





RESOURCE ARTICLE

Sampling strategy optimization to increase statistical power in landscape genomics: A simulation-based approach

Oliver Selmoni  | Elia Vajana  | Annie Guillaume  | Estelle Rochat  |
Stéphane Joost 

Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Correspondence

Stéphane Joost, Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Email: stephane.joost@epfl.ch

Funding information

Horizon 2020 Framework Programme, Grant/Award Number: 677353

Abstract

An increasing number of studies are using landscape genomics to investigate local adaptation in wild and domestic populations. Implementation of this approach requires the sampling phase to consider the complexity of environmental settings and the burden of logistical constraints. These important aspects are often underestimated in the literature dedicated to sampling strategies. In this study, we computed simulated genomic data sets to run against actual environmental data in order to trial landscape genomics experiments under distinct sampling strategies. These strategies differed by design approach (to enhance environmental and/or geographical representativeness at study sites), number of sampling locations and sample sizes. We then evaluated how these elements affected statistical performances (power and false discoveries) under two antithetical demographic scenarios. Our results highlight the importance of selecting an appropriate sample size, which should be modified based on the demographic characteristics of the studied population. For species with limited dispersal, sample sizes above 200 units are generally sufficient to detect most adaptive signals, while in random mating populations this threshold should be increased to 400 units. Furthermore, we describe a design approach that maximizes both environmental and geographical representativeness of sampling sites and show how it systematically outperforms random or regular sampling schemes. Finally, we show that although having more sampling locations (between 40 and 50 sites) increase statistical power and reduce false discovery rate, similar results can be achieved with a moderate number of sites (20 sites). Overall, this study provides valuable guidelines for optimizing sampling strategies for landscape genomics experiments.

KEYWORDS

false discovery rate, landscape genomics, sample size, sampling strategy, statistical power

1 | INTRODUCTION

Landscape genomics is a subfield of population genomics, with the aim of identifying genetic variation underlying local adaptation in natural and managed populations (Balkenhol et al., 2017; Joost et al., 2007;

Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015). The approach consists of analysing genomic diversity and environmental variability simultaneously in order to detect genetic variants associated with a specific landscape composition. Studies of this kind usually incorporate an analysis of population structure, such that neutral genetic variation can

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2019 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

be distinguished from adaptive variation (Rellstab et al., 2015). Over the last few years, the landscape genomic approach has become more widely used (see Table 1; Balkenhol et al., 2017; Rellstab et al., 2015). It is being applied to a range of species, including livestock (Colli et al., 2014; Lv et al., 2014; Pariset, Joost, Marsan, & Valentini, 2009; Stucki et al., 2017; Vajana et al., 2018), wild animals (Harris & Munshi-South, 2017; Manthey & Moyle, 2015; Stronen et al., 2015; Wenzel, Douglas, James, Redpath, & Piertney, 2016), insects (Crossley, Chen, Groves, & Schoville, 2017; Dudaniec, Yong, Lancaster, Svensson, & Hansson, 2018; Theodorou et al., 2018), plants (Abebe, Naz, & Léon, 2015; De Kort et al., 2014; Pluess et al., 2016; Yoder et al., 2014) and aquatic organisms (DiBattista et al., 2017; Hecht, Matala, Hess, & Narum, 2015; Laporte et al., 2016; Riginos, Crandall, Liggins, Bongaerts, & Trembl, 2016; Vincent, Dionne, Kent, Lien, & Bernatchez, 2013).

Sampling strategy plays a pivotal role in experimental research, and must be theoretically tailored to the aim(s) of a study (Rellstab et al., 2015; Riginos et al., 2016). In the context of landscape genomics, the sampling design should cover a spatial scale representative of both the demographic processes and the environmental variability experienced by the study population (Balkenhol et al., 2017; Leempoel et al., 2017; Manel et al., 2010; Rellstab et al., 2015). This is imperative to be able to properly account for the confounding effect of population structure, to provide a biologically meaningful contrast between the environmental variables of interest and to definitely allow the search for actual adaptive variants (Balkenhol et al., 2017; Manel et al., 2010; Rellstab et al., 2015). Consequently, extensive field sampling is generally required and needs to be coupled with high-throughput genome sequencing to characterize samples at a large number of loci (Balkenhol et al., 2017; Rellstab et al., 2015). Beyond these theoretical aspects, pragmatic choices need to be made with regard to financial and logistical constraints that are often imposed (Manel et al., 2010; Rellstab et al., 2015). A sampling strategy consists of: (a) sampling design (the spatial arrangement of the sampling locations, D); (b) the number of sampling locations (L); and (c) sample size (the number of individuals sampled, N ; Table 1). The care with which these parameters are defined affects the scientific output of an experiment as well as its costs (Manel et al., 2010; Rellstab et al., 2015).

The landscape genomics community has traditionally focused on formulating theoretical guidelines for collecting individuals throughout the study area. In this literature, particular emphasis has been placed on how spatial scales and environmental variation should be accounted for when selecting sampling sites (Leempoel et al., 2017; Manel, Albert, & Yoccoz, 2012; Manel et al., 2010; Rellstab et al., 2015; Riginos et al., 2016). Theoretical simulations have shown that performing transects along environmental gradients or sampling pairs from contrasting sites that are spatially close reduced false discovery rates (FDRs) caused by demographic processes confounding effects (De Mita et al., 2013; Lotterhos & Whitlock, 2015). However, in these studies the environment was described using a single variable, which oversimplifies the choice of sampling sites. In fact, in a real landscape genomics application, several variables are usually analysed in order to explore a variety of possible environmental pressures causing selection (Balkenhol et al., 2017). The concurrent use of several environmental descriptors also allows us to control for

the bias associated with collinear conditions (Rellstab et al., 2015). Furthermore, these studies have focused on the comparison of different statistical methods with the drawback of confronting only a few combinations of the elements determining the sampling strategy (De Mita et al., 2013; Lotterhos & Whitlock, 2015). Last but not least, the number of samples used in the simulations (between 540 and 1,800; Lotterhos & Whitlock, 2015) appears to be unrealistic for use in most real landscape genomic experiments (Table 1) and thus the guidelines proposed are scarcely applicable in practice.

There is therefore a need to identify pragmatic and realistic guidelines such that a sampling strategy is designed to maximize statistical power, minimize false discoveries, and optimize efforts and financial expenses (Balkenhol et al., 2017; Rellstab et al., 2015). In particular, the fundamental questions that need to be addressed are: (a) how to determine the spatial arrangement of sampling locations; (b) how to organize sampling effort (for instance preferring many samples at a few sites, or rather fewer samples at many sites); and (c) how many samples are required to obtain sufficient statistical power (Rellstab et al., 2015; Riginos et al., 2016).

Here, we investigate how the outcome of landscape genomic analyses is driven by the sampling strategy. We ran simulations using a fictive genetic data set encompassing adaptive genotypes shaped by real environmental variables. The simulations accounted for antithetic demographic scenarios encompassing strong or weak population structure. We proposed sampling strategies that differed according to three elements: sampling design approach (D), number of sampling locations (L) and sample size (number of samples, N). For each of these three elements, we measured their relative impacts on the analyses' true positive rates (TPRs) and FDRs, as well as their impact on the predictive positive value (PPV; Marshall, 1989) of the strongest adaptive signals.

2 | MATERIAL AND METHODS

The iterative approach we designed to test the different sampling strategies required that a new genetic data set encompassing neutral and adaptive variation was created at every run of the simulations. A simulated genomic data set can be constructed by means of software performing coalescent (backward-in-time) or forward-in-time simulations (Carvajal-Rodríguez, 2008). However, methods using coalescent simulations (e.g., `SPLATCHE2`; Ray, Currat, Foll, & Excoffier, 2010) did not match our needs as they cannot compute complex selective scenarios (e.g., those involving multiple environmental variables; Carvajal-Rodríguez, 2008). We could not use forward-in-time methods either, as they are slow and therefore not compatible with the computational requirements of our simulative approach (Carvajal-Rodríguez, 2008). We therefore developed a customized framework in the `R` environment (version 3.3.1; R Core Team, 2016) to compute both neutral and adaptive genetic variation based on gradients of population membership and environmental variations, respectively (Figure 1). Before running the simulations across the complete data set (the multivariate environmental landscape of Europe), we tested our approach

TABLE 1 Sampling design in landscape genomics studies: a nonexhaustive list of landscape genomics studies, highlighting different species and their related sampling strategies

Study	Species	Sampling design (<i>D</i>)	Sampling locations (<i>L</i>)	Sample size (<i>S</i>)
Colli et al. (2014)	Goat	Spatial and breed representativeness	10 sites	43
Pariset et al. (2009)	Goat	Spatial and breed representativeness	16 regions	497
Stucki et al. (2017) and Vajana et al. (2018)	Cattle	Spatial representativeness	51 regions	813
Harris and Munshi-South (2017)	White-footed mouse	Habitat representativeness	6 sites	48
Stronen et al. (2015)	Wolf	Opportunistic, population representativeness	59 sites	59
Wenzel et al. (2016)	Red grouse	Spatial representativeness	21 sites	231
Crossley et al. (2017)	Potato beetle	Habitat representativeness	16 sites	192
Dudaniec et al. (2018)	Damselfly	Environmental and spatial representativeness	25 sites	426
Theodorou et al. (2018)	Red-tailed bumblebee	Habitat representativeness	18 sites	198
Abebe et al. (2015)	Barley	Spatial representativeness	10 regions	260
De Kort et al. (2014)	Black alder	Spatial and habitat representativeness	24 populations	356
Pluess et al. (2016)	European beech	Spatial and environmental representativeness	79 populations	234
Yoder et al. (2014)	Barrelclover	Spatial representativeness	202 sites	202
DiBattista et al. (2017)	Stripey snapper	Spatial representativeness	51 sites	1,016
Hecht et al. (2015)	Chinook salmon	Spatial representativeness	53 sites	1,956
Laporte et al. (2016)	European eel	Spatial and environmental representativeness	8 sites	179
Vincent et al. (2013)	Atlantic salmon	Spatial representativeness	26 ^a rivers	641 ^a

^aNumbers from the Vincent et al. report (2013) concerning the non-pooled samples.

on a reduced data set and compared it to a well-established forward-in-time simulation software (CDPOP, version 1.3; Landguth & Cushman, 2010). This step allowed us to define the optimal parameters required to simulate two types of demographic scenarios: panmictic (no dispersal constraints, random mating) and structured (dispersal and mating limited by distance).

We then proceeded with the simulations on the environmental data set of Europe. At each iteration, a new genetic background encompassing neutral and adaptive variation was computed (Figure 1, steps 1 and 2). Subsequently, a sampling strategy was applied as a combination of sampling design (*D*), number of sampling locations (*L*) and sample size (*N*) (Figure 1, steps 3–5), resulting in the generation of a genetic data set that, coupled with environmental data, underwent a landscape genomics analysis (Figure 1, step 6). At the end of each iteration, three diagnostic parameters were calculated: TPR (i.e., statistical power) and FDR for the analysis, as well as the PPV of the strongest genotype–environment associations (Figure 1, step 7).

At the end of the simulations, we analysed how each element of the sampling strategy (*D*, *L*, *N*) affected the rates of the three diagnostic parameters (TPR, FDR, PPV) under the two demographic scenarios (with or without dispersal constraints). All scripts and data used to perform this analysis are publicly available on Dryad (<https://doi.org/10.5061/dryad.m16d23c>).

2.1 | Environmental data

As a base for our simulations, we quantified the environmental settings of Europe (Figure S1). We retrieved eight climatic variables from publicly available sources (annual mean temperature, mean diurnal range, temperature seasonality, mean temperature of wettest quarter, annual precipitation, precipitation seasonality, precipitation of warmest quarter and altitude; Table S1; Hijmans, Cameron, Parra, Jones, & Jarvis, 2005; Ryan et al., 2009). In order to work on a relevant geographical scale (Leempoel et al., 2017) while maintaining an acceptable computational speed, the landscape was discretized into grid cells of 50 × 50 km, using QGIS toolbox (version 2.18.13; QGIS development team, 2009). This resulted in 8,155 landscape sites. Average values of environmental variables were computed for each cell of the landscape using the QGIS zonal statistics tool.

2.2 | Computation of genotypes

For the creation of the genotype matrices, we developed an R-pipeline based on probability functions to compute genotypes from population membership coefficients and environmental values (Box S1). The theoretical fundamentals of this method are based on the observation that when the population is structured, neutral alleles tend

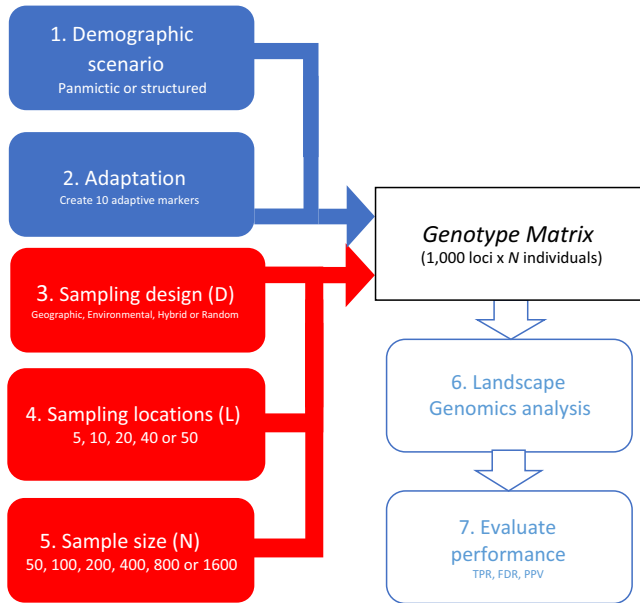


FIGURE 1 Workflow for each iteration of the simulative approach. The seven steps taken for every iteration. Starting with the blue boxes, the genetic set-up is established by selecting the demographic scenario (panmictic or structured), which determines the neutral structure, and by picking the environmental variables implied in adaptation. The environmental variable of interest and the strength of selection are randomly sampled for each of the 10 adaptive markers. Following this, the sampling strategy (here shown with red boxes) is set as a combination of design approach (geographical, environmental, hybrid or random), number of sampling locations (5, 10, 20, 40 or 50 locations) and sample size (50, 100, 200, 400, 800 or 1,600 samples). This results in the creation of a genotype matrix that undergoes a landscape genomics analysis. At the end of iterations, the statistical power (TPR) and false discovery rate (FDR) of the analysis and statistical predictive positive value of the strongest associations (PPV) are calculated to assess the performance of the sampling strategy [Colour figure can be viewed at wileyonlinelibrary.com]

to show similar spatial patterns of distribution (a feature commonly exploited in F_{ST} outlier tests; Luikart, England, Tallmon, Jordan, & Taberlet, 2003; and principal component analyses of genotype matrices; Novembre et al., 2008). Conversely, when a marker is under selection, its genotypic/allelic frequencies correlate with the environmental variable of interest (this is the basic concept of landscape genomics; see Balkenhol et al., 2017). For every iteration, 1,000 loci are computed: 10 are set to “adaptive,” and the remaining 990 to “neutral.” They are computed as follows:

2.2.1 | Neutral markers (Box S1a)

A parameter (m) is set to define the number of population membership gradients used in the simulations, where higher values of m result in more complex population structures. Every population membership gradient is simulated by randomly picking one to five landscape locations to represent the centre of the gradient. For each landscape location, the geographical distance to the gradient centres

(calculated using the R *dist* function) constitutes the membership coefficient. Next, a linear transformation converts this coefficient (Figure S1) for each sampling site into the probability of carrying a private allele for the population described ($pA|PS$). A second parameter (c , Box S2) defines this transformation, with values between 0.5 (random population structure) and 0 (strong population structure). The probability of $pA|PS$ is then used to draw (using the R-stat *sample* function) the bi-allelic genotype for each individual. This procedure is reiterated for every neutral locus assigned to a specific population membership coefficient. Each of the 990 neutral loci is then assigned to one of the m population membership coefficients (probability of assignment equal to $\frac{(1-c)}{\sum_{i=1}^m (1-c_i)}$) using the R *sample* function.

2.2.2 | Adaptive markers (Box S1b)

The probability of carrying an adaptive allele ($pA|Env$) is calculated through a linear transformation of a specific environmental gradient. This transformation is defined by two parameters. The first parameter (s_1) determines the amplitude of the transformation, and ranges between 0 (strong selective response) and 0.5 (neutral response; Box S2). The second parameter (s_2) shifts the baseline for allele frequencies, and ranges between -0.2 and 0.2 (weakening and strengthening the selective response, respectively; Box S2). Each of the 10 adaptive loci is randomly associated with one environmental variable. This implies that some environmental conditions can be associated with several genetic markers, and others with none. For every adaptive locus, the bi-allelic genotype is drawn (using the R-stat *sample* function) out of $pA|Env$.

2.3 | Evolutionary scenarios and parametrization

Two distinct demographic scenarios were chosen for this study: one involving a population that is not genetically structured (hereafter the “panmictic population scenario”), and one involving a structured population (hereafter the “structured population scenario”; see Box S2). To define the values of parameters m , c , s_1 and s_2 that allow the production of these two demographic scenarios, we ran a comparison of our customized simulation framework against simulations obtained using a well-established forward-in-time simulation software for landscape genetics called *cdPOP* (version 1.3; Landguth & Cushman, 2010).

This comparison was performed on a reduced data set composed of a 10-by-10 cell grid, covered with two dummy environmental variables extracted from the bioclim collection (Hijmans et al., 2005; Figure S1a,b). Each cell could host up to five individuals, where each individual was characterized at 200 single nucleotide polymorphisms (SNPs). In this set-up, we ran *cdPOP* using two distinct settings: the first that allowed for completely random dispersal and mating movements of individuals (i.e., panmictic population scenario), while the second setting restricted movements to neighbouring cells using a dispersal-cost based on distance (i.e., structured population scenario). In both scenarios, we applied

identical mortality constraints related to the two environmental variables, and set for each of them a genetic variant modulating fitness (Figure S1c,d). Fitness responses were constructed on an antagonistic pleiotropy model (i.e., adaptive tradeoffs, Lowry, 2012), using different intensities to represent moderate (Figure S1c) and strong selective constraints (Figure S1d). The following default CDPOP parameters were used for the remaining settings: five age classes with no sex-specific mortality, reproduction was sexual and with replacement, no genetic mutations, and epistatic effects or infections were allowed. The simulations ran for 100 generations and 10 replicates per demographic scenario were computed.

In parallel, we ran our customized algorithm to compute genotypes, using the same simplified data set as above. We iteratively tested all the possible combinations (hereafter “simulative variants”) of the parameters m (values tested: 1, 5, 10, 15, 20, 25), c (all possible ranges tested between: 0.1, 0.2, 0.3, 0.4, 0.5), s_1 (values tested: 0, 0.1, 0.2, 0.3, 0.4, 0.5) and s_2 (values tested: -0.2, -0.1, 0, 0.1, 0.2), and replicated each combination 10 times. Following this, we investigated which of the simulative variants provided the closest match with the allele frequencies observed in the CDPOP runs. The comparisons were based on three indicators of neutral structure:

2.3.1 | Principal component analysis (PCA) of the genotype matrix (Figure 2a)

A PCA of the genotype matrix was performed using the *prcomp* R function for each simulation (of both the CDPOP and the present customized method), where the differential of the variation explained by each principal component was then calculated. When the population is structured, the first principal component usually shows strong differences in the percentage of explained variation compared with the other components (Novembre et al., 2008). In contrast, when the population structure is absent, minor changes in this differential value emerge. The curve describing this differential value was then used for a pairwise comparison between the 10 replicates of each CDPOP scenario and the 10 replicates of each simulative variant (from the customized method). The curves were compared by calculating the root mean square error (RMSE), and the average RMSE was then used to rank simulative variants.

2.3.2 | F statistic (F_{ST} ; Figure 2b)

Five areas, which spanned four cells each, were selected to represent subpopulations of the study area: four areas located at the four corners of the 10-by-10 cell grid and the fifth located at the centre. For each simulation, we computed the pairwise F_{ST} (Weir & Cockerham, 1984) between these subpopulations using the *HIERFSTAT* R package (version 0.04; Goudet, 2005). An F_{ST} close to 0 indicates the absence of a genetic structure between subpopulations, while under a structured scenario this value tends to increase (Luikart et al., 2003). The distribution of all the F_{ST} values for the 10 CDPOP replicates was

compared to the distribution of the F_{ST} of 10 replicates of each simulative variant using the Kullback–Leibler divergence (KLD; Kullback & Leibler, 1951) analysis implemented in the *LAPLACESDEMON* R package (version 16.1.1; Statisticat & LCC, 2018). KLD was then used to rank simulative variants.

2.3.3 | Mantel test (Figure 2c)

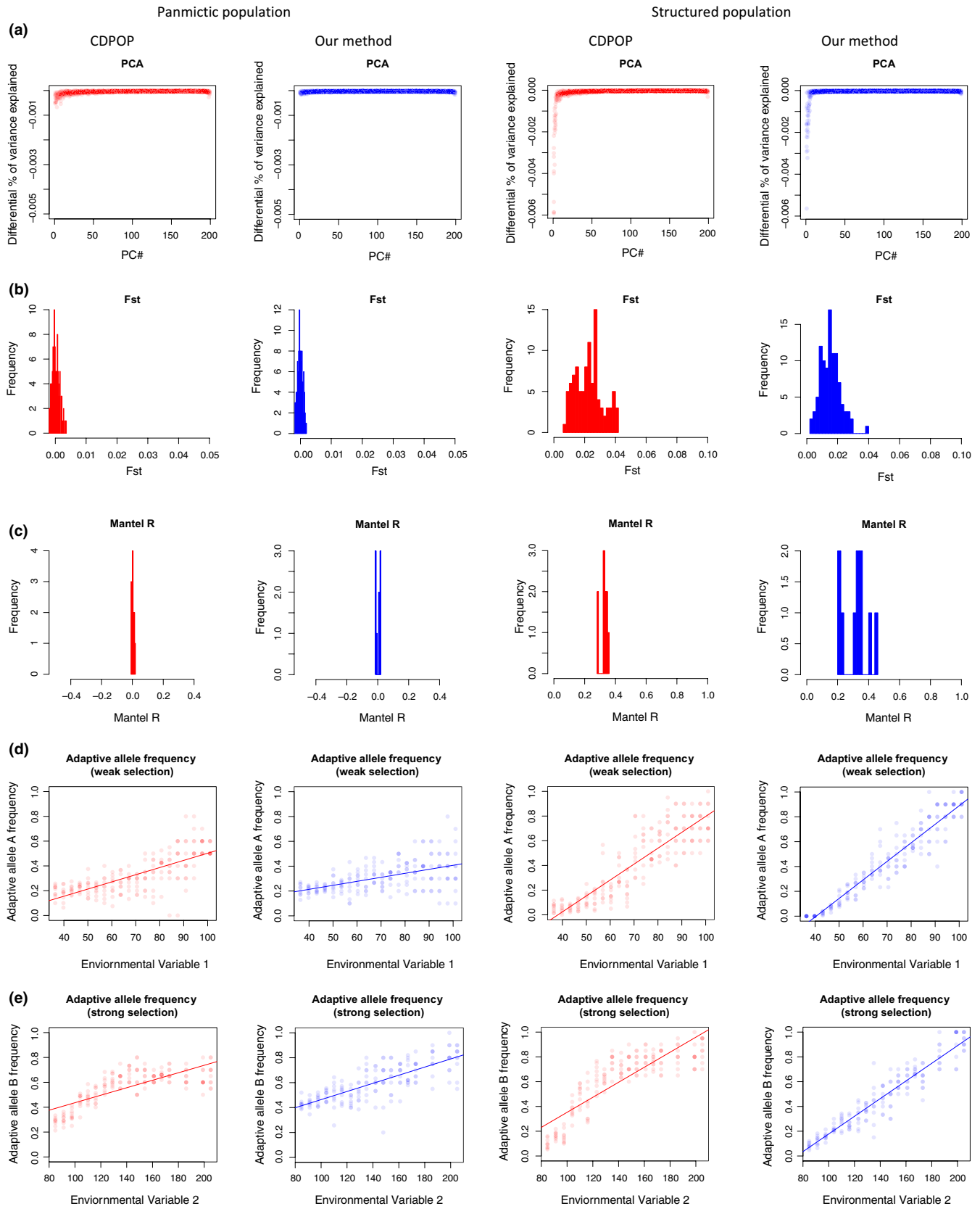
For each simulation, we computed the genetic and geographical distance between all individuals of the population applying the R *dist* function to the genotype matrix and the coordinates, respectively. Next, we calculated the Mantel correlation (mR; Mantel, 1967) between these two distance matrices using the *mantel.rtest* function implemented in the *ADE4* R package (version 1.7, Dray & Dufour, 2007). When mR is close to 0, it indicates the absence of correlation between the genetic and geographical distances, suggesting the absence of genetic structure (i.e., panmictic population scenario). In contrast, an mR closer to -1 or +1 indicates that genetic distances match geographical distances, as we would expect in a structured population scenario (Mantel, 1967). The average mR was calculated for each simulative variant and compared to the average mR measured in the two CDPOP scenarios. The resulting difference in mR (ΔmR) was used to rank simulative variants.

The three ranking coefficients (RMSE, KLD and ΔmR) were scaled using the *scale* R function and averaged, and the resulting value was used to rank simulative variants. In this way, it was possible to find one simulative variant with the best ranking when compared to the CDPOP panmictic population scenario, and another with the best ranking when compared to the CDPOP structured population scenario. These two simulative variants provided the values of m and c for the simulations on the complete data set.

Subsequently, we focused on the comparison of the values for the parameters defining the adaptive processes: s_1 and s_2 . For each CDPOP demographic scenario, we searched for the s_1 and s_2 combination that resulted in a simulative variant that best matched the allelic frequencies of each of the two genotypes implied in selection (moderate and strong). The environmental variable of interest was distributed in 20 equal intervals and within each interval the allelic frequencies of the adaptive genotype were computed. This resulted in the computation of a regression line for each simulation that described the allelic frequency of the adaptive genotype as a function of the environmental variable causing the selective constraint (Figure 2d,e). Next, we calculated the RMSE to compare this regression line between the CDPOP scenarios and the respective simulative variant (i.e., those with the optimal m and c according to the previous analyses) under different s_1 and s_2 combinations. For the two demographic scenarios, the ranges of s_1 and s_2 were ranked according to RMSE to represent a moderate to strong selection in the simulations for the complete data set.

2.4 | Sampling design

Four types of sampling design are proposed: three of them differently account for the characteristics of the landscape while one



randomly selects the sampling locations. The first is “geographical” (Figure 3a) and is defined through a hierarchical classification of the sites based on their geographical coordinates. The desired number of sampling locations (L) determines the number of clusters and the

geographical centre of each cluster is set as a sampling location. The goal of this strategy is to sample sites located as far apart as possible from each other in the geographical space to guarantee spatial representativeness.

FIGURE 2 Comparison of genotypes simulated with `CDPOP` and our method. Two distinct demographic scenarios were conceived, one with random mating (panmictic population) and one with dispersal costs related to distance (structured population). For each of them, `CDPOP` simulated the evolution of the population over 100 generations (red graphs) and replicated the same scenario 10 times. Simultaneously, we replicated the same scenarios using our simulative approach and show here the closest match (also replicated 10 times) to `CDPOP` simulations (blue graphs). Five methods for evaluating the genetic makeup are presented. (a) A principal component analysis is applied to the genotype matrix and the differential of the percentage of explained variation by each component is plotted for every replicate. (b) Pairwise F_{ST} analysis between five subpopulations is performed for every replicate and the resulting distribution of F_{ST} is shown. (c) Mantel correlation is calculated between a matrix of genetic and of geographical distances. The resulting Mantel R for every replicate is shown. (d, e) The allelic frequency of adaptive genotypes is shown as a function of the environmental variables causing selection (representing a case of moderate and strong selection, respectively) [Colour figure can be viewed at wileyonlinelibrary.com]

The second design type is “environmental” (Figure 3b). It is based on computation of distances depending on the values of environmental variables. The latter are first processed by a correlation filter: when two variables are found correlated to each other ($R > \pm 0.5$), one of them (randomly chosen) is excluded from the data set. The remaining uncorrelated descriptors are scaled ($SD = 1$) and centred (mean = 0) using the R *scale* function. The scaled values are used to perform a hierarchical clustering between the landscape sites. Like the previous design, the desired number of sampling locations (L) defines the number of clusters. For each cluster, the environmental centre is defined by an array containing the mean of the scaled environmental values. The Euclidean distances between this array and the scaled values of each site of the cluster are then computed. On this basis, the most similar sites to each centre are selected as sampling locations. This strategy aims to maximize environmental contrast between sampling locations and thus favours the detection of adaptive signals (Manel et al., 2012; Riginos et al., 2016).

The third design is “hybrid” (Figure 3c) and is a combination of the first two. It consists of dividing the landscape into k environmental regions and selecting within each of these regions two or more sampling locations based on geographical position. Initially, the environmental variables are processed as for the environmental design (correlation-filter and scaling) and used for hierarchical classification of the landscape sites. The next step is separating the landscape sites into k environmental regions based on this classification. The value of k allowed ranges between 2 and half of the desired number of sampling locations (L). We use the R package `NBCLUST` (version 3.0, Charrad, Ghazzali, Boiteau, & Niknafs, 2015) to find the optimal value of k within this range. The optimal k is then used to determine the k environmental regions. Next, the number of sampling locations (L) is equally divided across the k environmental regions. If k is not an exact divisor of L , the remainder of L/k is randomly assigned to environmental regions. The number of sampling locations per environment region (L_{ki}) can therefore be equal among environmental regions or, at worst, differ by 1 (e.g., if $L = 8$ and $k = 4$: $L_{k1} = 2$, $L_{k2} = 2$, $L_{k3} = 2$, $L_{k4} = 2$; if $L = 10$ and $k = 4$: $L_{k1} = 3$, $L_{k2} = 3$, $L_{k3} = 2$, $L_{k4} = 2$). Sampling locations within environmental regions are chosen based on geographical position. Geographical clusters within each environmental region are formed as in the geographical design, setting L_{ki} as the number of clusters. The landscape site spatially closer to the centre of each geographical cluster is selected as the sampling location. In this way, the procedure allows the replication of similar environmental

conditions at distant sites, therefore being expected to disentangle neutral and adaptive genetic variation and to promote the detection of variants under selection (Manel et al., 2012; Rellstab et al., 2015; Riginos et al., 2016).

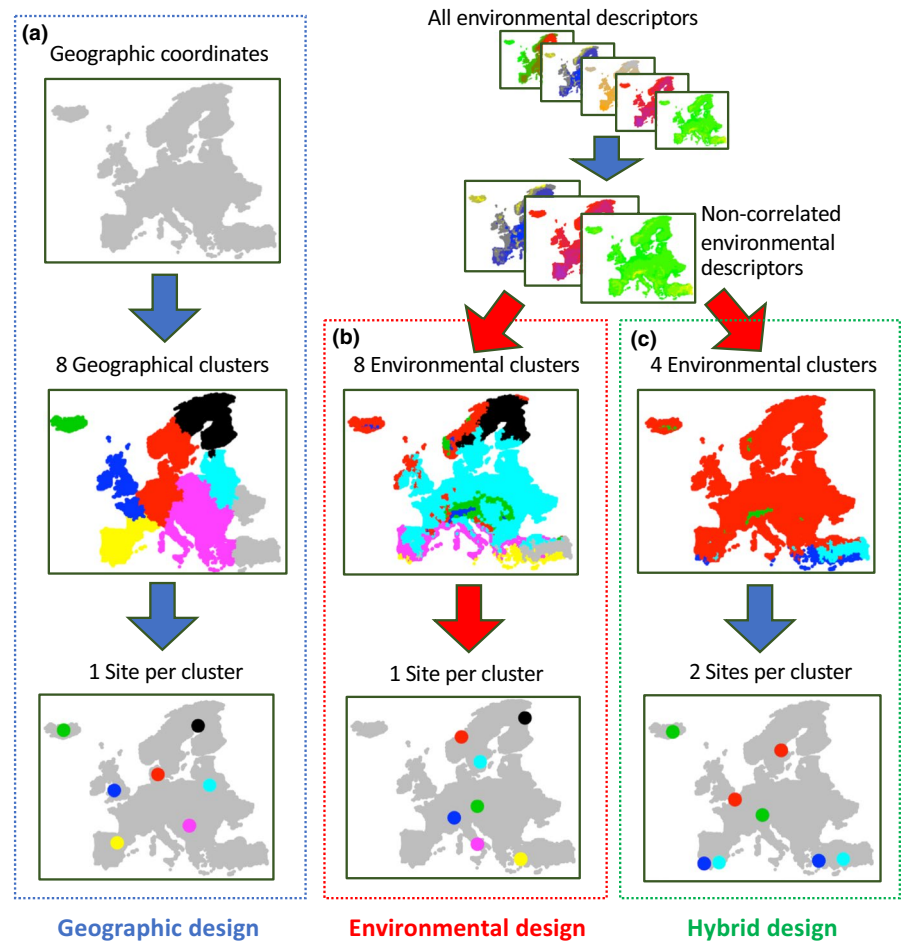
The fourth type of design is “random”: the sampling locations (L) are randomly selected across all the available landscape sites. In our simulations, we tested each type of sampling design with numbers comparable to those used in real experiments (see Table 1). We used five levels of sampling locations L (5, 10, 20, 40 and 50 locations) and six of sample sizes N (50, 100, 200, 400, 800 and 1,600 individuals). In iterations for which the sample size is not an exact multiple of the number of sites (e.g., 20 sites and 50 individuals), the total number of individuals was changed to the closest multiple (here 40 individuals). The scripts including these procedures were written in R using the functions embedded within the *stats* package (R Core Team, 2016).

2.5 | Landscape genomics analysis

We computed association models for each iteration with the *SAMBADA* software (version 0.6.0; Stucki et al., 2017). First, the simulated matrix of genotypes is filtered through a customized R function with minor allele frequency < 0.05 and major genotype frequency > 0.95 to avoid including rare or monomorphic alleles and genotypes, respectively. Second, a PCA is run on the filtered genotype matrix to obtain synthetic variables accounting for population structure (hereafter referred to as population structure variables; Patterson, Price, & Reich, 2006). Analysis of the eigenvalues of the PCA is carried out to assess whether the population structure is negligible for downstream analysis or not (Patterson et al., 2006). At each iteration, the algorithm runs a Tracy–Widom significance test of the eigenvalues, as implemented in the *ASSOCTESTS* R package (version 0.4, Wang, Zhang, Li, & Zhu 2017). Significant eigenvalues indicate the presence of non-negligible population structure: in these situations, the corresponding principal components will be used as covariables in the genotype–environment association study.

After filtering, *SAMBADA* is used to detect candidate loci for local adaptation. The software is able to run multivariate logistic regression models (Joost et al., 2007) that include population structure as a covariable, while guaranteeing fast computations (Duruz et al., 2019; Rellstab et al., 2015; Stucki et al., 2017). To ensure compatibility with our pipeline and increase computational speed, we integrated the *SAMBADA* method into a customized *PYTHON* script (version 3.5; Python

FIGURE 3 The three sampling design approaches accounting for landscape characteristics. The three maps illustrate how the eight sampling sites are chosen under three different sampling designs. Under a geographical strategy (a), sample location is selected using only geographical coordinates in order to maximize distance between sites. The environmental design (b) is computed using environmental variables (after filtering out highly correlated variables), in order to maximize the climatic distance between the chosen sites. The hybrid strategy (c) is a combination of the first two designs: first the landscape is divided into distinct environmental regions before choosing sites within each region that maximize spatial distance [Colour figure can be viewed at wileyonlinelibrary.com]



Software Foundation, 2018) based on the `PANDAS` (McKinney, 2010), `STATSMODELS` (Seabold & Perktold, 2010) and `MULTIPROCESSING` (Mckerns, Strand, Sullivan, Fang, & Aivazis, 2011) packages. Probability values related to the two statistics (G-score and Wald-score) associated with each association model are computed and subsequently corrected for multiple testing using the `R QVALUE` package (version 2.6; Storey, 2003). Models are deemed significant when showing $q < 0.05$ for both tests. When multiple models are found to be significant for the same marker, only the best one is kept (according to the G-score). The pipeline was developed in the R-environment using the `STATS` library.

2.6 | Simulations and evaluation of the performance

Each combination of demographic scenarios, sampling designs, number of sampling locations and sample sizes was replicated 20 times for a total of 4,800 iterations (Table 2). A new genetic matrix was randomly redrawn for each iteration to change the selective forces implying local adaptation and the demographic set-up determining the neutral loci. At the end of each iteration, three diagnostic parameters were computed:

- true positive rate of the analysis (TPR or statistical power): percentage of true associations detected to be significant;
- false discovery rate of the analysis (FDR): percentage of false association among those that are significant;
- positive predictive value (PPV; Marshall, 1989) of the 10 strongest associations: significant associations were sorted according to the association strength (β , the value of the parameter associated with environmental variable in the logistic model calculated by `sambada`). PPV represents the percentage of true associations among the best 10 associations according to β .

After the simulations, we calculated the median (Mdn) and interquartile range (IQR) of TPR, FDR and PPV under the different levels of the three elements underlying the sampling strategy (i.e., sampling design, number of sampling locations and sample size; Table 2). We also estimated how changes in these three elements drove alterations in TPR, FDR and PPV (i.e., effect size). We focused only on main effects (i.e., effects of single elements of the sampling strategy) because interactions effects (i.e., effects obtained combining two elements of the sampling strategy) appeared as minor after a preliminary visual inspection (Figure S2). Because TPR, FDR and PPV did not follow a normal distribution, we applied a bootstrap resampling technique ($r = 5,000$) to estimate their means and the related uncertainties under the different levels of each element of the sampling strategy (Dixon, 2006; Nakagawa & Cuthill, 2007). This step was performed in R, using the

boot library (version 1.3; Canty & Ripley, 2017; Davison & Hinkley, 1997). The effect size was then calculated as the difference in the mean values of TPR, FDR or PPV (and the related 95% confidence interval; Nakagawa & Cuthill, 2007) under different levels of the elements defining the sampling strategy. In the case of numerical elements (i.e., number of sampling locations and sample size), effect sizes were calculated as the changes in TPR, FDR and PPV along with the increments of the ordinal factor levels (e.g., change in TPR between sample sizes of 100 to 200, 200 to 400, 400 to 800, etc.). In the case of sample design, where the factor levels are not ordinal, we compared each design approach against a random sampling scheme.

3 | RESULTS

3.1 | Parameters of simulations

For the panmictic population scenario, the simulative variant best matching the $CDPOP$ results was obtained with the coefficients $m = 1$ and $c = 0.5$, whereas for the structured population scenario, the simulative variant was best at $m = 10$ and $c = \text{Unif}(0.2, 0.4)$ (Figure 2a–c; Box S2, Table S2a,b). In the panmictic population scenario, we found that the moderate selection case was best emulated by $s_1 = 0.4$ and $s_2 = -0.2$ and the strong selection by $s_1 = 0.3$ and $s_2 = +0.1$. In the structured population scenario, the moderate selection found its best match in the simulative variant with $s_1 = 0$ and

$s_2 = -0.1$, whereas the strong selection in the one set with $s_1 = 0$ and $s_2 = +0.2$ (Figure 2d,e; Box S2, Table S2c,d).

3.2 | True positive rate

In general, the panmictic population scenario simulations showed higher TPR ($\text{Mdn}_{\text{PAN}} = 40\%$ [IQR = 0%–90%]) than simulations performed under the structured population scenario ($\text{Mdn}_{\text{STR}} = 0\%$ [IQR = 0%–40%]; Figure 4a–c). For both scenarios, the largest effect sizes on TPR were generally related to changes in sample size (Table 3a). Smaller sample sizes ($N = 50, 100$) resulted in TPR close or equal to zero for both population scenarios (Figure 4c). Under the structured population scenario, an increase of TPR started from $N = 200$ (Table 3a), leading to an initial increase of ~4% TPR for every 10 additional samples. At $N = 400$, this increment progressively became less abrupt until reaching a maximal value at $N = 800$ ($\text{Mdn} = 100\%$ [IQR = 60%–100%]; Figure 4c; Table 3a). By comparison, the panmictic population scenario showed an increase in TPR starting at $N = 400$, with a more constant and less abrupt rate of increase (Figure 4c; Table 3a). Under this scenario, $N = 1,600$ was not sufficient to yield maximal TPR ($\text{Mdn} = 80\%$ [IQR = 60%–90%]; Figure 4c).

The effect sizes on TPR related to the number of sampling locations were less pronounced, compared to those of sample size (Table 3a; Figure 4b). Under both population scenarios, the largest increases in TPR were observed when passing from $L = 5$ to $L = 10$ (+7% and +32% TPR under panmictic and structured scenarios, respectively; Figure 4b; Table 3a). At higher numbers of sampling sites ($L = 20, 40$ and 50) the incremental rate of TPR was less evident but still positive under the structured scenario and close to zero under the panmictic one (Table 3a; Figure 4b).

Similar to the influence of sampling locations, the type of sampling design had a minor effect on TPR when compared to the effect of sample size (Table 3a; Figure 4a). When compared to the random approach, a hybrid design approach was seen to increase the TPR by +11% and +14% under panmictic and structured population scenarios, respectively (Figure 4a; Table 3a). Environmental design had slightly lower effect sizes on TPR (+10% and +12% under panmictic and structured population scenarios, respectively; Figure 4a; Table 3a), while those of geographical design were close to zero (Figure 4a; Table 3a).

3.3 | False discovery rate

False discoveries generally appeared at a higher rate under a panmictic population scenario ($\text{Mdn}_{\text{PAN}} = 100\%$ [IQR = 20%–100%]) than under a structured population scenario ($\text{Mdn}_{\text{STR}} = 63\%$ [IQR = 20%–100%]; Figure 4d–f). Sample size had the largest effects on FDR for both population scenarios (Table 3b; Figure 4f). For the panmictic population scenario, median FDR was 100% at smaller sample sizes ($N = 50, 100$ and 200 ; Figure 4f), but between $N = 200$ and $N = 400$, the FDR began to decrease by ~2% for every 10 additional samples taken (Table 3b). The reduction in FDR was

TABLE 2 Table of factors varying in the simulative approach

Factor	Number of levels	Levels
Demographic scenarios	2	Panmictic population, structured population
Sampling design (D)	4	Geographical, environmental, hybrid, random
Sampling locations (L)	5	5, 10, 20, 40, 50
Sampling size (N)	6	50, 100, 200, 400, 800, 1,600
Replicates	20	
Total	4,800	

Note: Two different demographic scenarios are possible, one in which there is no neutral genetic structure (panmictic population) and one in which there is a structured variation (structured population). We then used sampling strategies emulating those observed in real experiments. Three different sampling design approaches accounting for landscape characteristics are proposed: one maximizing the spatial representativeness of samples (geographical), one maximizing the environmental representativeness (environmental) and one that is a combination of both (hybrid). A fourth sampling design picks sampling locations randomly. The numerical ranges we used were comparable to those from real experiment: five levels for number of sampling locations spanning from five to 50 sites, and six levels of sample sizes (i.e., total number of samples) from 50 to 1,600 samples. For each combination of the aforementioned factors, 20 replicates were computed differing in the number and types of selective forces driving adaptation. In total, 4,800 simulations were computed.

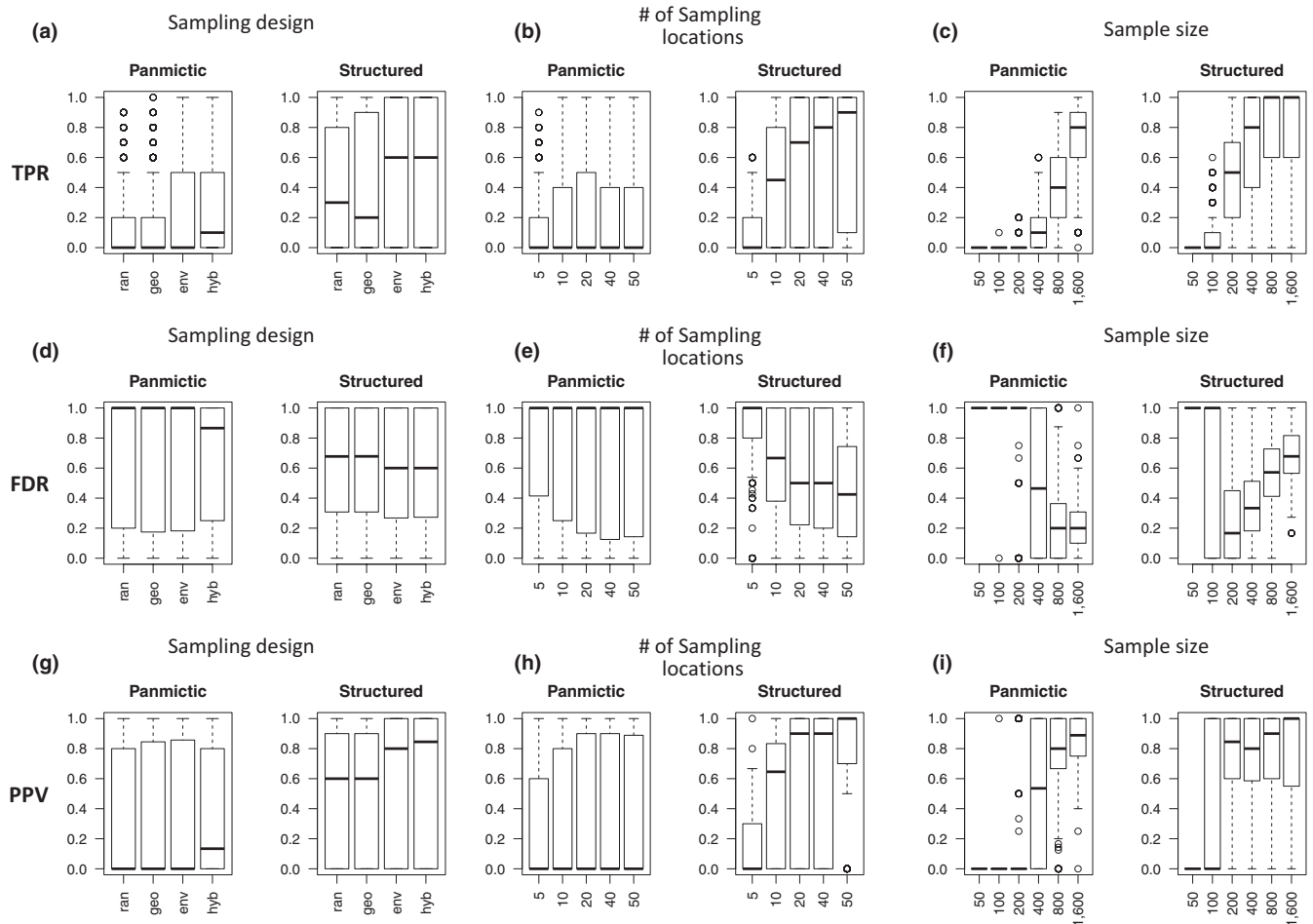


FIGURE 4 Effects of sampling strategy on the landscape genomics simulations. The plots display how the performance of landscape genomics experiments is driven by changes in the elements defining the sampling strategy. Three diagnostic parameters are used to measure the performance of each strategy: true positive rate (TPR; a–c) and false discovery rate (FDR; d–f) for the analysis and the positive predictive value of the 10 strongest significant association models (PPV; g–i). For each diagnostic parameter, we show the effect of sampling design (a, d, g; ran = random, geo = geographical, env = environmental, hyb = hybrid), number of sampling locations (b, e, h; 5, 10, 20, 40 or 50 sites) and sample size (c, f, i; 50, 100, 200, 400, 800 or 1,600 individuals) under two demographic scenarios: panmictic and structured population [Colour figure can be viewed at wileyonlinelibrary.com]

less abrupt after $N = 400$, and quasinull after $N = 800$ (Table 3b). At $N = 1,600$, median FDR was 20% [IQR = 10%–30%] (Figure 4f). The structured population scenario produced a different pattern: the largest median FDR was found at smaller sample sizes ($N = 50$ and 100), before a steep decrease was observed closer to $N = 200$ (Figure 4f; Table 3b). At larger sample sizes ($N = 400, 800$ and 1,600), FDR showed a logarithmic increase in growth rate where, at its most abrupt (between $N = 100$ and 400), there was an increase of +0.9% FDR for every 10 additional samples (Figure 4f; Table 3b). For the structured population scenario, $N = 1,600$ resulted in a median FDR of 68% [IQR = 57%–82%].

Under both population scenarios, the effect of sampling location number on FDR was weaker than the effect of sample size (Table 3b; Figure 4e). Similar to the pattern for TPR, the effects were stronger when passing from $L = 5$ to $L = 10$ (leading to a decrease of FDR of 7% and 23% under panmictic and structured population scenarios, respectively; Figure 4e; Table 3b), than between higher numbers of sampling locations ($L = 20, 40$ and 50; Table 3b).

Sampling design showed effects on FDR, but it was not as strong as the influences of sample size and sampling locations (Table 3b; Figure 4d). When compared to a random sampling scheme, both the environmental and the hybrid sampling designs showed comparable decreases in FDR (hybrid design: –3% and –6%; environmental design: –4% and –5% under panmictic and structured population scenarios, respectively), while the geographical design showed negligible changes (Figure 4d; Table 3b).

3.4 | Positive predictive value

The PPVs of the 10 strongest significant associations (hereafter simply referred to as PPV) were generally higher under the structured population scenario ($Mdn_{PAN} = 70\%$ [IQR = 0%–100%]) than under the panmictic population scenario ($Mdn_{PAN} = 0\%$ [IQR = 0%–80%]; Figure 4g–i). As with TPR and FDR, changes in sample size had the strongest influence on PPV under both population scenarios (Table 3c; Figure 4i). Under the panmictic population scenario,

TABLE 3 Effect sizes of elements defining the sampling strategy on TPR, FDR and PPV

Design	(a) TPR		(b) FDR		(c) PPV	
	Panmictic	Structured	Panmictic	Structured	Panmictic	Structured
Geo	0.0052 (0.0049–0.0056)	0.0101 (0.0094–0.0107)	-0.0167 (-0.0174 to -0.016)	0.0085 (0.0079–0.0091)	0.0158 (0.0151–0.0164)	0.0018 (0.0011–0.0025)
Env	0.1037 (0.1032–0.1042)	0.1173 (0.1167–0.118)	-0.045 (-0.0457 to -0.0444)	-0.0453 (-0.0459 to -0.0447)	0.0569 (0.0562–0.0576)	0.1139 (0.1132–0.1146)
Hyb	0.1145 (0.114–0.115)	0.1351 (0.1344–0.1357)	-0.0327 (-0.0333 to -0.032)	-0.0555 (-0.0561 to -0.0549)	0.048 (0.0473–0.0487)	0.1271 (0.1264–0.1278)
Nb. of sites						
5–10	0.0746 (0.0742–0.0751)	0.321 (0.3205–0.3215)	-0.0751 (-0.0757 to -0.0744)	-0.2288 (-0.2293 to -0.2283)	0.0849 (0.0842–0.0856)	0.3526 (0.3521–0.3532)
10–20	0.0281 (0.0275–0.0286)	0.1241 (0.1234–0.1248)	-0.0459 (-0.0466 to -0.0451)	-0.1008 (-0.1014 to -0.1001)	0.0453 (0.0445–0.046)	0.1326 (0.1319–0.1333)
20–40	-0.0043 (-0.0049 to -0.0038)	0.0286 (0.0278–0.0293)	-0.0021 (-0.0029 to -0.0013)	0.0063 (0.0057–0.007)	~0	-0.0132 (-0.0139 to -0.0124)
40–50	-0.0044 (-0.005 to -0.0038)	0.0353 (0.0345–0.0361)	0.0061 (0.0053–0.0068)	-0.0636 (-0.0642 to -0.0629)	-0.0097 (-0.0105 to -0.0089)	0.0759 (0.0751–0.0767)
Sample size						
50–100	~0	0.08 (0.0799–0.0802)	-0.0025 (-0.0026 to -0.0024)	-0.35 (-0.3506 to -0.3493)	0.0025 (0.0024–0.0026)	0.3502 (0.3495–0.3508)
100–200	0.0165 (0.0164–0.0165)	0.3911 (0.3906–0.3916)	-0.134 (-0.1345 to -0.1336)	-0.3488 (-0.3496 to -0.348)	0.1337 (0.1332–0.1342)	0.3582 (0.3574–0.359)
200–400	0.1089 (0.1087–0.109)	0.2126 (0.212–0.2132)	-0.3896 (-0.3904 to -0.3889)	0.0933 (0.0927–0.094)	0.39 (0.3893–0.3908)	0.0138 (0.0131–0.0144)
400–800	0.2801 (0.2798–0.2805)	0.0937 (0.0931–0.0944)	-0.2217 (-0.2224 to -0.2211)	0.1813 (0.1808–0.1818)	0.2264 (0.2258–0.2271)	0.025 (0.0244–0.0256)
800–1600	0.3031 (0.3026–0.3035)	0.0137 (0.013–0.0143)	-0.0411 (-0.0415 to -0.0406)	0.1107 (0.1103–0.1111)	0.0836 (0.0832–0.0841)	0.018 (0.0174–0.0186)

Note: The table shows the changes in the averages of the three diagnostic parameters of the analysis (TPR: true positive rate, a; FDR: false discovery rate, b; PPV: positive predictive value of among the 10 strongest significant association models, c) for every element determining the sampling strategy (sampling design, number of locations and sample size) under the two demographic scenarios, panmictic and structured population. In the case of sampling design, the changes refer to a comparison against the random sampling design (and positive values represent an increase in the diagnostic parameter in comparison to the random case, and vice versa for negative values). In situations concerning the number of sampling sites and sample size, two levels are compared and positive values indicate an increase in the diagnostic parameter under the second term of comparison (and vice versa for negative values). The 95% confidence intervals are shown in parentheses.

median PPV was ~0% for the smaller sample sizes ($N = 50, 100$ and 200 ; Figure 4i), after which patterns of increase were observed: from $N = 200$ to 400 there was an increase of PPV of ~+2% for every 10 additional samples, and from $N = 800$ to $1,600$ PPV continued to increase although it was less abrupt, resulting in a median PPV of 88% (IQR = 75%–100%) at $N = 1,600$ (Figure 4i, Table 3c). Under the structured population scenario, fewer samples were required to observe a similar increment: while PPV was close to 0 for $N = 50$ to $N = 100$, PPV increased by +7% for every 10 additional samples (Figure 4i; Table 3c). The increment of PPV became gradually weaker when transitioning between higher levels ($N = 400, 800$ and $1,600$) and led to a median PPV of 100% (IQR = 57.5%–100%) at $N = 1,600$.

Similar to TPR and FDR, the effect of sampling location number on PPV was weaker than that of sample size (Table 3c; Figure 4h). This effect was particularly evident under the structured population scenario, where an increase in the number of sampling locations increased PPV strongly (Figure 4h). The strongest PPV increment was observed between $L = 5$ and 10 , where each additional sampling location raised the PPV by +7% (Figure 4h; Table 3c). With more sampling locations ($L = 20, 40$ and 50) the rate of increase of PPV remained but was weaker (Figure 4h; Table 3c). In the panmictic population scenario, an increase in the number of sampling locations produced weaker changes in PPV (Figure 4h; Table 3c).

The sampling design used resulted in rate changes for PPV, despite being less strong than when compared to the other elements of the sampling strategy (Table 3c; Figure 4g). When compared to a random sampling scheme, the hybrid design and the environmental design increased PPV by +6% and +5% under the panmictic population scenario and by +13% and +12% under the structured one, respectively (Figure 4g; Table 3c). In contrast, geographical design did not result in pronounced changes of PPV (Figure 4g; Table 3c).

4 | DISCUSSION

The simulations presented in this study highlight that sampling strategy clearly drives the outcome of a landscape genomics experiment, and that the demographic characteristics of the studied species can significantly affect the analysis. Despite some limitations that will be discussed below, the results obtained make it possible to answer three questions that researchers are confronted with when planning this type of research investigation.

4.1 | How many samples are required to detect any adaptive signal?

In line with the findings of previous studies (e.g., Lotterhos & Whitlock, 2015), our results suggest that sample size is the key factor in securing the best possible outcome for a landscape genomics analysis. Where statistical power is concerned, there is an unquestionable advantage in increasing the number of samples under the scenarios tested. When focusing on the panmictic population scenario, we found a lack of statistical power in simulations for $N \leq 200$,

while detection of true positives increased significantly for $N \geq 400$ (Figure 4c). As we progressively doubled sample size ($N = 800, 1,600$), TPR linearly doubled as well (Figure 4c). Under the structured population scenario, this increase in statistical power started at $N \geq 100$ and followed a logarithmic trend that achieved the maximum power at $N \geq 800$ (Figure 4c).

These results show that it is crucial to consider the population's demographic background to ensure sufficient statistical power in the analyses, as advised by several reviews in the field (Balkenhol et al., 2017; Manel et al., 2012; Rellstab et al., 2015). In fact, the allelic frequencies of adaptive genotypes respond differently to a single environmental constraint under distinct dispersal modes (Figure 2d,e). When individual dispersal is limited by distance (structured population scenario), the allelic frequencies of adaptive genotypes are the result of several generations of selection, resulting in the progressive disappearance of nonadaptive alleles from areas where selection acts. When the dispersal of individuals is completely random (panmictic population scenario), the same selective force operates only within the last generation, such that even nonadaptive alleles can be found where the environmental constraint acts. Under these premises, a correlative approach for studying adaptation (such as *SAMBADA*) is more likely to find true positives under a structured population scenario rather than under a panmictic one.

The dichotomy between structured and panmictic populations also emerges when analysing FDRs. Under the panmictic population scenario, increasing the number of individuals sampled reduced FDR, while the inverse pattern was seen under a structured population scenario (Figure 4f). The issue of high false positive rates under structured demographic scenarios is well acknowledged in landscape genomics (De Mita et al., 2013; Rellstab et al., 2015). Population structure results in gradients of allele frequencies that can mimic, and be confounded by, patterns resulting from selection (Rellstab et al., 2015). As sample size increases, the augmented detection of true positives is accompanied by the (mis)detection of false positives. Under the panmictic population scenarios, these confounding gradients of population structure are absent (Figure 2a–c) and high sample sizes accentuate the detection of true positives only (Figure 4f).

Working with FDR up to 70% (Figure 4f) might appear excessive, but this should be contextualized in the case of landscape genomics experiments. The latter constitute the first step towards the identification of adaptive loci, which is generally followed by further experimental validations (Pardo-Díaz, Salazar, & Jiggins, 2015). Most landscape genomics methods test single-locus effects (Rellstab et al., 2015). This framework is efficient for detecting the few individual loci that provide a strong selective advantage, rather than the many loci with a weak individual effect (for instance those making up a polygenic adaptive trait; Pardo-Díaz et al., 2015). Therefore, when researchers are faced with a high number of significant associations, they tend to focus on the strongest ones (Rellstab et al., 2015), as we did here by measuring the PPV of the 10 strongest associations. By relying on this diagnostic parameter, we showed that increasing sample size ensures that the genotypes more strongly associated

with environmental gradients are truly due to adaptive associations (Figure 4i). Under these considerations, acceptable results are obtainable with moderate sample sizes: a median PPV of at least 50% was found with simulations with $N = 400$ and $N = 200$ under the panmictic and structured population scenario, respectively.

Each landscape genomic experiment is unique in terms of environmental and demographic scenarios, which is why it is not possible to propose a comprehensive mathematical formula to predict the expected TPR, FDR and PPV based solely on sample size. When working with a species with a presumed structured population (e.g., wild land animals), we advise against conducting experiments with fewer than 200 sampled individuals, as the statistical requirements to detect true signals are unlikely to be met. Panmixia is extremely rare in nature (Beveridge & Simmons, 2006), but long-range dispersal can be observed in many species such as plants (Nathan, 2006) and marine organisms (Riginos et al., 2016). When studying species of this kind, it is recommended to increase sample size to at least 400 units.

4.2 | How many sampling sites?

Increasing the number of samples inevitably raises the cost of an experiment, largely resulting from sequencing and genotyping costs (Manel et al., 2010; Rellstab et al., 2015). Additionally, fieldwork rapidly increases the cost of a study in cases where sampling has to be carried out across landscapes with logistic difficulties and physical obstacles. Therefore, it is both convenient and economical to optimize the number of sampling locations to control for ancillary costs.

De Mita et al. (2013) suggested that increasing the number of sampling locations would raise power and reduce false discoveries. The present study partially supports this view. A small number of sampling locations ($L = 5$) was found to reduce TPR and PPV while increasing FDR, compared to using more sampling locations ($L = 10, 20, 40$ and 50 ; Figure 4b,e,h). This is not surprising, because when sampling at a small number of locations the environmental characterization is likely to neglect some contrasts and ignore confounding effects between collinear variables (Leempoel et al., 2017; Manel et al., 2010). This was particularly evident under the structured population scenario (Figure 4b,e,h). In contrast, we found that higher numbers of sampling locations ($L = 40$ and 50) provided little benefits in terms of TPR, FDR and PPV, compared to a moderate number of locations ($L = 20$; Figure 4b,e,h). These discrepancies with previous studies are probably due to differences in the respective simulative approaches applied (we use several environmental descriptors instead of one) and the characteristics of the statistical method we employed to detect signatures of selection. In fact, as a number of sampling locations is sufficient to portray the environmental contrasts of the study area, adding more locations does not bring additional information and therefore does not increase statistical power. The implications of the information described above are considerable because the cost of fieldwork can be drastically reduced with marginal countereffects on statistical power and false discoveries.

4.3 | Where to sample?

Compared with random or opportunistic approaches, sampling designs based on the characteristics of the study area are expected to improve the power of landscape genomics analysis (Lotterhos & Whitlock, 2015). We developed three distinct methods to choose sampling locations accounting for geographical and/or environmental information (geographical, environmental and hybrid designs). We compared these design approaches between themselves and with random sampling schemes. The approach based on geographical position (geographical design) resulted in statistical power similar to the random designs (Figure 4a,d,f), while those based on climatic data (environmental and hybrid design) displayed remarkably higher TPRs and PPV and slightly lower FDR (Figure 4a,d,f). These beneficial effects on the analysis were accentuated under the structured demographic scenario.

These results match previous observations: methods conceived to take advantage of environmental contrasts facilitate the detection of adaptive signals (Manel et al., 2012; Riginos et al., 2016). Furthermore, the hybrid design prevents the sampling of neighbouring sites with similar conditions, therefore avoiding the superposition between adaptive and neutral genetic variation (Manel et al., 2012). This is likely to explain why the hybrid design slightly outperformed the environmental approach (Figure 4a,d,f).

Therefore, we strongly advise using a sampling scheme accounting for both environmental and geographical representativeness. Without bringing any additional cost to the analysis, this approach can boost statistical power by up to 14% under a complex demographic scenario (Table 3a), in comparison to a regular (geographical) or random sampling scheme.

4.4 | Limitation

The preliminary run of comparison with a well-established forward-in-time simulation software (CDPOP) showed the pertinence of our customized simulative approach (Figure 2). The neutral genetic variation appeared as random under the panmictic population scenario (no skew in the PCA graph, F_{ST} close to 0, mR close to 0) and structured under the structured population scenario (skew in the PCA graph, F_{ST} higher than 0, mR different from 0; Figure 2a–c). Adaptive allele frequencies also matched theoretical expectations: the responses along the environmental gradients were more stressed under the structured population scenario than under the panmictic one (Figure 2d,e).

Nonetheless, the use of forward-in-time simulations on the complete data set (used by De Mita et al., 2013; Lotterhos & Whitlock, 2015) would probably have resulted in more realistic scenarios. To be used in a framework such as that proposed here, the forward-in-time methods should be compatible with a large number of spatial locations (i.e., potential sampling sites), hundreds of individuals per location and a genetic data set counting at least 1,000 loci, of which 10 are set as adaptive against distinct

environmental variables. Importantly, all these requirements should be fulfilled at a reasonable computational speed (with our method, for instance, genotypes are computed in a few seconds). As far as we know, there is no existing software meeting these criteria.

The framework we presented here is based on an artificial genomic architecture encompassing 10 adaptive loci and 990 neutral loci. Given the generally high rates of false positives in landscape genomics (Balkenhol et al., 2017; Rellstab et al., 2015), it is hard to estimate a realistic percentage of SNPs implied in local adaptation from the literature. Besides, this percentage is driven by various factors specific to the biology of the studied species/population (e.g., life cycle duration, genome size, mutation rate, population size, extent of selective pressures; Dittmar, Oakley, Conner, Gould, & Schemske, 2016) and to the methods applied (e.g., genotyping strategy; Rellstab et al., 2015). Furthermore, not all adaptive genotypes are the same (Dittmar et al., 2016) and, as a consequence, diversified landscape genomics methods exist. Our framework relied on *SAMBADA*, a well-established method that assumes that (a) genotype–environment association follows a logistic response and (b) a few genotypes have large effects (Stucki et al., 2017). Not all the landscape genomics methods are based on these assumptions, however, and the guidelines described in this work might not be relevant for all these methods.

5 | CONCLUSIONS

The present work provides guidelines for optimizing the sampling strategy in the context of landscape genomic experiments. Our simulations highlight the importance of considering the demographic characteristics of the studied species when deciding the sampling strategy to be used. For species with limited dispersal, we suggest working with a minimum sample size of 200 individuals to achieve sufficient power for landscape genomic analyses. When species display long-range dispersal, this number should be raised to at least 400 individuals. The costs induced by a large number of samples can be balanced by reducing those related to fieldwork. In cases where a moderate number of sampling locations (20 sites) is sufficient to portray the environmental contrasts of the study area, there is only minimal statistical benefit in sampling a larger number of sites (40 or 50). Furthermore, we describe an approach for selecting sampling locations while accounting for environmental characteristics and spatial representativeness, and show its beneficial effects on the detection of true positives.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for useful comments and suggestions. We acknowledge funding from the IMAGE (Innovative Management of Animal Genetic Resources) project funded under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 677353.

AUTHOR CONTRIBUTIONS

O.S. and S.J. designed the research; O.S. performed the research; O.S., E.V., A.G., E.R. and S.J. analysed the results and wrote the paper. All the authors undertook revisions, contributed intellectually to the development of this manuscript and approved the final manuscript.

DATA AVAILABILITY STATEMENT

All scripts and data used to perform this analysis are publicly available on Dryad (<https://doi.org/10.5061/dryad.m16d23c>).

ORCID

Oliver Selmoni  <https://orcid.org/0000-0003-0904-5486>

Elia Vajana  <https://orcid.org/0000-0003-1340-3389>

Annie Guillaume  <https://orcid.org/0000-0002-2188-2861>

Estelle Rochat  <https://orcid.org/0000-0002-7978-5239>

Stéphane Joost  <https://orcid.org/0000-0002-1184-7501>

REFERENCES

- Abebe, T. D., Naz, A. A., & Léon, J. (2015). Landscape genomics reveal signatures of local adaptation in barley (*Hordeum vulgare* L.). *Frontiers in Plant Science*, 6(October), 813. <https://doi.org/10.3389/fpls.2015.00813>
- Balkenhol, N., Dudaniec, R. Y., Krutovsky, K. V., Johnson, J. S., Cairns, D. M., Segelbacher, G., ... Joost, S. (2017). Landscape genomics: Understanding relationships between environmental heterogeneity and genomic characteristics of populations. In O. Rajara (Eds.) *Population genomics* (pp. 261–322). Cham, Switzerland: Springer.
- Beveridge, M., & Simmons, L. W. (2006). Panmixia: An example from Dawson's burrowing bee (*Amegilla dawsoni*) (Hymenoptera: Anthophorini). *Molecular Ecology*, 15(4), 951–957. <https://doi.org/10.1111/j.1365-294X.2006.02846.x>
- Canty, A., & Ripley, B. (2017). *boot: Bootstrap R (S-Plus) Functions*. Retrieved from <https://CRAN.R-project.org/package=boot>
- Carvajal-Rodríguez, A. (2008). Simulation of genomes: A review. *Current Genomics*, 9(3), 155–159. <https://doi.org/10.2174/138920208784340759>
- Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2015). *NBCLUST: An R Package for Determining the relevant number of clusters in a data set*. *Journal of Statistical Software*, 61(6), 1–36. <https://doi.org/10.18637/jss.v061.i06>
- Colli, L., Joost, S., Negrini, R., Nicoloso, L., Crepaldi, P., Ajmone-Marsan, P., ... Zundel, S. (2014). Assessing the spatial dependence of adaptive loci in 43 European and Western Asian goat breeds using AFLP markers. *PLoS ONE*, 9(1), e86668. <https://doi.org/10.1371/journal.pone.0086668>
- Crossley, M. S., Chen, Y. H., Groves, R. L., & Schoville, S. D. (2017). Landscape genomics of Colorado potato beetle provides evidence of polygenic adaptation to insecticides. *Molecular Ecology*, 26(22), 6284–6300. <https://doi.org/10.1111/mec.14339>
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., & Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the

- tree species *Alnus glutinosa*. *Molecular Ecology*, 23(19), 4709–4721. <https://doi.org/10.1111/mec.12813>
- De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013). Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, 22(5), 1383–1399. <https://doi.org/10.1111/mec.12182>
- DiBattista, J. D., Travers, M. J., Moore, G. I., Evans, R. D., Newman, S. J., Feng, M., ... Berry, O. (2017). Seascape genomics reveals fine-scale patterns of dispersal for a reef fish along the ecologically divergent coast of Northwestern Australia. *Molecular Ecology*, 26(22), 6206–6223. <https://doi.org/10.1111/mec.14352>
- Dittmar, E. L., Oakley, C. G., Conner, J. K., Gould, B. A., & Schemske, D. W. (2016). Factors influencing the effect size distribution of adaptive substitutions. *Proceedings of the Royal Society B: Biological Sciences*, 283(1828), 20153065. <https://doi.org/10.1098/rspb.2015.3065>
- Dixon, P. M. (2006). Statistical and numerical computing. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics*. Chichester, UK: John Wiley & Sons.
- Dray, S., & Dufour, A.-B. (2007). The ADE4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4), 1–20. <https://doi.org/10.18637/jss.v022.i04>
- Dudaniec, R. Y., Yong, C. J., Lancaster, L. T., Svensson, E. I., & Hansson, B. (2018). Signatures of local adaptation along environmental gradients in a range-expanding damselfly (*Ischnura elegans*). *Molecular Ecology*, 27(11), 2576–2593. <https://doi.org/10.1111/mec.14709>
- Duruz, S., Sevane, N., Selmoni, O., Vajana, E., Leempoel, K., Stucki, S., ... Joost, S. (2019). Rapid identification and interpretation of gene-environment associations using the new R.SAMBADA landscape genomics pipeline. *Molecular Ecology Resources*, 19, 1355–1365. <https://doi.org/10.1111/1755-0998.13044>
- Goudet, J. (2005). HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, 5(1), 184–186. <https://doi.org/10.1111/j.1471-8286.2004.00828.x>
- Harris, S. E., & Munshi-South, J. (2017). Signatures of positive selection and local adaptation to urbanization in white-footed mice (*Peromyscus leucopus*). *Molecular Ecology*, 26(22), 6336–6350. <https://doi.org/10.1111/mec.14369>
- Hecht, B. C., Matala, A. P., Hess, J. E., & Narum, S. R. (2015). Environmental adaptation in Chinook salmon (*Oncorhynchus tshawytscha*) throughout their North American range. *Molecular Ecology*, 24(22), 5573–5595. <https://doi.org/10.1111/mec.13409>
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), 1965–1978. <https://doi.org/10.1002/joc.1276>
- Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., & Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation. *Molecular Ecology*, 16(18), 3955–3969. <https://doi.org/10.1111/j.1365-294X.2007.03442.x>
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. <https://doi.org/10.1214/aoms/1177729694>
- Landguth, E. L., & Cushman, S. A. (2010). CDPOP: A spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, 10(1), 156–161. <https://doi.org/10.1111/j.1755-0998.2009.02719.x>
- Laporte, M., Pavey, S. A., Rougeux, C., Pierron, F., Lauzent, M., Budzinski, H., ... Bernatchez, L. (2016). RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic eels. *Molecular Ecology*, 25(1), 219–237. <https://doi.org/10.1111/mec.13466>
- Leempoel, K., Duruz, S., Rochat, E., Widmer, I., Orozco-terWengel, P., & Joost, S. (2017). Simple rules for an efficient use of geographic information systems in molecular ecology. *Frontiers in Ecology and Evolution*, 5, 33. <https://doi.org/10.3389/fevo.2017.00033>
- Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, 24(5), 1031–1046. <https://doi.org/10.1111/mec.13100>
- Lowry, D. B. (2012). Local adaptation in the model plant. *New Phytologist*, 194(4), 888–890. <https://doi.org/10.1111/j.1469-8137.2012.04146.x>
- Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, 4(12), 981–994. <https://doi.org/10.1038/nrg1226>
- Lv, F.-H., Agha, S., Kantanen, J., Colli, L., Stucki, S., Kijas, J. W., ... Ajmone Marsan, P. (2014). Adaptations to climate-mediated selective pressures in sheep. *Molecular Biology and Evolution*, 31(12), 3324–3343. <https://doi.org/10.1093/molbev/msu264>
- Manel, S., Albert, C. H., & Yoccoz, N. G. (2012). Sampling in landscape genomics. *Methods in Molecular Biology*, 888, 3–12. https://doi.org/10.1007/978-1-61779-870-2_1
- Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storer, A., Rosenberg, M. S., ... Fortin, M.-J. (2010). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, 19(17), 3760–3772. <https://doi.org/10.1111/j.1365-294X.2010.04717.x>
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2), 209–220. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/6018555>
- Manthey, J. D., & Moyle, R. G. (2015). Isolation by environment in White-breasted Nuthatches (*Sitta carolinensis*) of the Madrean Archipelago sky islands: A landscape genomics approach. *Molecular Ecology*, 24(14), 3628–3638. <https://doi.org/10.1111/mec.13258>
- Marshall, R. J. (1989). The predictive value of simple rules for combining two diagnostic tests. *Biometrics*, 45(4), 1213. <https://doi.org/10.2307/2531772>
- Mckerns, M. M., Strand, L., Sullivan, T., Fang, A., & Aivazis, M. A. G. (2011). Building a framework for predictive science. *Proc. of the 10th Python In Science Conf*. Retrieved from <https://arxiv.org/abs/1202.1056>
- McKinney, W. (2010). *Data structures for statistical computing in PYTHON*. Retrieved from <http://conference.scipy.org/proceedings/scipy2010/mckinney.html>
- Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, 82(4), 591–605. <https://doi.org/10.1111/j.1469-185X.2007.00027.x>
- Nathan, R. (2006). Long-distance dispersal of plants. *Science*, 313(5788), 786. <https://doi.org/10.1126/science.1124975>
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218), 98–101. <https://doi.org/10.1038/nature07331>
- Pardo-Diaz, C., Salazar, C., & Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, 6(4), 445–464. <https://doi.org/10.1111/2041-210X.12324>
- Pariset, L., Joost, S., Marsan, P., & Valentini, A. (2009). Landscape genomics and biased F_{ST} approaches reveal single nucleotide polymorphisms under selection in goat breeds of North-East Mediterranean. *BMC Genetics*, 10(1), 7. <https://doi.org/10.1186/1471-2156-10-7>
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12), 2074–2093. <https://doi.org/10.1371/journal.pgen.0020190>
- Pluess, A. R., Frank, A., Heiri, C., Lalagüe, H., Vendramin, G. G., & Oddou-Muratorio, S. (2016). Genome-environment association study suggests local adaptation to climate at the regional scale in *Fagus sylvatica*. *New Phytologist*, 210(2), 589–601. <https://doi.org/10.1111/nph.13809>
- Python Software Foundation. (2018). *Python language reference, version 3.5*. Retrieved from www.python.org

- QGIS development team (2009). QGIS geographic information system. Open source geospatial foundation project. Retrieved from <http://www.qgis.org/>
- R Core Team. (2016). R: A language and environment for statistical computing. Retrieved from <https://www.r-project.org/>
- Ray, N., Currat, M., Foll, M., & Excoffier, L. (2010). SPLATCHE2: A spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics*, 26(23), 2993–2994. <https://doi.org/10.1093/bioinformatics/btq579>
- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, 24(17), 4348–4370. <https://doi.org/10.1111/mec.13322>
- Riginos, C., Crandall, E. D., Liggins, L., Bongaerts, P., & Tremblay, E. A. (2016). Navigating the currents of seascape genomics: How spatial analyses can augment population genetic studies. *Current Zoology*, 62, <https://doi.org/10.1093/cz/zow067>
- Ryan, W. B. F., Carbotte, S. M., Coplan, J. O., O'Hara, S., Melkonian, A., Arko, R., ... Zensky, R. (2009). Global multi-resolution topography synthesis. *Geochemistry, Geophysics, Geosystems*, 10(3), n/a–n/a. <https://doi.org/10.1029/2008GC002332>
- Seabold, S., & Perktold, J. (2010). Statsmodels: econometric and statistical modeling with Python. In *9th Python in Science Conference* (pp. 57–61). Retrieved from <http://statsmodels.sourceforge.net/>
- Statistat & LCC. (2018). *LaplacesDemon: Complete environment for Bayesian inference*. Retrieved from Bayesian-Inference.com
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31(6), 2013–2035. <https://doi.org/10.1214/aos/1074290335>
- Stronen, A. V., Jędrzejewska, B., Pertoldi, C., Demontis, D., Randi, E., Niedziałkowska, M., ... Czarnomska, S. D. (2015). Genome-wide analyses suggest parallel selection for universal traits may eclipse local environmental selection in a highly mobile carnivore. *Ecology and Evolution*, 5(19), 4410–4425. <https://doi.org/10.1002/ece3.1695>
- Stucki, S., Orozco-terWengel, P., Bruford, M. W., Colli, L., Masembe, C., Negrini, R., ... NEXTGEN Consortium (2017). High performance computation of landscape genomic models integrating local indices of spatial association. *Molecular Ecology Resources*, 17(5), 1072–1089. <https://doi.org/10.1111/1755-0998.12629>
- Theodorou, P., Radzevičiūtė, R., Kahnt, B., Soro, A., Grosse, I., & Paxton, R. J. (2018). Genome-wide single nucleotide polymorphism scan suggests adaptation to urbanization in an important pollinator, the red-tailed bumblebee (*Bombus lapidarius* L.). *Proceedings of the Royal Society B: Biological Sciences*, 285(1877), 20172806. <https://doi.org/10.1098/rspb.2017.2806>
- Vajana, E., Barbato, M., Colli, L., Milanese, M., Rochat, E., Fabrizi, E., ... Ajmone-Marsan, P. (2018). Combining landscape genomics and ecological modelling to investigate local adaptation of indigenous Ugandan cattle to East Coast fever. *Frontiers in Genetics*, 9, 385. <https://doi.org/10.3389/FGENE.2018.00385>
- Vincent, B., Dionne, M., Kent, M. P., Lien, S., & Bernatchez, L. (2013). Landscape genomics in atlantic salmon (*salmo salar*): Searching for gene–environment interactions driving local adaptation. *Evolution*, 67(12), 3469–3487. <https://doi.org/10.1111/evo.12139>
- Wang, L., Zhang, W., Li, Q., & Zhu, W. (2017). *AssocTests: Genetic association studies*. Retrieved from <https://CRAN.R-project.org/package=AssocTests>
- Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6), 1358. <https://doi.org/10.2307/2408641>
- Wenzel, M. A., Douglas, A., James, M. C., Redpath, S. M., & Pieltney, S. B. (2016). The role of parasite-driven selection in shaping landscape genomic structure in red grouse (*Lagopus lagopus scotica*). *Molecular Ecology*, 25(1), 324–341. <https://doi.org/10.1111/mec.13473>
- Yoder, J. B., Stanton-Geddes, J., Zhou, P., Briskine, R., Young, N. D., & Tiffin, P. (2014). Genomic signature of adaptation to climate in *Medicago truncatula*. *Genetics*, 196(4), 1263–1275. <https://doi.org/10.1534/genetics.113.159319>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Selmoni O, Vajana E, Guillaume A, Rochat E, Joost S. Sampling strategy optimization to increase statistical power in landscape genomics: A simulation-based approach. *Mol Ecol Resour*. 2020;20:154–169. <https://doi.org/10.1111/1755-0998.13095>