreno-Letelier et al., 2020). Briefly, genomic studies suggest that teosintes have high genetic diversity, show patterns of isolation by distance and environment, and show strong patterns of local adaptation (Pyhäjärvi et al., 2013; Aguirre-Liguori et al., 2017, 2019a,b; Fustier et al., 2017, 2019). The vast genomic resources and biological knowledge makes teosintes an ideal system to study the importance of sampling design in analyses of genetic diversity, isolation patterns, and identification of candidate SNPs.

Datasets

Simulated Datasets

The majority of tests that have analyzed the effect of sampling design on the estimations of summary statistics using genomic information have been performed with simulated data (**Table 1**). The advantage of using simulated data is that it allows for modeling an evolutionary process based on known demographic parameters. Here we simulated two large genomic datasets to analyze the effect of sampling design on the estimations of summary statistics and then compared the results with two empirical genomic datasets and one microsatellite dataset.

We used Fastsimcoal 2 (Excoffier and Foll, 2011; Excoffier et al., 2013) to simulate two demographic models consisting of 50 populations divided into two genetic clusters fitting a model of hierarchical structure (i.e., two subspecies of teosintes). Populations belonging to the old genetic cluster (i.e., *parviglumis*) coalesced with their common ancestor approximately 140,000 generations ago (Ross-Ibarra et al., 2009). Populations belonging to the young genetic cluster (i.e., *mexicana*) coalesced with their common ancestor approximately 20,000 generations ago (Aguirre-Liguori et al., 2019a). We set the time of divergence between the two genetic clusters at ∼20,000 generations ago (Aguirre-Liguori et al., 2017).

The effective population size of the old genetic cluster was set to ∼5,000 individuals and was 1.5 times higher than the young genetic cluster. For the first model (the hierarchical model with high gene flow), migration between populations belonging to the same genetic clusters were set at a 0.001 probability of a gene moving from one population to the other back in time. The migration between populations belonging to different genetic clusters were 10 times smaller (0.0001). For the second model (the hierarchical model with low gene flow), gene flow did not occur between populations belonging to different genetic clusters.

To incorporate variation in the demographic parameters across populations, we used the *norm* function in R to create a normal distribution for each demographic parameter (N_e, *m, T,* and inbreeding index) with the mean values detailed above. Next, for each population we sampled a random value for each parameter.

We created the fastsimcoal inputs using the *fscWrite* function of the *strataG* package of R. For each model we used the command line *fsc26 –i input –n 1 –g –I* to simulate 30,000 SNPs (with an infinite site model) and 15 diploid individuals per population. Finally, we used *strataG* (Archer et al., 2016) and the *adegenet* (Jombart, 2008) pacakage of R 3.6.1 (R Core Team, 2019) to create for each simulated dataset a *genind* and a *hierfstat* input object for further analyses.

Empirical Datasets

For the empirical datasets, we combined the MaizeSNP50 Genotyping BeadChip data published by Pyhäjärvi et al. (2013) and Aguirre-Liguori et al. (2017) to obtain a total dataset

consisting of 49 populations, 24 belonging to *mexicana* and 25 to *parviglumis* (**Supplementary Figure S1**), including between 12 and 15 individuals per population, and 33,454 SNPs. Since the MaizeSNP50 Genotyping BeadChip was designed to maximize variation in maize, it has ascertainment bias (Albrechtsen et al., 2010). Therefore, this dataset is expected to include SNPs that are in high frequency across distant teosintes populations and thus might overestimate genetic diversity and underestimate genetic differentiation.

We also downloaded the DArTseq data from Aguirre-Liguori et al. (2019a), which are composed of pooled DNA of 47 populations (∼12 individuals per population), 21 belonging to *parviglumis* and 26 to *mexicana* (**Supplementary Figure S1**), and 9,735 SNPs. The DArTseq dataset was obtained by initially cutting the DNA using restriction enzymes (Sansaloni et al., 2011; Ren et al., 2015) and has lower ascertainment bias (see Aguirre-Liguori et al., 2019a). This dataset has a frequency spectrum with lower bias than the 50K dataset and is expected to generate more robust demographic inferences (Albrechtsen et al., 2010). We called the BeadChip and the DArTseq datasets the 50K and DTS datasets, respectively.

To be able to compare deviations obtained from MPS and microsatellite markers (**Table 1**), we also used the microsatellite dataset from Gasca-Pineda et al. (2020), which includes 527 individuals distributed across 29 populations, 14 belonging to *parviglumis* and 15 to *mexicana* (**Supplementary Figure S1**). This microsatellite dataset consists of 22 loci and 355 alleles.

For each population and dataset, we downloaded the longitude and latitude at which they grow (Supporting Information in Aguirre-Liguori et al., 2017, 2019a; Gasca-Pineda et al., 2020). We also obtained the score of the first principal component (PC1) describing temperature for each population. The environmental data were obtained from 19 bioclimatic variables downloaded from WorldClim at a 30°arc Resolution, and the PCA was performed using the prcomp function in R across all variables and populations.

These three datasets share many populations (**Supplementary Figure S1**). The microsatellite dataset is a subsample of the 50K dataset and therefore shares all populations with the 50K dataset. The 50K and DTS datasets shared 29 populations. Also, the three datasets are distributed along the entire geographic and environmental distribution of teosintes (Hufford et al., 2012; Pyhäjärvi et al., 2013; Aguirre-Liguori et al., 2017, 2019a). They are composed of many individuals per population (between 9 and 26 individuals per population) and contain a large number of SNPs or microsatellite markers, distributed along the 10 chromosomes of teosinte. Importantly, the 50K and DTS datasets are the largest genomic datasets based on population sampling (not accessions) that have been developed so far in teosintes. Therefore, we considered these datasets (and the microsatellite dataset) as the samples representing the most accurate data (i.e., the "real" data for the purpose of this work) and estimated the deviations in the estimations of summary statistics, landscape genetics, and tests for local adaptation, generated by sampling a different number of loci, the number of individuals, and the number of populations.

We used *adegenet* and *hierfstat* (Goudet, 2005) packages of R to generate *genind*, *genpop*, and *hierfstat* objects to manipulate the data. All these objects are indexable, and therefore allow subsampling random individuals, SNPs, microsatellite markers, subspecies, and/or populations.

For all subsamplings we combined the *mexicana* and *parviglumis* populations. However, complex demographic scenarios can bias the estimations of divergence between populations when only few populations are sampled (Chikhi et al., 2010; Heller et al., 2013; Robinson et al., 2014). For instance, hierarchical structure increases the F_{ST} between populations belonging to different genetic groups (Slatkin and Voelm, 1991) and reduced sampling can bias estimations of divergence if more populations are sampled within one genetic cluster than between genetic clusters. Alternatively, incomplete lineage sorting can underestimate the amount of divergence between populations belonging to different genetic clusters (Lack et al., 2010; Orozco-Terwengel et al., 2011; Jones, 2019). To test the effect of sampling bias associated with complex demographic structures, we also tested the effect of sampling design by analyzing each subspecies separately. Since the patterns were similar between the entire datasets and the subspecies datasets, for simplicity we present the results of the combined datasets and show results of each subspecies as **Supplementary Information**.

Estimation of Population Genetics Parameters

For each simulated dataset, the entire genomic datasets, and each subsampling within dataset (see below for descriptions of the subsamplings), we used the *basic.stats* function of the *hierfstat* package in R to calculate the sample H_S and F_{IS} and F_{ST}. For each summary statistic, we obtained the mean value across loci for each population.

For the empirical datasets, we also used environmental data to analyze landscape genetic associations. For the genomic datasets, we used the *BEDASSLE* package (Bradburd et al., 2013) in R to calculate the pairwise F_{ST} between populations (Weir and Hill, 2002). For the microsatellite datasets, we used the *pairwise.fst* function of the *hierfstat* package in R to calculate Nei's pairwise F_{ST} between populations.

We tested patterns of isolation by distance (IBD) and IBE using multiple regressions of distance matrices (*MRM* function from the *ecodist* package; Goslee and Urban, 2007) to test the association between pairwise genetic distance (F_{ST}) as a response variable and the environmental and geographic distances as predictive variables. We performed 1,000 permutations in each test. The advantage of MRM tests is that they allow for simultaneous testing in both the environmental and geographic distances, and determine the relative contribution of each variable (Lichstein, 2007).

Finally, we tested the central abundance hypothesis (CAH), which suggests that genetic diversity should reduce as a function of the distance from the geographic or niche centroid (Eckert et al., 2008; Martínez-Meyer et al., 2012; Lira-Noriega and Manthey, 2014). For the CAH tests, we used simple linear regressions (*lm* function in R) to test the association between

*H*s as the response variable and the distance to the niche and geographic centroids as independent variables (which were estimated as the Euclidian distances from the geographic and niche centroids; for more details of the methods see Aguirre-Liguori et al., 2017).

Sampling Designs

First we analyzed the effect of sampling design with the estimation of *H*s, F_{IS}, and F_{ST} using the simulated dataset. Since we controlled the demographic parameters of the simulations, we were able to generate an expectation of how sampling design would affect the estimation of summary statistics. Next, we used the empirical dataset to validate the simulated results.

Subsampling of the Number of Loci and the Number of Individuals per Population

We tested the effect of sampling a different number of SNPs or microsatellite markers per population. We used R custom scripts (available as Supporting Information- function num_locs) to extract data from the entire empirical and genomic datasets: for the DTS dataset 100 (∼1%), 1,000 (∼10%), and 5,000 (∼51%) random SNPs; for the 50K dataset and the simulated datasets100 (∼0.3%), 1,000 (∼3%), or 15,000 (∼45%) random SNPs; and for the microsatellite dataset 5 (∼22%), 10 (∼45%), and 15 (∼68%) random markers. For the simulated, the 50K, and the microsatellite datasets, we also tested the effect of sampling different estimates of individuals per populations. We extracted randomly for each population 3, 6, and 9 individuals [available as Supporting Information- function num_inds()]. This was not performed on the DTS dataset, because it was based on pooled DNA.

For the number of SNPs, the number of microsatellite markers, and the number of individuals per population, we re-sampled randomly and without replacement each set 1,000 times, we generated *genind*/*hierfstat*/*BEDASSLE* input objects and estimated the summary statistics described above (*H*s, F_{IS}, F_{ST}, IBE, IBD, CAH associations). For each parameter and each replicate we obtained the mean summary statistic across populations and generated a distribution based on 1,000 summaries corresponding to each subsampling.

Subsampling the Number of Populations

To test the effect of the number of populations in the estimation of the parameters described above (*H*s, F_{IS}, F_{ST}, IBE, IBD, CAH associations), we performed random sampling designs. For the simulated and the genomic datasets, we sampled 5 (∼10%), 10 (∼20%), 20 (∼40%), 30 (∼61%), and 40 (∼81%) random populations from the 49 (50K) and 47 (DTS) populations described above (**Supplementary Figure S1**). For the microsatellite dataset, we sampled 5 (∼17%), 10 (∼34%), and 20 (∼69%) random populations from the 29 populations described above (**Supplementary Figure S1**). Again, we generated 1,000 subsamples without replacement [available as Supporting Information- function num_pops ()], and for each replicate we generated *genind*/*genpop*/*hierfstat*/*BEDASSLE* inputs and in each case, we calculated the estimates described above, and for summary statistics we estimated the mean across populations.

Tradeoff Between Number of Individuals and Populations

To test the tradeoff between the number of individuals and number of populations, for the 50K dataset we also tested three sets of sampling designs changing the number of individuals sampled per population, going from fewer individuals in many populations to many individuals sampled in fewer populations. For three scenarios (3 individuals and 49 populations; 6 individuals and 24 populations; 9 individuals and 10 populations) we generated 1,000 subsamples and estimated the parameters described above.

Comparison Between Samplings and Between Datasets

For each subsampling, we compared qualitatively the simulations to the "real" dataset, to determine the deviations generated by different sampling designs. To be able to compare between different datasets (including the simulated and empirical datasets), we also compared the magnitude of the deviation between sampling designs and between datasets using the relative error between each subsampling and the estimated "real" summary statistics described above. The relative error was calculated as $(X_{est} - X_{sim})/X_{est}$, where X_{est} is the summary statistic estimated for the "real data" set and X_{sim} is the summary statistic estimated for a given subsampling.

Test for Local Adaptation

Detecting outlier SNPs is challenging, since high genetic structure can inflate false positives (Schoville et al., 2012; De Mita et al., 2013; Tiffin and Ross-Ibarra, 2014). We tested the effect of varying the number of populations in detecting outlier loci. For this, we subsampled without replacement 5, 10, 20, and 30 random populations from the entire 50K dataset (49 populations and between 12 and 15 individuals per population, see Aguirre-Liguori et al., 2017 for more details). Since outlier analyses are time-consuming, we only generated 10 replicates of each sampling design, and we subsampled 10,000 SNPs from the 50K dataset. We also ran the analysis 10 times with the 49 populations to have a comparable number of replicates. We chose 10,000 SNPs to reduce computing time and because our results (see below) show that over 1,000 SNPs are enough to identify adequately the genetic structure between populations, and therefore reduce false positives.

For each sample, we used *Bayescenv* (De Villemereuil and Gaggiotti, 2015) to identify outlier SNPs associated to PC1 (as in Aguirre-Liguori et al., 2017, 2019a). *Bayescenv* decomposes F_{ST} based on a signal shared between all loci (β), a signal specific to each locus (α), and the association of the SNP with the environmental variable tested (γ). We used default parameters to run the analyses and we defined outlier SNPs as those that had $q\text{-}val < 0.05$, which is a conservative approximation to detect outlier loci (De Villemereuil and Gaggiotti, 2015). For each replicate of each sampling design, we recorded the highest F_{ST} value for a SNP and the number of SNPs that had $q\text{-}val < 0.05$.

We used the entire dataset to identify outlier SNPs. We considered this dataset as presenting the "real outlier SNPs" representing the local adaptation to all environmental conditions

in which teosintes grow. We tested whether different sampling designs based on a different number of populations sampled would identify a subset of the outlier SNPs detected for the entire dataset. We used the *intersect* function in R to detect the SNPs that were considered as "outlier" for all replicates in each sampling design. We also used the *venn* function of the *gplots* package in R to identify SNPs that were the candidate (q-val) for all sampling designs (5, 10, 20, 30, and 49 populations) and the 10 replicates.

RESULTS

Summary Statistics for the Entire Datasets

Simulated Datasets

We generated two simulated datasets with hierarchical structures, but with different levels of gene flow between populations belonging to different genetic clusters.

For the two models, we found that estimated genetic diversity was high and the fixation index low (**Table 2** and **Supplementary Figure S2**), as it has been found in teosintes. As expected, we found that the hierarchical model with high gene flow had a lower mean F_{ST} than the hierarchical model with low gene flow (**Table 2** and **Supplementary Figure S2**). Importantly, we found large variance between populations for *H*s and F_{IS}, which is similar to what has been observed in teosintes (Aguirre-Liguori et al., 2017). The values of F_{ST} were close to 0, and for many replicates we found negative values approximate to 0 (see **Supplementary Figure S2**). Negative F_{ST} values occur when sample size corrections are used, and are usually considered to be 0. However, to be able to compare the relative error associated to sampling design, we recorded the F_{ST} estimated from *hierfstat*.

In brief, we consider that the simulated datasets were adequate datasets to generate expectations of how sampling designs would affect the estimation of summary statistics.

Empirical Datasets

We considered the entire datasets (50K, DTS, and microsatellite) as those revealing the "real" or most accurate patterns of genetic diversity across teosinte populations. **Table 2** shows the mean H_S, F_{IS}, and F_{ST} across populations, patterns of isolation by distance and environment, and the test of central abundance, estimated for the DTS, 50K, and microsatellite datasets (the distribution across different sampling designs are found in **Supplementary Table S1**).

We found striking differences among the datasets for the estimated mean across populations of *H*s, F_{ST}, and F_{IS} (**Figure 1** and **Table 2**). We detected that the DTS dataset presents low mean genetic diversity across populations (*H*s = 0.13), the 50K intermediate values (*H*s = 0.225), and the microsatellite data high values (*H*s = 0.69). In parallel fashion, we found that DTS shows the highest mean genetic structure across populations (F_{ST} = 0.393), followed by the 50K dataset (F_{ST} = 0.246), and finally the microsatellite dataset (F_{ST} = 0.11). We were not able to calculate F_{IS} for the DTS dataset (as they were derived from pooled DNA), but we also found differences between the estimated mean using the 50K (F_{IS} = 0.065) and microsatellite datasets (F_{IS} = 0.19).

In contrast to the summary statistics, for the three datasets we found similar patterns of IBD and IBE (**Figure 2**), based on the MRM tests (**Figure 2** and **Table 2**). For the three datasets, we observed that patterns of IBD and IBE were positive, indicating that there is isolation by distance or by environment. Finally, for the three experimental datasets we observed negative associations between genetic diversity and the distance to the geographic and niche centroids (**Figure 3** and **Table 2**), indicating that as populations grow further away from the center of their geographic distribution or the optimum ecological conditions, their genetic diversity is lower.

Varying the Number of Sampled Individuals

As mentioned above, this test was performed with the simulated datasets, the 50K, and the microsatellite datasets, since they were based on individual samples. The DTS dataset was generated from pooled DNA and therefore individual genotypes were not known.

For the two simulated datasets, we found that subsampling fewer individuals increased the variance and relative error in the estimation of *H*s, F_{IS}, and F_{ST} across 1,000 replicates (**Supplementary Figure S2**). Importantly, for F_{IS} estimations we found that when fewer individuals were sampled, the mean value across the 1,000 replicates was lower than the complete dataset,

TABLE 2 | Summary statistics estimated for the DTS, 50K, and microsatellite datasets of Mexican wild maize.

Mean estimate	Hierarchical high flow	Hierarchical low flow	DTS	50K	Microsatellite
H_S	0.26 (0.05)	0.32 (0.03)	0.130 (0.05)	0.225 (0.04)	0.691
F_{IS}	0.02 (0.18)	0.01 (0.18)		0.069 (0.04)	0.182
F_{ST}			0.393	0.246	0.106
MRM: geographic (β)			0.027	0.025	0.013
MRM: environmental (β)			0.011	0.011	0.004
CAH: geographic (β)			−0.014	−0.014	−0.041
CAH: environmental (β)			−0.006	−0.008	−0.031

The numbers in parenthesis correspond to standard deviation of the mean values. For the mean and maximum and minimum values across 1,000 replicates of each sampling designs (see ***Supplementary Table S1****).*

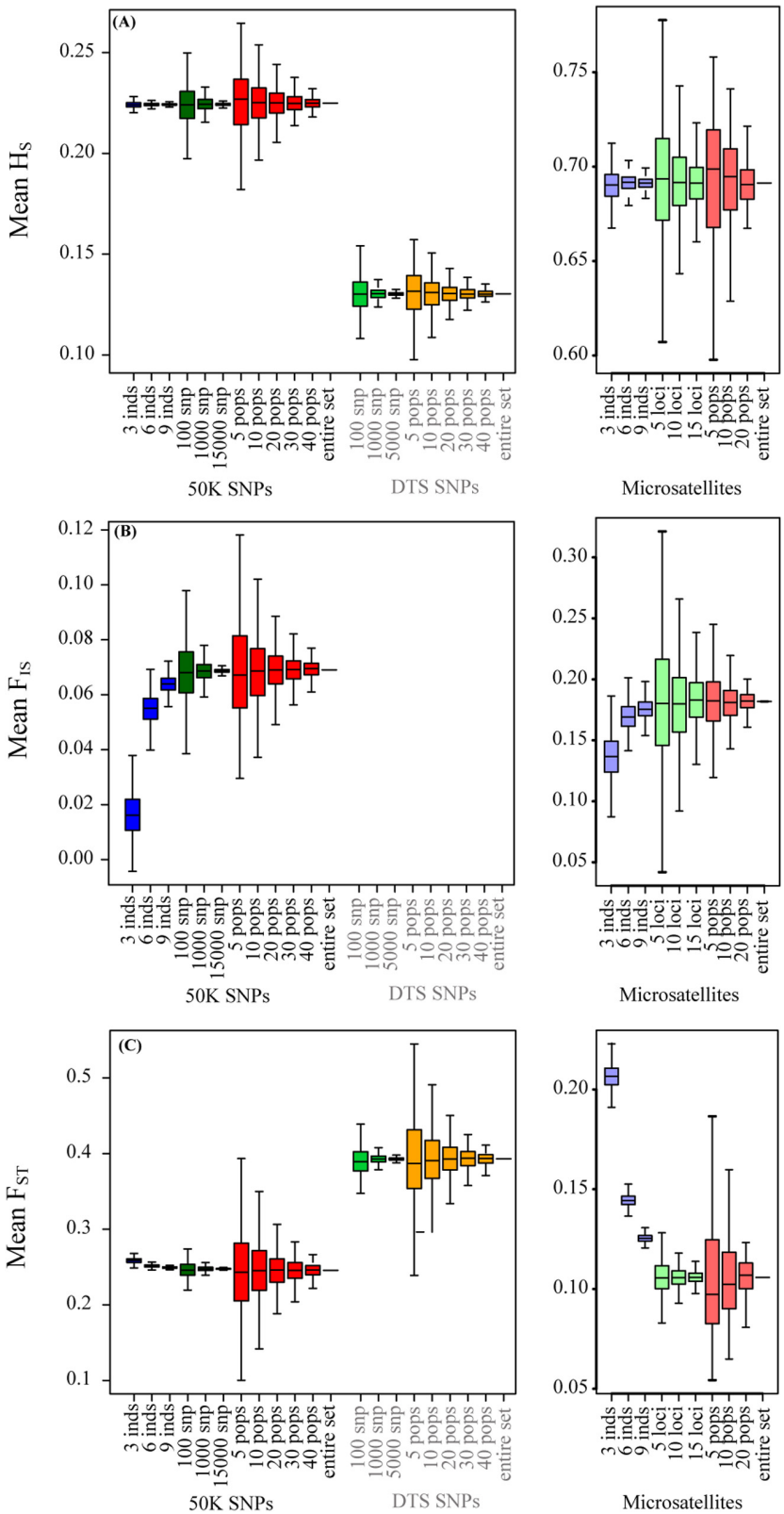

FIGURE 1 | The effect of sampling designs on the estimation of summary statistics for genomic (left panels) and microsatellite (right panels) datasets: **(A)** H_S; **(B)** F_{IS}; **(C)** F_{ST}. Boxplots show the distribution of mean summaries estimated for 1,000 replicate simulations varying the number of individuals, number of SNPs, and number of populations sampled. F_{IS} was not possible to obtain for the DTS dataset because it is based on pooled data.

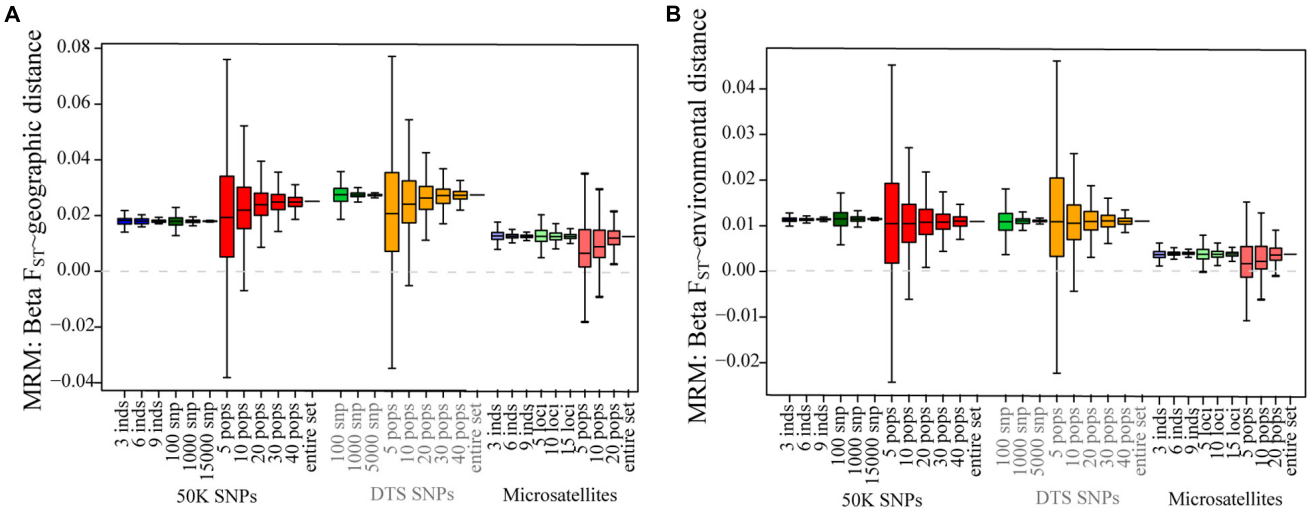

FIGURE 2 | The effect of sampling designs on the analysis of patterns of isolation for genomic and microsatellite datasets: **(A)** IBD-MRM test; **(B)** IBE-MRM test. Boxplots show the distribution of associations estimated for 1,000 simulations varying the number of individuals, number of SNPs, and number of sampled populations. The dotted gray line shows the 0 value.

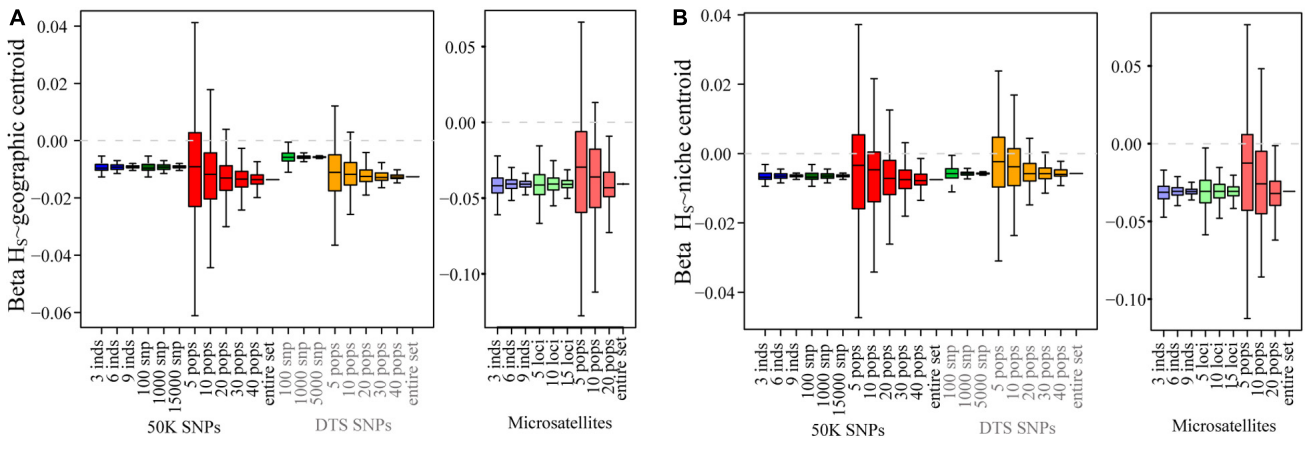

FIGURE 3 | The effect of sampling designs on the estimation of the central abundance hypothesis for genomic and microsatellite datasets: **(A)** the association between distance to the geographic centroid and *Hs*; **(B)** the association between distance to the niche centroid and *Hs*. Boxplots show the distribution of associations estimated for 1,000 simulations varying the number of individuals, number of SNPs, and number of sampled populations. The dotted gray line shows the 0 value.

indicating that this summary statistic was the most sensitive to the number of individuals sampled.

These patterns were similar for the empirical 50K and microsatellite datasets. Briefly, we found that sampling fewer individuals did not strongly affect the estimates of *Hs* (**Figure 1A** and **Supplementary Figure S3**). Also, in accordance to the simulated data, we found that changing the number of sampled individuals generated strong deviations and a higher relative error for the estimation of F_{IS} for the 50K and microsatellite datasets (**Figure 1B** and **Supplementary Figures S3, S6**).

In contrast to the simulated datasets, we found that sampling a different number of individuals generated moderate deviations for estimates of F_{ST} for the 50K dataset (**Figure 1C** and **Supplementary Figures S3, S6**) and large deviations for the estimates of F_{ST} for the microsatellite dataset (**Figure 1C** and **Supplementary Figures S3, S6**). We found that the relative error was higher for F_{IS} estimations when using the 50K dataset, but higher for microsatellites when estimating F_{ST} (**Supplementary Figure S6**). Importantly, in both datasets we found that sampling fewer individuals reduced the F_{IS} estimates and increased the F_{ST} estimates (**Figures 1B,C**).

For the empirical datasets, we also estimated the effect of sampling a different number of individuals in different landscape genetic tests. Even if we found increased variance in the estimation of F_{ST}, it was interesting to note that we did not find strong deviations for summary statistics describing patterns of isolation by distance and environment (**Figure 3** and **Supplementary Figure S4**). Finally, in accordance to low

variance in the estimation of *Hs*, we found low biases in the associations between genetic diversity and ecological variables (**Figure 3** and **Supplementary Figure S5**). The ranges of the maximum and minimum values were close to the "real" estimates for the 50K dataset.

It is interesting to note that F_{IS} was the only summary statistic that was very sensitive to the number of individuals sampled for the simulated datasets, and the empirical 50K and microsatellite datasets. Since this pattern is shared between the simulated and the empirical datasets, the differences are not associated with increased missing data or null alleles. To identify what could be generating these differences, we analyzed the F_{IS} values for all loci across populations. We found that fewer individuals sampled generated larger "NaN (Not a Number)" values across loci. This is the result of loci appearing as fixed since a larger number of individuals are need to sample low frequency alleles. We correlated the mean F_{IS} across populations and the number of "NaN" values estimated per loci per population and found that the reduction in F_{IS} correlated with an increased number "NaN" obtained when fewer individuals were sampled (**Supplementary Figure S7**).

Varying the Number of Sampled Loci

Next, we evaluated the effect of sampling a different number of loci in the estimation of summary statistics. For the two simulated datasets, we found that sampling fewer loci slightly increased the relative error and the variance in the estimation of the summary statistics across replicates (**Supplementary Figure S2**). Interestingly, the variance and relative error was stronger for *Hs* estimations when 100 loci were sampled than for the other summary statistics (**Supplementary Figure S2**). Sampling 1,000 and 15,000 random SNPs did not generate strong differences.

Sampling a different number of loci for the three empirical datasets showed similar patterns to the simulated datasets. In all cases, sampling fewer loci increased slightly the variance and the relative error across replicates for all estimates (**Figures 1–3** and **Supplementary Figures S3–S6**). Importantly, we found that the variance and relative error across replicates was higher for the microsatellite dataset (especially when sampling five microsatellites, **Supplementary Figure S6**), followed by the DTS dataset, and finally when sampling each subspecies using the 50K dataset (**Supplementary Figures S3–S5**).

Even though decreasing the number of loci increased the variance, it was interesting to note that estimated distributions fell close to the estimates for "real" datasets (**Supplementary Table S1**). For H_S and F_{IS}, sampling fewer microsatellite loci produced an important increase in the variance across replicates.

We also tested the effect of sampling a different number of loci in the estimation of landscape genetic statistics. In these tests we only found deviations with respect to the "real" datasets for the association between H_S and the geographic centroid, where we found that sampling fewer DTS and 50K SNPs reduced the association (β got closer to 0).

Importantly, sampling 1,000 or 15,000 of the 50K SNPs; 1,000 or 5,000 of the DTS SNPs; and 10 or 15 loci of the microsatellite dataset generated similar summary statistics and reduced the relative error in the estimation of the parameter (**Supplementary Figure S6**).

Varying the Number of Sampled Populations

Finally, we tested the effect of sampling a different number of populations in the estimation of summary statistics. For the simulated datasets, we found that sampling fewer populations increased the variance and relative error across the three summary statistics. For all summary statistics we found that the mean value across the 1,000 replicates of the different number of populations sampled were similar to the entire dataset, except for the distribution of F_{ST} in the hierarchical model with low gene flow where we found a lower mean value as fewer populations were sampled (**Supplementary Figure S2**).

For the empirical datasets, varying the number of populations generated similar mean values to those found for the "real" datasets. Also, for the three datasets, sampling a small number (5 and 10) of populations generated deviation and importantly increased the relative error compared to the real estimated values for all summary statistics and in particular for patterns of isolation by distance and environments and patterns associated with the tests of centrality (**Figures 2**, **3**).

The variance and relative error across datasets dropped after approximately 30 populations for the genomic datasets (see ranges in **Supplementary Table S1** and **Supplementary Figures S5, S6**), but remained high for the microsatellite dataset. The relative error was in general higher for samplings using the DTS dataset, except for patterns associated with the test of centrality, for which the 50K dataset presented a higher error (**Supplementary Figure S6**). Also we found that when sampling fewer populations, the microsatellite dataset presented a lower relative error than the 50K and DTS datasets (**Supplementary Figure S6**).

Importantly, sampling fewer populations generated a high variance and higher relative error in the estimation of all parameters, except for F_{IS} and F_{ST} for the microsatellite dataset (**Figures 1B,D**) and F_{IS} estimates for the 50K dataset (**Figure 1** and **Supplementary Figures S3–S6**). In these cases, the deviations across replicates when sampling fewer populations was lower than the variance generated by sampling a different number of individuals (**Figure 1** and **Supplementary Figure S6**).

For patterns of IBD and IBE, we also found that sampling fewer than 10 populations generated in some replicates incorrect association estimates. The real value showed positive associations, isolation by distance or environment (**Table 2**), but for 5 and 10 sampled populations we found that up to 18 and 7% of sample replicates generated negative associations, respectively (**Supplementary Table S2**; see changes in signs in **Supplementary Table S1**). For tests of association between H_S and ecological variables, this was even more sensitive for the genomic datasets, since we found up to 4.6% of positive estimates (when the entire dataset was negative) for 30 populations when testing association between *Hs* and the distance to the niche centroid. However, we found that the DTS was less sensitive to deviation for associations between ecological variables and summary statistics (**Figure 3**).

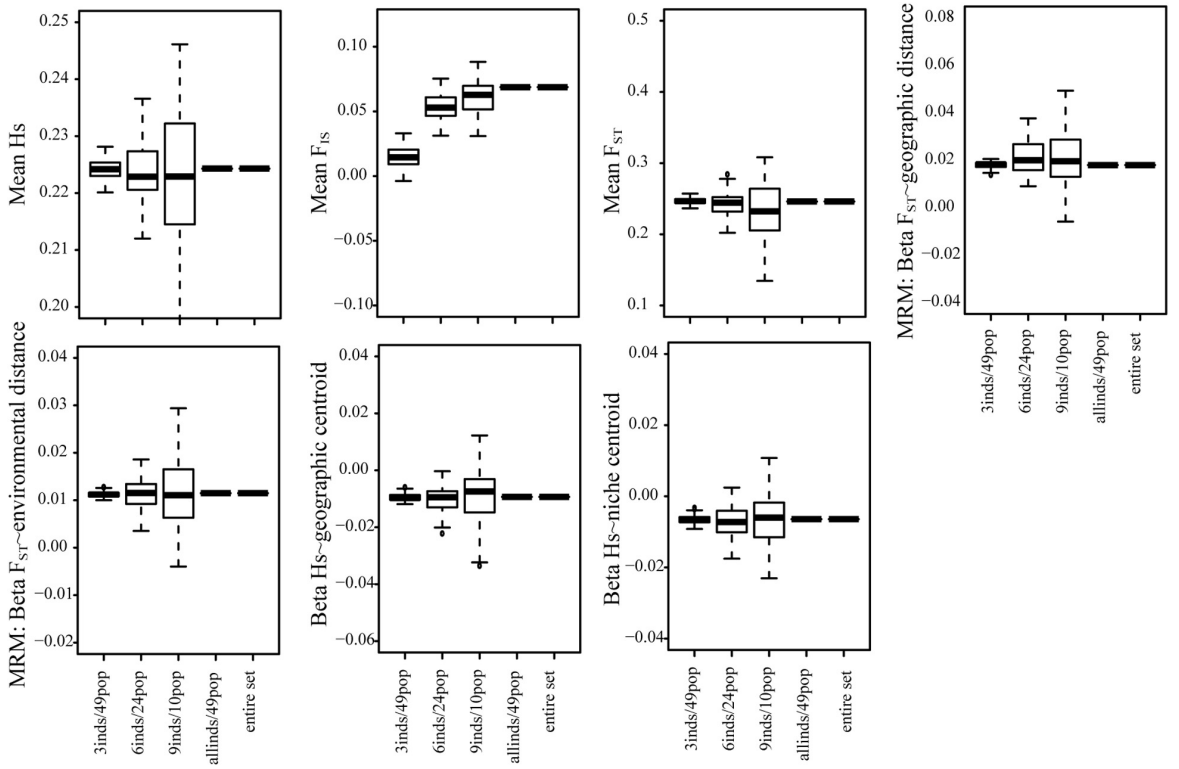

FIGURE 4 | The tradeoff between the number of individuals and the number of populations sampled for all summary statistics using the 50K dataset. We tested the effect of sampling more individuals in fewer populations and fewer individuals in many populations.

Finally, we found that the microsatellite dataset was more sensitive for isolation by distance and environment patterns (we found up to 30% of samples showed opposing patterns to the entire dataset when sampling 5 populations; **Supplementary Table S2**), and less sensitive for tests of CAH (we found up to 1.8% of opposing results when sampling 20 populations; **Supplementary Table S2**).

Tradeoff Between Number of Individuals and Number of Populations

For the 50K dataset, we also contrasted the effect of sampling fewer individuals in many populations or many individuals in fewer populations (3 individuals/49 populations, 6 individuals in 24 populations, and 9 individuals in 10 populations). For all summary statistics, except F_{IS}, we found that sampling more populations but fewer individuals generated more accurate results and lower biases (**Figure 4** and **Supplementary Figure S8**).

Varying the Number of Populations in the Identification of Candidate SNPs

Figure 5A shows that for 5 and 10 sampled populations, the maximum F_{ST} for a locus found by *bayescenv* across replicates was higher than for the rest of the sampling designs. We also tested the number of candidate SNPs across replicates. We found more candidate SNPs when sampling a higher number of populations (**Figure 5B**).

We also evaluated how many shared outlier SNPs were identified by all replicates and sampling designs. Interestingly, we only found 1 SNP that was identified for the 10 replicates of the 30 and 49 populations sampling designs (**Figure 5C**). The lack of shared SNPs could be explained by the nature of *Bayescenv*, that identifies genome by environment associations. If the populations that were sampled in each test have different ecological settings, we would not expect to find the same outlier SNPs. Therefore, we also analyzed each replicate independently to identify how many outlier SNPs were shared with the 49 sampling designs (**Supplementary Table S3**). For replicates of 5 and 10 populations, less than 10% of SNPs were shared with the SNPs identified for 49 populations. For 20 populations, one replicate identified 46% of shared SNPs with 49 populations; and for 30 populations 4 replicates identified > 51% of shared outlier SNPs (**Supplementary Table S3**).

DISCUSSION

It remains challenging for many researchers to generate large genomic samples, posing a tradeoff between the information obtained with MPS and the number of populations sampled (Meirmans, 2015). Here we used two simulated datasets to estimate the effect of sampling size on the estimation of

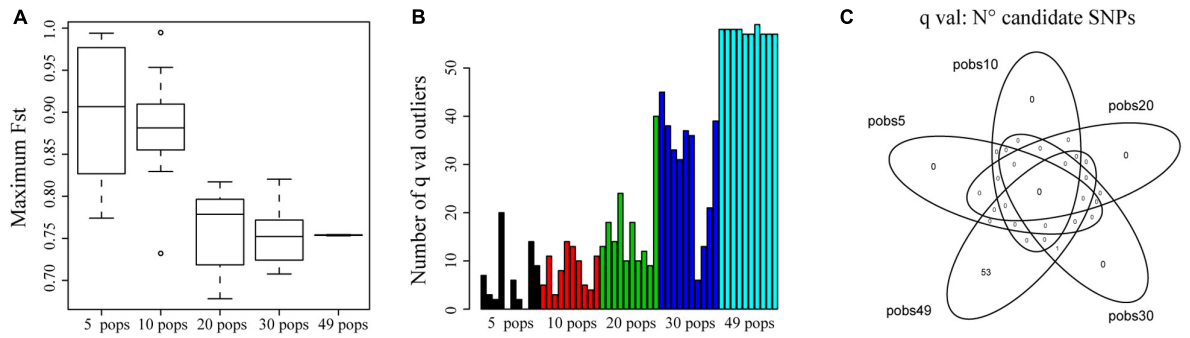

FIGURE 5 | The effect of sampling a different number of populations on the identification of outlier SNPs. **(A)** Distribution of the highest F_{ST} identified for a locus across simulations. **(B)** Number of outlier SNPs (*q-val*) identified for each replicate for a different number of sampled populations. **(C)** The Venn diagram shows the number of shared SNPs identified across replicates and the number of populations sampled using q-val to identify outliers.

summary statistics. To confirm the effect on simulated results, we used empirical genomic datasets and a microsatellite dataset obtained for a large sample of wild maize, the teosintes (*Zea maiz* ssp. *parviglumis* and *Zea mays* ssp. *mexicana*), to test if the deviations generated by different sampling designs, while estimating classic population genomics and landscape genomics estimates. Depending on the objectives, and the amount of data that can be produced using genomic platforms, we propose some suggestions for sampling designs that could be considered according to our results (**Table 3**). It is important to consider that these recommendations might be more reliable for species with life history traits similar to Mexican wild maize, and caution should be taken since life history can have an important effect in summary statistics (Hamrick and Godt, 1996; Nybom, 2004).

Comparing Datasets

We used Fastsimcoal 2 as a tool to simulate large samplings and complex demographic scenarios similar to teosintes. Except for F_{ST} estimations, we found that the values of genetic diversity and variance between populations (**Table 2**) were similar between the simulated and empirical dataset.

For the empirical datasets, we found that the most evident differences between datasets were associated to H_S, F_{ST}, and F_{IS} estimates. These differences are expected, given the properties of each dataset. First, we found that the microsatellites had the highest H_S and lowest F_{ST}, which is a well-known pattern, and can be explained by their large number of alleles and mutation rates (Ellegren, 2004). Second, the 50K dataset had higher genetic diversity and lower F_{ST} compared to the DTS dataset. This is concordant with the design of the 50K dataset to detect highly polymorphic SNPs in maize, and therefore has ascertainment bias (Albrechtsen et al., 2010). In contrast, the DTS dataset was generated using restriction enzymes (similar to GBS and other reduced representation methods, including RADtags), and has lower ascertainment bias (Sansaloni et al., 2011; Ren et al., 2015).

Sampling a Different Number of Loci

In general, it has been suggested that if fewer populations and individuals are sampled, increasing the number of loci can increase the accuracy of estimates (Oyler-McCance et al., 2013; Peterman et al., 2016; Flesch et al., 2018; see summary and references in **Table 1**). In this study, we observed that after increasing the number of SNPs from 1,000, or 10 microsatellites to 5,000 (DTS), 15,000 (50K) SNPs, or 20 microsatellite loci we found similar patterns for all summary statistics and reduced variance and relative errors estimates across replicates (**Figures 1–3** and **Supplementary Figure S6**).

These are interesting observations, since depending on the study, it may be convenient to reduce genome or microsatellite coverage to increase the number of sampled populations, especially when patterns of isolation and demographic history are analyzed. However, it is important to notice that if the objective is to find targets of selection then, increasing the number of SNPs is critical in detecting stronger neutral expectations and to reduce false positives (De Mita et al., 2013), as well as increase the probability of finding SNPs that fall within coding or regulating regions (Metzker, 2010; Ekblom and Galindo, 2011; Glenn, 2011).

Sampling a Different Number of Individuals

We were able to compare the effect of sampling a different number of individuals for the simulated, 50K, and microsatellite datasets. For these datasets, we corroborated that sampling fewer individuals increased the variation across sampling (Miyamoto et al., 2008; Sinclair and Hobbs, 2009; Hale et al., 2012; Sánchez-Montes et al., 2017; see summary and references in **Table 1**), but more importantly, it underestimated the F_{IS} inbreeding estimation and overestimated the F_{ST} (**Figure 1B** and **Supplementary Figures S2, S3, S6**).

These results suggest that for genomic datasets, as long as many populations are sampled, and *Hs*, F_{ST}, patterns of isolation, or patterns associated to ecological variables are tested, the number of individuals is not as sensitive as the number of populations sampled covering a large portion of the distribution (Willing et al., 2012; Landguth and Schwartz, 2014). In fact, we found that it is more convenient to sample fewer individuals in as many populations as possible than the opposite (**Figure 4** and **Supplementary Figure S8**).

TABLE 3 | Recommendations for sampling designs depending on study objectives.

Estimate	Number of individuals	Number of loci	Number of populations	Considerations
H_S	Not sensitive (> 6 individuals)	Sensitive (>1,000 SNP loci; > 15 microsatellite loci)	Sensitive (> 20 populations)	Increase the number of loci and populations. Genomic dataset is less sensitive than microsatellite dataset.
F_{IS}	Very sensitive (>9 individuals)	Sensitive (>1,000 SNP loci; > 15 microsatellite loci)	Sensitive (> 20 populations)	Increase the number of individuals over loci and populations. If fewer populations are available, increase the number of individuals in those populations.
F_{ST}	Microsatellite dataset was very sensitive (> 20 individuals). 50K dataset: Not sensitive (>9 individuals)	Not very sensitive (>1,000 SNPs; > 15 loci)	Sensitive (> 20 populations)	Increase the number of populations over the number of SNPs or individuals.
IBD and IBE MRM tests	Not sensitive (> 3 individuals)	Not sensitive (>1,000 SNPs, > 15 loci)	Very sensitive (>20 populations)	Sample as many populations as possible even if fewer individuals or loci are sampled.
CAH tests	Not very sensitive (> 3 individuals for genomic datasets; > 6 individuals for microsatellite datasets)	Sensitive depending on the dataset (>1,000 DTS SNPs, > 100 50K SNPs, > 15 microsatellite loci)	Very sensitive (> 30 populations)	Increasing the number of populations is more important than increasing the number of loci or individuals. Microsatellites are more sensitive than genomic datasets to the number of loci and individuals, although less sensitive to the number of populations sampled.
Tests of selection using bayescenv	Not tested	Sensitive (as many as possible)	Very sensitive (>30–40 populations)	As many SNPs as possible are needed to differentiate outlier loci, also to increase the probability of finding loci within selective regions. Increase as much as possible the number of populations, covering the largest geographic and environmental distribution. A possibility is to use pooled-sample DNA.

The numbers within parenthesis indicate how many individuals, populations, or loci should be sampled for different objectives according to our simulations.

On the contrary, studies that depend on F_{IS} values (i.e., genetic analyses in conservation studies; or studies that aim at detecting non-random mating), or that are performed using microsatellite data, should sample as many individuals as possible to reduce the bias generated by missing data (Flesch et al., 2018) or by identifying low frequency alleles (**Supplementary Figure S7**). If testing local adaptation is not a priority, then sampling fewer populations, but with many individuals (>20) might be more important, and with special focus on sampling many individuals belonging to populations that are of particular interest for the research group (i.e., threatened or vulnerable populations).

For endangered species for which fewer populations and individuals exist, if it is a priority to obtain their genetic parameters, new bioinformatics tools have been developed to estimate the demographic history based on fewer individuals sampled (Gronau et al., 2011; Schiffels and Wang, 2020). The problem is that these methods rely on large amounts of SNPs, which can be challenging to obtain if no reference genomes are available (Glenn, 2011). In such cases, it might be more important to conserve the few individuals that exist irrespective of their genetic diversity.

Sampling a Different Number of Populations

We tested the effect of randomly sampling a different number of populations for all datasets. Sampling above 10 populations did not generate strong deviations between sampling designs and the "real" sample for the three datasets. However, we found that the number of populations was strongly associated with the accuracy and a reduction in the relative error of the mean estimates across replicates (**Figures 1–3**, **Supplementary Figures S2–S6** and **Supplementary Table S1**). Sampling a different number of populations with the microsatellite dataset generated a lower variance and relative error across replicates than the genomic datasets when estimating patterns of isolation using the MRM test (**Figures 2A,B** and **Supplementary Figure S6**).

Importantly, we found that sampling fewer populations in some cases can result in opposite associations (negative instead of positive) compared to the real dataset for patterns of isolation and ecological associations (i.e., less than 10 populations for patterns of isolation, and less than 30 populations for ecological associations). Although these incorrect associations were recorded only for a few replicates (**Supplementary Table S2**), it is important to notice that an overestimation of false associations could result by not sampling the entire geographic and environmental distribution (see also Chao et al., 2014; Rico, 2017).

The fact that fewer populations increase variance across replicates of genomic datasets is important, because many genomic studies usually sample fewer populations in order to increase the genomic coverage (Meirmans, 2015). Our results are concordant with different studies performing simulations that have shown that increasing the number of populations increases

the accuracy in estimates of summary statistics and especially in landscape genetics studies (Schwartz and McKelvey, 2009; Landguth and Schwartz, 2014; see summary and references in **Table 1**). In fact, we found that it is more convenient to sample more populations with fewer individuals than fewer populations with many individuals (**Figure 4** and **Supplementary Figure S8**). Thus, we propose that if detecting local adaptation is not an objective and F_{IS} is not being measured, it is more important to sample many populations (∼30) even if fewer individuals per population are considered (**Figure 4**) and fewer SNPs are obtained.

Sampling a Different Number of Populations for Detecting Outlier SNPs

An important advantage of MPS is that it allows detecting candidate loci under selection. However, an important limitation of incorrect sampling while detecting candidate loci is that demographic history and complex genetic structure can increase false positives (Schoville et al., 2012; De Mita et al., 2013; Tiffin and Ross-Ibarra, 2014). Since adaptive loci could be important for conservation (Allendorf et al., 2010) and to respond to environmental change (Bay et al., 2018; Exposito-Alonso et al., 2018; Aguirre-Liguori et al., 2019b), many efforts have been made to reduce false positives and to better detect genes that could be under selection.

When we sampled fewer populations, it was interesting to notice that mean and maximum F_{ST} values across loci were higher (**Figure 5A**), supporting that sampling fewer populations reduced the efficacity of estimating adequate estimates of real F_{ST} patterns, increasing the potential of false positives (De Mita et al., 2013). Interestingly, we found that as more populations were sampled and F_{ST} was estimated more accurately, more outlier SNPs were identified (**Figure 5B**). However, it is important to notice that the majority of the replicates did not identify the same outlier SNPs than the "real" dataset (**Figure 5C** and **Supplementary Table S3**). While this can be associated with false negatives, we rather consider that the lack of shared SNPs could correspond to the identification of outlier loci associated with different ecological settings. However, it was relevant to note that even for 30 populations, where we expect more populations to be shared with the real dataset, we still found a replicate that had only 5 shared SNPs with the entire dataset. From these analyses, we conclude that increasing the number of populations (>30) and SNPs is very important for detecting candidate SNPs since it allows the genetic structure to be defined more accurately and increases the power of the analysis (De Mita et al., 2013); and that it is important to cover the largest geographic and environmental distribution. Also, it is relevant to consider that environmental settings can have important implications on the SNPs that are identified as outliers. Our tests did not identify a strong candidate across replicates. Therefore, if genetic rescue is an objective (i.e., for conservation), it is important to perform experimental studies to corroborate the relevance of candidate SNPs (Kardos and Shafer, 2018; Bell et al., 2019).

These are important observations, especially when not many populations can be sampled either because organisms have limited distributions (Chao et al., 2014; Smith and Wang, 2014) or because there is a tradeoff between the amount of SNPs that can be obtained using MPS and the number of populations that can be genotyped (Meirmans, 2015). Methods such as *Bayescenv* (De Villemereuil and Gaggiotti, 2015), *Bayescan* (Foll and Gaggiotti, 2008), and *Bayenv* (Coop et al., 2010) do not rely on genotype counts, but rather on allelic counts. Therefore, they are not sensitive to the correct estimates of F_{IS} and one alternative can be to use a pooled sample approach to increase the number of loci and the number of populations genotyped.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

JA-L and LEE conceived and designed the work. JA-L and JL-S performed and analyzed the genomic analyses, and wrote the first version of the manuscript. JG-P generated, performed, and analyzed the microsatellite datasets. All authors analyzed and interpreted the combined data and reviewed the manuscript.

FUNDING

JA-L thanks the "Programa de Doctorado en Ciencias Biomédicas, UNAM" and the scholarship provided by CONACYT (grant no. 255770). This work was funded by grants CB2011/167826 (Investigación Científica Básica), CN-10-393 (UC MEXUS-CONACYT), and M12?A03 ECOS Nord France – CONACYT-ANUIES 207571.

ACKNOWLEDGMENTS

We thank Yocelyn Gutiérrez Guerrero and Alberto Villasante Barahona for support in generating the microsatellite datasets. We thank Sarah Hearne and CIMMYT for generating the DTS dataset. We thank Erika Aguirre Planter and Laura Espinosa Asuar for technical support. Finally, we thank Gabriela Castellanos-Morales, Erika Aguirre-Planter, and the two reviewers for comments to the manuscript. This manuscript was submitted at *biorxiv* (https://www.biorxiv.org/content/10.1101/2020.03.06.980888v1.abstract).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00870/full#supplementary-material

REFERENCES

Aguirre-Liguori, J. A., Aguirre-Planter, E., and Eguiarte, L. E. (2016). "Genetics and ecology of wild and cultivated maize: domestication and introgression," in *Ethnobotany of Mexico*, eds R. Lira, A. Casa, and J. Blancas (New York, NY: Springer), 403–416. doi: 10.1007/978-1-4614-6669-7_16

Aguirre-Liguori, J. A., Gaut, B. S., Jaramillo-Correa, J. P., Tenaillon, M. I., Montes-Hernández, S., García-Oliva, F., et al. (2019a). Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (*Zea mays parviglumis* and *Zea mays mexicana*). *Mol. Ecol.* 28, 2814–2830. doi: 10.1111/mec.15098

Aguirre-Liguori, J. A., Ramírez-Barahon, S., Tiffin, P., and Eguiarte, L. E. (2019b). Climate change is predicted to disrupt patterns of local adaptation in wild and cultivated maize. *Proc. R Soc. Lond. B Biol. Sci.* 286:20190486. doi: 10.1098/rspb.2019.0486

Aguirre-Liguori, J. A., Tenaillon, M. I., Vázquez-Lobo, A., Gaut, B. S., Jaramillo-Correa, J. P., Montes-Hernandez, S., et al. (2017). Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Mol. Ecol.* 26, 4226–4240. doi: 10.1111/mec.14203

Albrechtsen, A., Nielsen, F. C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* 27, 2534–2547. doi: 10.1093/molbev/msq148

Allendorf, F. W., Hohenlohe, P. A., and Luikart, G. (2010). Genomics and the future of conservation genetics. *Nat. Rev. Genet.* 11, 697–709. doi: 10.1038/nrg2844

Archer, F., Adams, P., and Schneiders, B. (2016). STRATAG: An R package for manipulating, summarizing and analysing population genetic data. *Mol. Ecol. Resour.* 17, 5–11. doi: 10.1111/1755-0998.12559

Bay, R. A., Harrigan, R. J., Le Underwood, V., Gibbs, H. L., Smith, T. B., and Ruegg, K. (2018). Genomic signals of selection predict climate-driven populations declines in a migratory bird. *Science* 359, 83–86. doi: 10.1126/science.aan4380

Bell, D. A., Robinson, Z. L., Funk, W. C., Fitzpatrick, S. W., Allendorf, F. W., Tallmon, D., et al. (2019). The exciting potential and remaining uncertainties of genetic rescue. *T. Ecol. Evol.* 34, 1070–1079. doi: 10.1016/j.tree.2019.06.006

Bradburd, G. S., Ralph, P. L., and Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* 67, 3258–3273. doi: 10.1111/evo.12193

Chao, A., Gotelli, N., Hsieh, T., Sander, E., Ma, K., Colwell, R., et al. (2014). Rarefaction and extrapolation with Hill estimates: A framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* 84, 45–67. doi: 10.1890/13-0133.1

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., and Beaumont, M. A. (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. doi: 10.1534/genetics.110.118661

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185, 1411–1423. doi: 10.1534/genetics.110.114819

De Mita, S., Thuillet, A. C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., et al. (2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399. doi: 10.1111/mec.12182

De Villemereuil, P., and Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods Ecol. Evol.* 6, 1248–1258. doi: 10.1111/2041-210X.12418

Dubois, Q., Lebigre, C., Schtickzelle, N., and Turlure, C. (2017). Sex, size and timing: Sampling design for reliable population genetic analyses using microsatellite data. *Methods Ecol. Evol.* 9, 1036–1048. doi: 10.1111/2041-210X.12948

Eckert, C. G., Samis, K. E., and Lougheed, S. C. (2008). Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Mol. Ecol.* 17, 1170–1180. doi: 10.1111/j.1365-294X.2007.03659.x

Ekblom, R., and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107, 1–15. doi: 10.1038/hdy.2010.152

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435–445. doi: 10.1038/nrg1348

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet* 9:e1003905. doi: 10.1371/journal.pgen.1003905

Excoffier, L., and Foll, M. (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27, 1332–1334. doi: 10.1093/bioinformatics/btr124

Exposito-Alonso, M., Vasseur, F., Ding, W., Burbano, H. A., and Weigel, D. (2018). Genomic basis and evolutionary potential for extreme drought adaptation in *Arabidopsis thaliana*. *Nat. Ecol. Evol.* 2, 352–358. doi: 10.1038/s41559-017-0423-0

Flesch, E., Rotella, J., Thomson, J., Graves, T., and Garrot, R. (2018). Evaluating sample size to estimate genetic management metrics in the genomics era. *Mol. Ecol. Resour.* 18, 1077–1091. doi: 10.1111/1755-0998.12898

Foll, M., and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics* 180, 977–993. doi: 10.1534/genetics.108.092221

Fumagalli, M. (2013). Assessing the effect of sequencing Depth and sample size in population genetics inferences. *PLoS One* 8:e79667. doi: 10.1371/journal.pone.0079667

Fustier, M. A., Bradenburg, J. T., Boitard, S., Lapeyronnie, J., Eguiarte, L. E., Vigouroux, Y., et al. (2017). Signatures of local adaptation in lowland and highland teosintes from whole-genome sequencing of pooled samples. *Mol. Ecol.* 26, 2738–2756. doi: 10.1111/mec.14082

Fustier, M. A., Martinez-Ainsworth, N. E., Aguirre-Liguori, J. A., Venon, A., Corti, H., Rousselet, A., et al. (2019). Common gardens in teosintes reveal the establishment of a syndrome of adaptation to altitude. *PLoS Genet.* 15:e1008512. doi: 10.1371/journal.pgen.1008512

Gasca-Pineda, J., Gutierrez-Guerrero, Y. T., Aguirre-Planter, E., and Eguiarte, L. E. (2020). The role of historical and contemporary environmental factors in the distribution of genetic diversity in the teosinte in Mexico. *bioRxiv* [Preprint]. doi: 10.1101/820126

Glenn, T. C. (2011). Fieldguide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11, 759–769. doi: 10.1111/j.1755-0998.2011.03024.x

González-Ramos, J., Agell, G., and Uriz, M. (2015). Microsatellites from sponges genomes: the number necessary for detecting genetic structure in *Hemimycale columella* populations. *Aquat. Biol.* 24, 25–34. doi: 10.3354/ab00630

Goslee, S. C., and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *J. Stat. Softw.* 22, 1–19. doi: 10.18637/jss.v022.i07

Goudet, J. (2005). Hierfstat, a package for R to compute and test variance components and F-statistics. *Mol. Ecol. Notes.* 5, 184–186. doi: 10.1111/j.1471-8286.2004.00828.x

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* 43, 1031–1034. doi: 10.1038/ng.937

Hale, M., Burg, T., and Steeves, T. (2012). Sampling from microsatellite-based popullationgenetic studies: 25 to 30 individuals is enough to accurately estimate allele frecuencies. *PLoS One* 7:e45170. doi: 10.1371/journal.pone.0045170

Hamrick, J. L., and Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plant species. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 351, 1291–1298. doi: 10.1098/rstb.1996.0112

Heller, R., Chikhi, L., and Siegismund, H. R. (2013). The confounding effect of population structure on bayesian skyline plot inferences of demographic history. *PloS One* 8:e62992. doi: 10.1371/journal.pone.0062992

Hufford, M. B., Lubinksy, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., and Ross-Ibarra, J. (2013). The genomic signature of crop-wild introgression in maize. *PLoS Genet.* 9:e1003477. doi: 10.1371/journal.pgen.1003477

Hufford, M. B., Martínez-Meyer, E., Gaut, B. S., Eguiarte, L. E., and Tenaillon, M. I. (2012). Inferences from the historical distribution of wild and domesticated maize provide ecological and evolutionary insight. *PLoS One* 7:e47659. doi: 10.1371/journal.pone.0047659

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jones, G. R. (2019). Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst. Biol.* 68, 19–31. doi: 10.1093/sysbio/syy041

Kardos, M., and Shafer, A. B. A. (2018). The peril of gene-targeted conservation. *T. Ecol. Evol.* 33, 827–839. doi: 10.1016/j.tree.2018.08.011

Lack, J. B., Pfau, R. S., and Wilson, G. M. (2010). Demographic history and incomplete lineage sorting obscure population genetic structure of the Texas mouse (*Peromyscus attwateri*). *J. Mammal.* 91, 314–325. doi: 10.1644/09-MAMM-A-242.1

Landguth, E., Fedy, B., Oyler-McCance, S., Gareys, A., Emel, S., Mumma, M., et al. (2012). Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Mol. Ecol. Resour.* 12, 276–284. doi: 10.1111/j.1755-0998.2011.03077.x

Landguth, E., and Schwartz, M. (2014). Evaluating sample allocation and effort in detecting population differentiation for discrete and continuosly distributed individuals. *Conserv. Genet.* 15, 981–992. doi: 10.1007/s10592-014-0593-0

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecol.* 188, 177–131. doi: 10.1007/s11258-006-9126-3

Lira-Noriega, A., and Manthey, J. D. (2014). Relationship of genetic diversity and niche centrality: a survey and analysis. *Evolution* 68, 1082–1093. doi: 10.1111/evo.12343

Martínez-Meyer, E., Díaz-Porras, D., Peterson, T. A., and Yañez-Arenas, C. (2012). Ecological niche structure and range-wide abundance patterns of species. *Biol. Lett.* 9:20120637. doi: 10.1098/rsbl.2012.0637

Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them. *Mol. Ecol.* 24, 3223–3231. doi: 10.1111/mec.13243

Metzker, M. L. (2010). Sequencing technologies -the next generation. *Nat. Rev.* 11, 31–46. doi: 10.1038/nrg2626

Miyamoto, N., Fernández-Manjarrés, J., Morand-Prieur, M. E., and Frascaria-Lacoste, N. (2008). What sampling is needed for reliable estimates of genetic diversity in *Fraxinus excelsior* L. (Oleaceae)? *Ann. For. Sci.* 65:403. doi: 10.1051/forest:2008014

Moreno-Letelier, A., Aguirre-Liguori, J. A., Piñero, D., Vázquez-Lobo, A., and Eguiarte, L. E. (2020). The relevance of gene flow with wild relatives in understanding the domestication process. *Roy. Soc. Open Sci.* 7:191545. doi: 10.1098/rsos.191545

Nazareno, A., Bemmels, J., Dick, C., and Lohmann, L. (2017). Minimum samples sizes for population genomics: an empirical study from an amazonian plant species. *Mol. Ecol. Res.* 17, 1136–1147. doi: 10.1111/1755-0998.12654

Nybom, H. (2004). Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Mol. Ecol.* 13, 1143–1155. doi: 10.1111/j.1365-294X.2004.02141.x

Orozco-Terwengel, P., Corander, J., and Schloetterer, C. (2011). Genealogical lineage sorting leads to significant, but incorrect Bayesian multilocus inference of population structure. *Mol. Ecol.* 20, 1108–1121. doi: 10.1111/j.1365-294X.2010.04990.x

Oyler-McCance, S., Fedy, B., and Landguth, E. (2013). Sample design effects in landscape genetics. *Conserv. Genet.* 14, 275–285. doi: 10.1007/s10592-012-0415-1

Peterman, W., Brocato, E., Semlitsch, R., and Eggert, L. (2016). Reducing bias in population and landscape genetics inferences: The effects of sampling related individuals and multiple life stages. *PeerJ* 4:e1813. doi: 10.7717/peerj.1813

Pruett, C., and Winker, K. (2008). The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *J. Avian Biol.* 39, 252–256. doi: 10.1111/j.0908-8857.2008.04094.x

Puckett, E. E., and Eggert, L. S. (2016). Comparison of SNP and microsattellite genotyping panels for spatial assignament of individuals to natal range: a case of study using the American black bear (*Ursus americanus*). *Biol. Conserv.* 193, 86–93. doi: 10.1016/j.biocon.2015.11.020

Pyhäjärvi, T., Hufford, M. B., Mezmouk, S., and Ross-Ibarra, J. (2013). Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* 5, 1594–1609. doi: 10.1093/gbe/evt109

R Core Team (2019). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: R Core Team.

Ren, R., Ray, R., Li, P., Xu, J., Zhang, M., Liu, G., et al. (2015). Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. *Mol. Genet. Genomics* 290, 1457–1470. doi: 10.1007/s00438-015-0997-7

Rico, Y. (2017). Using computer simulation to asses sampling effects on spatial genetical structure in forest tre species. *New For.* 48, 225–243. doi: 10.1007/s11056-017-9571-y

Robinson, J. D., Coffman, A. J., Hickerson, M. J., and Gutenkunst, R. N. (2014). Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evol. Biol.* 14:254. doi: 10.1186/s12862-014-0254-4

Ross-Ibarra, J., Tenaillon, M., and Gaut, B. S. (2009). Historical divergence and gene flow in the genus Zea. *Genetics* 181, 1399–1413. doi: 10.1534/genetics.108.097238

Sánchez-Montes, G., Ariño, A., Vizmanos, J., Wang, J., and Martínez-Solano, I. (2017). Effects of sample size and full sibs on genetic diversity characterization: a case study of three syntopic Iberian Pond-Breeding amphibians. *J. Hered.* 108, 535–543. doi: 10.1093/jhered/esx038

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., et al. (2011). Diversity arrays technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proc.* 5:P54. doi: 10.1186/1753-6561-5-S7-P54

Schiffels, S., and Wang, K. (2020). "MSMC and MSMC2: the multiple sequentially markovian coalescent," in *Statistical Population Genomics. Methods in Molecular Biology*, ed. J. Dutheil (New York, NY: Humana). doi: 10.1007/978-1-0716-0199-0_7

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive genetic variation on the landscape: methods and cases. *Ann. Rev. Ecol. Evol. Syst.* 43, 23–43. doi: 10.1146/annurev-ecolsys-110411-160248

Schwartz, M., and McKelvey, K. (2009). Why sampling scheme matters: The effect of smpling scheme on lanscape genetic results. *Conserv. Genet.* 10, 441–452. doi: 10.1007/s10592-008-9622-1

Sinclair, E., and Hobbs, R. (2009). Samle size efects on estimates of population genetics structure: Implications for ecological restoration. *Restor. Ecol.* 17, 837–844. doi: 10.1111/j.1526-100X.2008.00420.x

Slatkin, M., and Voelm, L. (1991). F_{ST} in a hierarchical island model. *Genetics* 127, 627–629.

Smith, O., and Wang, J. (2014). When can noninvasive samples provide sufficient information in conservation genetics studies? *Mol. Ecol. Res.* 14, 1011–1023. doi: 10.1111/1755-0998.12250

Stapley, J., Reger, J., Feulner, P. G., Smadja, C., Galindo, J., Ekblom, R., et al. (2010). Adaptation genomics: the next generation. *Trends Ecol. Evol.* 25, 705–712. doi: 10.1016/j.tree.2010.09.002

Tiffin, P., and Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* 29, 673–680. doi: 10.1016/j.tree.2014.10.004

van Heerwaarden, J., Doebley, J. F., Briggs, W. H., Glaubitz, J. C., Goodman, M. M., Sánchez-González, J. J., et al. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1088–1092. doi: 10.1073/pnas.1013011108

van Meier, J. I., Sousa, V. C., Marques, D. A., Selz, O. M., Wagner, C. E., Excoffier, L., et al. (2017). Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia cichlid* species after hybridization. *Mol. Ecol.* 26, 123–141. doi: 10.1111/mec.13838

Weir, B., and Hill, W. (2002). Estimating F-Statics. *Ann. Rev. Genet.* 36, 721–750. doi: 10.1146/annurev.genet.36.050802.093940

Willing, E. M., Dreyer, C., and Van Oosterhout, C. (2012). Estimates of genetic fifferentiation measured by FST do not necessarely require large sample sizes when using many SNP markers. *PLoS One* 7:e42649. doi: 10.1038/sj.ejhg.5200519

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.