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

Running title: Sampling strategy in landscape genomics

Oliver Selmoni[1], Elia Vajana[1], Stéphane Joost[1]

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

Abstract

An increasing number of studies use landscape genomics to investigate local adaptation in wild and domestic populations. This method consists of simultaneously characterizing genetic and environmental variability across a landscape in order to depict how genotype frequencies are shaped by adaptive forces. The implementation of this approach requires the sampling phase to consider the complexity of environmental settings and the burden of logistic constraints. These aspects can have a major impact on reproducibility and costs but they are often underestimated in the literature dedicated to sampling strategies.

In this study, we compute simulated genomic datasets on a real environmental setting and distinct demographic scenarios to reiterate landscape genomics experiments with distinct sampling strategies. The latter differ by design approaches (aiming at enhancing environmental and/or geographic representativeness), number of sampling locations and sample sizes. We then propose an analytical framework to evaluate the effect of these elements on statistical power and false discovery rates of the analysis.

Our results highlight the importance of sample size in designing landscape genomic studies, where a minimal threshold of 200 individuals is required to detect at least part of the true positives. Furthermore, our results show that sampling at less than 10 sites clearly impairs statistical power, but also that the optimal number of sampling locations is not necessarily the highest possible. Last, we show that the designs that account for environmental representativeness result in the highest statistical power. Overall, this study provides valuable guidelines for optimizing sampling strategies in landscape genomics experiments.

Keywords

Sampling strategy – Landscape genomics – statistical power – false discovery rate – Sample size

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 analyzing 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 be distinguished from adaptive variation (Rellstab et al., 2015). Over the last few years, the landscape genomic approach is becoming more widely used (see Tab. 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, & Treml, 2016a; 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 high number of loci (Balkenhol et al., 2017; Rellstab et al., 2015). Beyond these theoretical aspects, pragmatic choices need to be made with regards to financial and logistic constraints that are often imposed (Manel et al., 2010; Rellstab et al., 2015). A sampling strategy is constituted of: i) sampling design (the spatial arrangement of the sampling locations, D); ii) the number of sampling locations (L); and iii) sampling size (the number of individuals sampled, N; Tab. 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 et al., 2010; Manel et al., 2012; Rellstab et al., 2015; Riginos et al., 2016). Theoretical simulations have shown that performing transects along environmental gradients or sampling pairs of contrasting sites which are spatially close reduced false discovery rates 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 analyzed 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 to control for the bias associated with collinear conditions (Rellstab et al., 2015). Furthermore, these studies 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 1800; Lotterhos & Whitlock, 2015) appear to be unrealistic for use in most of real landscape genomic experiments (Tab.1) and thus the guidelines proposed are scarcely applicable in practice.

For these reasons, there is 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 money expenses (Balkenhol et al., 2017; Rellstab et al., 2015). In particular, the fundamental questions that need to be addressed are: i) how to determine the spatial arrangement of sampling locations; ii) how to organize sampling effort (for instance preferring many samples at few sites, or rather fewer samples at many sites); and iii) how many samples are required to obtain sufficient statistical power (Rellstab et al., 2015; Riginos et al., 2016).

In this paper, we investigate how the outcome of landscape genomic analyses is driven by the sampling strategy. We ran simulations using a fictive genetic dataset 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 differ according to the sampling design approach (D), the number of sampling locations (L) and number of samples (N). A full factorial analysis of variance was applied to evaluate the effect of each factor on statistical power and false discovery rate (FDR) produced by the analysis.

Material & Methods

The iterative approach we designed for testing sampling strategies required that a new genetic dataset encompassing neutral and adaptive variation was created at every run of the simulations. A simulated genomic dataset 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 (for ex. SPLATCHE2; Ray, Currat, Foll, & Excoffier, 2010) did not match our needs as they cannot compute complex selective scenarios (for instance those involving multiple environmental variables; Carvajal-Rodríguez, 2008). We could not use forward-in-time methods (for ex. CDPOP; Landguth & Cushman, 2010) neither, as they are slow and therefore not compatible with the computational requirements of our simulative approach (Carvajal-Rodríguez, 2008). For these reasons, we developed a customized framework in the R environment (version 3.3.1; R Core Team, 2016) to compute neutral and adaptive genetic variation from gradients of population membership and environmental variations, respectively (Fig. 1).

Each iteration starts with the establishment of a genetic setup defined by the presence or absence of a population structure (Fig. 2, step 1) and by a given number of adaptive loci (Fig. 2, step 2). Subsequently, a sampling strategy is applied as a combination of sampling design (D), number of sampling locations (L) and sample size (N) (Fig.2, steps 3, 4 and 5), resulting in the generation of a genetic dataset that, coupled with environmental data, undergoes a landscape genomics analysis (Fig.2, step 6). Then the performance of each sampling strategy is evaluated by means of a factorial analysis of variance to investigate how the true positive rate (TPR, *i.e.* statistical power) and the false discovery rate (FDR) are influenced by the elements defining the sampling strategy (D, L, N; Fig. 2, step 7). These elements and the parameters defining the genetic setup varied at every iteration (Tab. 2). This approach was applied to a fictive case study mimicking a wild species living across the environmental settings of Europe. All scripts and data used to perform this analysis are publicly available on Dryad (doi:10.5061/dryad.28sf38g).

Environmental data

As a base for our simulations, we quantified the environmental settings of Europe (Fig. S1). We retrieved eight climatic variables from publicly available sources (e.g. mean annual temperature, average precipitation, and altitude; Tab 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 50x50 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.

Population structure

A population structure was artificially computed (Fig S1). Gradient of population membership was created using an image editing software (Microsoft PowerPoint). A greyscale gradient (*i.e.* pixel values ranging between 0 and 255) was manually drawn on a rectangular area so that extreme membership values corresponded to black (value=0) and white (value=255) pixels, while the rest of the pixels represented admixed membership coefficients (Fig. S1). A unique gradient spanning the length of the region was drawn so that one population was placed at one extreme (north-west of Europe) and a second population on the opposite end (south-east of Europe; Fig. S1). In between, individuals were progressively admixed to these two populations. The resulting image - whose pixel size can be any - was then imported in QGIS using the GDAL Georeferencer tool (Open Source Geospatial Foundation, 2018) and its extent was fitted to the one of the study area. Values from the imported layer were then assigned to each cell of the landscape grids using the mean function of the QGIS zonal statistics tool.

Computation of genotypes

For the creation of the genotype matrices, we developed an R-pipeline based on probability functions to compute genotypes out of population membership coefficient and environmental values (Box S1). The theoretical fundaments of this method stand on the observation that when the population is structured, neutral alleles tend to show similar spatial patterns of distribution (a feature commonly exploited in Fst outlier tests; Luikart et al., 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). Based on these assumptions, we developed two distinct evolutionary scenarios, one implying a population which is not genetically structured (hereafter referred to as the “panmictic population scenario”), and one implying a structured population (hereafter referred to as the “structured population scenario”; see Box S2).

For every iteration, 1’000 loci were computed: five to 20 were set to “adaptive”, while the remaining were set to “neutral”. They were computed as follows:

- Neutral markers (Box S1A): a linear transformation converts the coefficient of population structure (Fig. S1) at each sampling site into a probability of carrying a private allele of the population ($pA|PS$). A parameter (c, Box. S2) defines this transformation and returns a constant $pA|PS$ equal to 0.5 under the panmictic population scenario (c=0.5), while this probability varies together with the coefficient of population structure (c=0.2) under the structured scenario.

147 These probabilities are then used to draw (using the R-stat *sample* function) the bi-allelic genotype of each individual
148 and this procedure is re-iterated for each neutral locus.

149 - Adaptive markers (Box S1B): in addition to *pA*|*PS*, an additional probability (*pA*|*Env*) is calculated with the same
150 method but using a specific environmental variable instead of the coefficient of population membership. In this case,
151 the parameter defining the relationship (*s,* Box. S2) determines the strength of the selective force and is re-drawn for
152 every adaptive genotype from a uniform distribution ranging from 0.1 to 0.3 (corresponding to strong and weak
153 selection, respectively). Each adaptive locus is randomly associated with one environmental variable. This implies that
154 some environmental conditions can be associated with several genetic markers, while others can be associated with
155 none. For every adaptive locus, the bi-allelic genotype is drawn (using the R-stat *sample* function) out of a combined
156 probability (*pA*) resulting from the weighted mean of *pA*|*PS* and *pA*|*ENV*. The parameter (γ) defining the weight
157 assigned to *pA*|*PS* (in %) is drawn from a uniform distribution ranging from 0 to 0.5 (corresponding to low and high
158 weights, respectively; Box S2).

159

160 *Sampling design*

161 Four types of sampling designs are proposed: three of them differently account for the characteristics of the landscape while
162 one randomly selects the sampling locations. The first is "geographic" (Fig. 2A) and is defined through a hierarchical
163 classification of the sites based on their geographic coordinates. The desired number of sampling locations (L) determines the
164 number of clusters and the geographical center of each cluster is set as a sampling location. The goal of this strategy is to
165 sample sites located as far as possible from each other in the geographical space to guarantee spatial representativeness.

166 The second design type is "environmental" (Fig. 2B). It is based on the computation of distances depending on the values of
167 environmental variables. The latter are first processed by a correlation filter: when two variables are found correlated to each
168 other (R>±0.5), one of them (randomly chosen) is excluded from the dataset. The remaining un-correlated descriptors are
169 scaled (sd=1) and centered (mean=0) using R *scale* function. The scaled values are used to perform a hierarchical clustering
170 between the landscape sites. Like the previous design, the desired number of sampling locations (L) defines the number of
171 clusters. For each cluster, the environmental center is defined by an array containing the mean of the scaled environmental
172 values. Then, the Euclidean distances between this array and the scaled values of each site of the cluster are computed. On this
173 basis, the most similar sites to each center are selected as sampling locations. This strategy aims to maximize environmental
174 contrast between sampling locations and thus favors the detection of adaptive signals (Manel et al., 2012; Riginos et al., 2016).

175 The third design is "hybrid" (Fig. 2C) and is a combination of the first two. It consists in dividing the landscape in *k*
176 environmental regions and selecting within each of these regions two or more sampling locations based on geographic position.
177 Initially, the environmental variables are processed as in the environmental design (correlation-filter and scaling) and used for
178 the hierarchical classification of the landscape sites. The next step is separating the landscape sites in k environmental regions
179 based on this classification. The allowed value of k ranges between 2 and half of the desired number of sampling locations (L).
180 We use the R package NbClust (version 3.0, Charrad, Ghazzali, Boiteau, & Niknafs, 2015) to find the optimal value of k within
181 this range. The optimal k is then used to determine the k environmental regions. Next, the number of sampling locations (L) is
182 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
183 environmental regions. The number of sampling locations per environment region (L_{ki}) can therefore be equal among
184 environmental regions or, at worst, differ by one (for ex. 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,
185 L_{k3}=2, L_{k4}=2). Sampling locations within environmental regions are chosen based on geographical position. Geographical clusters
186 within each environmental region are formed as in the geographic design, setting L_{ki} as the number of clusters. The landscape
187 site spatially closer to the center of each geographical cluster is selected as sampling location. In such a way, the procedure
188 allows the replication of similar environmental conditions at distant sites, being therefore expected to disentangle neutral and
189 adaptive genetic variation and to promote the detection of variants under selection (Manel et al., 2012; Rellstab et al., 2015;
190 Riginos et al., 2016).
191 The fourth type of design is "random": the sampling locations (L) are randomly selected across all the available landscape sites.
192 In our simulations, we tested each type of sampling design with numbers comparable to the ones used in real experiments (see
193 Tab. 1). We used 5 levels of sampling locations L (5, 10, 20, 40 and 50 locations) and 6 of sample sizes N (50, 100, 200, 400, 800
194 and 1600 individuals). In iterations for which the sampling size is not an exact multiple of the number of sites (for ex., 20 sites
195 and 50 individuals), the total number of individuals was changed to the closest multiple (here 40 individuals). The scripts
196 including these procedures were written in R using the functions embedded within the *stats* package (R Core Team, 2016).
197
198 *Landscape genomics analysis*
199 We computed association models for each iteration with the SamBada software (version 0.6.0; Stucki et al., 2017). First, the
200 simulated matrix of genotypes is filtered through a customized R function with minor allele frequency <0.05 and major
201 genotype frequency >0.95 to avoid including rare or monomorphic alleles and genotypes, respectively. Secondly, a principal
202 component analysis (PCA) is run on the filtered genotype matrix to obtain synthetic variables accounting for population

203 structure (hereafter referred to as population structure variables; Patterson, Price, & Reich, 2006). The analysis of the
204 eigenvalues of the PCA is carried out in order to assess whether the population structure is negligible for downstream analysis
205 (Patterson et al., 2006). At each iteration, the algorithm checks the difference between the first two eigenvalues: if the second
206 one is at least 80% of the first one, population structure is considered absent. Even though rudimentary, this method is
207 expected to be sufficient to detect the straightforward neutral genetic backgrounds produced by the simulations. It is
208 compatible with the computational speed required in such an iterative framework, compared with a statistical testing approach
209 (Patterson, Price, & Reich, 2006).

210 After filtering, SamBada is used to detect candidate loci for local adaptation. The software is able to run multivariate logistic
211 regression models (Joost et al., 2007) that include population structure as a co-variable, while guaranteeing fast computations
212 (Rellstab et al., 2015; Stucki et al., 2017). To ensure compatibility with our pipeline and increase computational speed, we
213 integrated the SamBada method in a customized python script (version 3.5; van Rossum, 1995) based on the Pandas
214 (McKinney, 2010), Statsmodels (Seabold & Perktold, 2010) and Multiprocessing (Mckerns, Strand, Sullivan, Fang, & Aivazis,
215 2011) packages. *P*-values related to the two statistics (G-score and Wald-score) associated with each association model are
216 computed and subsequently corrected for multiple testing using the R *q-value* package (version 2.6; Storey, 2003). Models are
217 deemed significant when showing a $q<0.05$ for both tests. When multiple models turn out to be significant for the same
218 marker, only the best one is kept (according to the G-score). The pipeline was developed in the R-environment using the *stats*
219 library.

220

221 *Simulations and evaluation of the performance*

222 Each combination of demographic scenarios, sampling designs, number of sampling locations and sample sizes was replicated
223 20 times for a total number of 4'800 simulations (Tab. 2). A new genetic matrix is randomly redrawn for each simulation to
224 change the selective forces implying local adaptation and re-draw the neutral loci. At the end of each iteration, two diagnostic
225 parameters are returned:

226 - True Positive Rate (TPR): percentage of true associations detected to be significant;
227 - False Discovery Rate (FDR): percentage of false association among the significant ones.

228 After the simulations, a factorial analysis of variance (ANOVA) was performed using the *anova* function implemented in the R
229 stats package to investigate how TPR and FDR are influenced by the three elements determining the sampling strategy (*i.e.*
230 sampling design, number of sampling locations and sampling size; Tab. 2). The observed values of a diagnostic parameter across

231 all iterations (TPR or FDR) were set as the dependent variable and the three factors determining the sampling strategy as the
232 independent ones. For each factor (main effect) and for each combination of factors (interaction effect), the ANOVA tested
233 whether the variance of the dependent variable was due to differences between the levels of the factor (or the combination of
234 factors in case of interactions) or not. Furthermore, we applied the one-way ANOVA to test whether the two demographic
235 scenarios significantly affected TPR and FDR. Effects were deemed to be significant when the associated p-value was < 0.001.
236
237 **Results**
238 *ANOVA of True Positive Rate (TPR)*
239 The simulations involving the panmictic demographic scenario showed higher statistical power, compared to those ran under
240 the structured scenario ($p<10^{-13}$ under ANOVA; TPR_{PAN}=0.36±0.37, TPR_{STR}= 0.27±0.33; Fig. 3A, 3C). In both scenarios, the three
241 elements defining the sampling strategy (sampling design: D, sample size: N, number of sampling locations: L) resulted as
242 significantly driving TPR as main effects ($p_{PAN}<10^{-15}$, $p_{STR}<10^{-15}$; Tab 3a).
243 Sample size appeared as a main driving factor of statistical power ($p_{PAN}<10^{-15}$, $p_{STR}<10^{-15}$; Tab 3a). Small sample sizes (N=50, 100)
244 resulted in TPRs close or equal to zero ($TPR_{PAN-N50}$=0±0, $TPR_{STR-N50}$=0±0, $TPR_{PAN-N100}$= 0.01±0.03, $TPR_{STR-N100}$=0±0.01; Fig. 3A, 3C).
245 In contrast, increases of sample size starting from N=200 showed progressive augmentation of statistical power (TPR_{PAN-}
246 $_{N200}$=0.15±0.13, $TPR_{PAN-N400}$= 0.43±0.22, $TPR_{PAN-N800}$=0.69±0.22, $TPR_{PAN-N1600}$= 0.87±0.18; Fig. 3A; $TPR_{STR-N200}$=0.06±0.08 TPR_{STR-}
247 $_{N400}$= 0.28±0.19, $TPR_{STR-N800}$=0.57±0.23, $TPR_{STR—N1600}$=0.77±0.22; 3C).
248 A second important driver of a high TPR was the number of sampling locations ($p_{PAN}<10^{-15}$, $p_{STR}<10^{-15}$; Tab 3a). The main
249 contrasts differentiated simulations sampling at 5 locations ($TPR_{PAN—L5}$=0.26±0.29, $TPR_{STR—L5}$=0.16±0.22) from those sampling at
250 10 or more ($TPR_{PAN—L10}$=0.35±0.36, $TPR_{PAN—L20}$=0.39±0.39, $TPR_{PAN—L40}$=0.39±0.39, $TPR_{PAN—L50}$=0.39±0.39, Fig. 4A;
251 $TPR_{STR-L10}$=0.28±0.32, $TPR_{STR—L20}$=0.31±0.35, $TPR_{STR—L40}$=0.32±0.36, $TPR_{STR—L50}$=0.33±0.37, Fig. 4C).
252 The sampling design approach also resulted as having an important effect on statistical power ($p_{PAN}<10^{-15}$, $p_{STR}<10^{-15}$; Tab 3a).
253 Under both demographic scenarios, statistical power under the geographic sampling design was comparable to the one of a
254 random approach ($TPR_{PAN—GEO}$=0.29±0.33, $TPR_{PAN—RAN}$=0.29±0.33, Fig. 3A; $TPR_{STR—GEO}$=0.25±0.3, $TPR_{STR—RAN}$=0.23±0.30, Fig. 3A).
255 In contrast, environmental and hybrid strategies showed higher values of TPR ($TPR_{PAN—ENV}$=0.41±0.39, $TPR_{PAN—HYB}$=0.43±0.4, Fig.
256 3A; $TPR_{STR—ENV}$=0.30±0.35, $TPR_{STR—HYB}$=0.33±0.37, Fig. 3C).
257 The interaction effects of the three elements defining the sampling strategy (D, L, N) resulted to drive TPR under both scenarios
258 ($p_{PAN}<10^{-5}$, $p_{STR}<10^{-13}$; Tab 3a). None of the interactions between two elements reversed the patterns observed in main effects

259 of the elements alone (Fig. 3A, 3C, Fig. S2A, S2C, Fig. S3A, S3C). To the contrary, interactions appeared to accentuate the
260 patterns observed as main effects (Fig. 3A, 3C, Fig. S2A, S2C, Fig. S3A, S3C).
261
262 *ANOVA of False Discoveries (FDR)*
263 The analysis of variance of false discoveries shows that simulations under the structured demographic scenario resulted in
264 higher FDR ($p<10^{-15}$; FDR_{STR}= 0.83±0.30), compared to those under the panmictic population (FDR_{PAN}=0.5±0.41; Fig. 3B, 3D). As
265 main effects are concerned, sample size and number of sampling locations resulted as significantly driving false discoveries
266 under both scenarios ($p_{PAN}<10^{-15}$, $p_{STR}<10^{-12}$; Tab 3b).
267 Sample sizes resulted as an important driver of false discoveries ($p_{PAN}<10^{-15}$, $p_{STR}<10^{-15}$; Tab 3b). When sample sizes were small
268 (N=50, 100), FDR was systematically close to its maximal value ($FDR_{PAN-N50}$=1±0, $FDR_{STR-N50}$=1±0, $FDR_{PAN-N100}$= 0.92±0.27, TPR_{STR-}
269 $_{N100}$=0.99±0.1; Fig. 3B, 3D). For higher sample sizes, we observed opposite effects on FDR between the two demographic
270 scenarios. In the panmictic scenario, FDR decreased or remained constant along with increments of sample size
271 ($FDR_{PAN-N200}$=0.37±0.42, $FDR_{PAN-N400}$= 0.22±0.23, $FDR_{PAN-N800}$= 0.23±0.16, $FDR_{PAN-N1600}$= 0.26±0.14; Fig. 3B) while in the structured
272 population scenario the augmentation of number of samples raised false discoveries ($FDR_{STR-N200}$=0.65±0.43, FDR_{STR-}
273 $_{N400}$=0.61±0.36, $FDR_{STR-N800}$=0.81±0.24, $FDR_{STR-N1600}$=0.92±0.15; Fig. 3D).
274 The number of sampling locations resulted as an important factor in determining the rates of false discoveries ($p_{PAN}<10^{-15}$,
275 $p_{STR}<10^{-12}$; Tab 3b). Under the panmictic demographic scenario, increasing the number of sampling locations reduced FDR
276 (FDR_{PAN-L5}= 0.6±0.36, $FDR_{PAN-L10}$=0.49±0.4, $FDR_{PAN-L20}$=0.47±0.42, $FDR_{PAN-L40}$=0.46±0.42, $FDR_{PAN-L50}$=0.47±0.42; Fig. 4B). A similar
277 effect was observed under the structured demographic scenario, even though FDR were comparable for medium-high numbers
278 of sampling location (L≥20; FDR_{STR-L5}= 0.89±0.23, $FDR_{STR-L10}$=0.85±0.28, $FDR_{STR-L20}$=0.81±0.32, $FDR_{STR-L40}$=0.82±0.32,
279 $FDR_{STR-L50}$=0.79±0.33; Fig. 4B).
280 Sampling design is significantly driving FDR only under the structured demographic scenario (p_{PAN}=0.010, $p_{STR}<10^{-15}$; Tab 3b).
281 Under this demographic set-up, environmental design resulted in FDRs comparable to those of a random design
282 ($FDR_{STR-ENV}$=0.77±0.33, $FDR_{STR-RAN}$=0.75±0.35) whereas the hybrid and geographic approaches resulted in higher rates
283 ($FDR_{STR-HYB}$=0.9±0.24, $FDR_{STR-GEO}$=0.9±0.24). Under the panmictic scenario, FDRs were comparable regardless of the design
284 approach ($FDR_{PAN-ENV}$=0.47±0.4, $FDR_{PAN-RAN}$=0.51±0.43, $FDR_{PAN-HYB}$=0.5±0.38, $FDR_{PAN-GEO}$=0.51±0.43).

285 The interaction effects between sampling design and sample size (DxN) and the one between sample size and number of
286 sampling locations (NxL) resulted as significantly determining FDR under any demographic scenario ($p_{PAN}<10^{-9}$, $p_{STR}<10^{-15}$; Tab
287 3b). In contrast, the interaction between design and number of sampling locations (DxL) determined false discoveries only
288 under the structured scenario ($p_{PAN}=0.010$, $p_{STR}<10^{-6}$; Tab 3b). As in the case of TPR, the mains effects on FDR of single elements
289 of sampling strategy (D, L, N) appeared be accentuated and not reversed, when measuring their effects in interactions (Fig. 3B,
290 3D, Fig. S2B, S2D, Fig. S3B, S3D).
291
292 **Discussion**
293 Our simulations highlight the fact that sampling strategy clearly drives the outcome of a landscape genomics experiment,
294 especially when the studied population shows a structure in its neutral genetic variation. Despite some limitations that we
295 discuss below, the results obtained make it possible to answer three questions researchers are confronted to when planning
296 this type of research investigation.
297
298 *How many samples are required to detect any adaptive signal?*
299 As per the findings of previous research (e.g. Lotterhos & Whitlock, 2015), our results suggest that sample size is the key factor
300 in securing the best possible outcome for a landscape genomics analysis. Where statistical power is concerned, there is an
301 unquestionable advantage in increasing the number of samples under the scenarios tested. We found a distinct lack of
302 statistical power in simulations where N=50 or 100 (even in absence of confounding effects due to demographic processes),
303 while detection of true positives increased significantly for N≥200 (Fig. 3A, 3C). As we progressively doubled sample size
304 (N=200, 400, 800, 1600), we observed that the statistical power increased following a logarithmic trend under both
305 demographic scenarios (Fig. 3A, 3C). The plateau of this incremental (TPR=1) was consistently reached in simulations under the
306 panmictic scenarios sampling 1600 individuals (Fig. 3A). When the population is structured, more samples are probably
307 required to reach maximal statistical power (Fig. 3C).
308 As we consider sample sizes above 200 individuals, we observe that FDR progressively raises along with the number of
309 individuals (Fig. 3B, 3D). This is likely to be a trade-off of high sample sizes: the increased detection of true positive is
310 accompanied by the (mis-)detection of false positives. As observed in previous studies, this effect is accentuated when the
311 demographic set-up is complex (De Mita et al., 2013; Rellstab et al., 2015), while it is marginal when population structure is
312 absent (Fig. 3B, Tab. 3B). The drawback of false discoveries is generally accepted because landscape genomics experiments

313 constitute the first step toward the identification of adaptive loci, which is in general followed by further experimental
314 validations (Pardo-Diaz, Salazar, & Jiggins, 2015).

315 Each landscape genomic experiment is unique in terms of environmental and demographic scenarios and this is the reason why
316 it is not possible to propose a comprehensive mathematical formula to predict expected TPR and FDR based only on the
317 sampling size. Nonetheless, we advise against setting up experiments with fewer than 200 samples as the statistical
318 requirements to detect true signals will unlikely be met.

319

320 *How many sampling sites?*

321 A large number of samples inevitably raises the cost of an experiment, largely due to sequencing and genotyping (Manel et al.,
322 2010; Rellstab et al., 2015). Additionally, field work rapidly increases the cost of a study in cases where sampling has to be
323 carried out across landscapes with logistic difficulties and physical obstacles. Therefore, it is both convenient and economical to
324 optimize the number of sampling locations to control for ancillary costs. De Mita et al. (2013) suggested that increasing the
325 number of sampling locations would raise power and reduce false discoveries. Our findings only partially support this view. A
326 small number of sampling locations (L=5) was found reducing TPR and increasing FDR, compared with higher numbers (L=10,
327 20, 40 and 50; Fig. 4). However, high numbers of sampling locations (L=40 and 50) provided little benefits in terms of FDR and
328 almost no benefits in terms of TPR, compared to moderate numbers (L=10 and 20; Fig. 4). These discrepancies with previous
329 studies are probably due to differences in the respective simulative approaches applied (we use several environmental
330 descriptors instead of one) and the characteristics of the statistical method we employed to detect signatures of selection. In
331 fact, as a number of sampling locations is sufficient to portray the environmental contrasts of the study area, adding more
332 locations does not bring additional information and therefore does not increase statistical power. To the contrary, logistic
333 regression is known to be sensitive to data sparsity (i.e. many sampling sites with few samples per sites; Greenland,
334 Mansournia, & Altman, 2016), but such negative effects were not observed for the highest number of sampling locations we
335 tested (L=50, Fig. 4).

336 The results we produced tend to show that there is no remarkable benefit to collect a small number of samples per site at a
337 large number of locations, compared to the opposite. It seems inappropriate to include more than 20 sampling sites if no
338 specific reason justifies it (e.g. important environmental differences, complex population structure). It is also important to
339 mention that, as we observed, an excessive reduction in the number of sampling locations reduces statistical power and raises
340 false discoveries. When sampling at a small number of locations, the environmental characterization of samples is likely to

341 neglect some contrasts and ignore confounding effects between collinear variables (Leempoel et al., 2017; Manel et al., 2010).
342 The implications of the information described above are considerable since the cost of field work can be drastically reduced
343 with marginal countereffects on statistical power and false discoveries.
344
345 *Where to sample?*
346 Compared with random or opportunistic approaches, sampling designs based on the characteristics of the study area are
347 expected to improve the power of landscape genomics analysis (Lotterhos & Whitlock, 2015). We developed three distinct
348 methods to choose sampling locations accounting for geographical and/or environmental information (geographic,
349 environmental and hybrid designs). We confronted these design approaches between them and with random sampling
350 schemes. The approach based on geographic position (geographic design) resulted in statistical power similar to random
351 designs (Fig. 3A, 3C), while the one based on climatic data (environmental design) displayed remarkably higher TPRs (Fig. 3A,
352 3C). In comparison to the latter, the sampling design accounting for both environmental information and geographic position
353 (hybrid design) resulted in a statistical power generally comparable or higher (Fig. 3A, 3C). These beneficial effects on TPR were
354 accentuated under the structured demographic scenario, when moderate sample sizes (N=200, 400) were used (Fig. 3C). These
355 results match previous observations: methods conceived to take advantage of environmental contrasts facilitate the detection
356 of adaptive signals (Manel et al., 2012; Riginos et al., 2016). Furthermore, the hybrid design prevents the sampling of
357 neighboring sites with similar conditions, therefore avoiding the superposition between adaptive and neutral genetic variation
358 (Manel et al., 2012). This is likely to explain why the hybrid design outperformed the environmental approach (Fig. 3).
359 As false discoveries are concerned, the three design approaches we developed (geographic, environmental and hybrid) did not
360 display any particular benefit. When population structure was absent, FDRs under these designs resulted as comparable to
361 those observed in random sampling schemes (Fig. 3B). In simulations using the structured population scenario (Fig. 3D), the
362 hybrid and geographic designs resulted in even higher rates of false discoveries, compared to random designs. The
363 environmental design is therefore the only one providing FDRs comparable to random sampling schemes, regardless of the
364 demographic scenario. The high FDR associated to the hybrid design is likely to be a trade-off of its higher statistical power, as
365 similarly observed for high sample sizes. In the environmental design, however, this trade-off appeared as less pronounced. As
366 mentioned earlier, high rates of false discoveries are usually tolerated because of the exploratory nature of landscape genomics
367 studies (Pardo-Diaz et al., 2015). For the reasons mentioned in this section, we strongly encourage the application of a design

368 approach that emphasizes environmental contrast (environmental or hybrid), as these two configurations provide the highest
369 statistical power.
370
371 *Limitations*
372 This paper addresses fundamental questions to elaborate an efficient sampling strategy. The simulative approach developed
373 and the results produced are constrained by several limitations related to the computation of genotypes. The main one
374 concerns the neutral genetic variation generated, which is calculated out of textbook demographic scenarios; it therefore
375 represents a simplified view of reality (Riginos et al., 2016). The use of *forward-in-time* simulations that mimic the demographic
376 history of a population (used by De Mita et al., 2013; Lotterhos & Whitlock, 2015) would probably have resulted in more
377 realistic scenarios. In order to be used in a framework as the one we proposed, the *forward-in-time* methods should be
378 compatible with a large number of spatial locations (*i.e.* potential sampling sites) and a genetic dataset counting at least one
379 thousand loci, of which up to 20 as adaptive against a dozen of environmental variables. Importantly, all these requirements
380 should be fulfilled at a reasonable computational speed. To our knowledge, there are no existing software meeting these
381 criteria.
382
383 *Conclusions*
384 The present work provides guidelines for optimizing the sampling strategy in the context of landscape genomic experiments.
385 Our simulations highlight the importance of working with a minimal sample size of 200 individuals to achieve sufficient power in
386 landscape genomic analyses. The costs induced by a large number of samples can be balanced by reducing those related to field
387 work. As long as a moderate number of sampling locations (10 or 20 sites) is sufficient to portray the environmental contrasts
388 of the study area, there is only little statistical benefit at sampling at higher number of sites (40 or 50). Furthermore, we
389 describe two approaches for selecting the sampling locations accounting for environmental characteristics and show their
390 benefic effects on the detection of true positives.
391

392 **Acknowledgements**

We thank Annie Guillaume for the useful comments and suggestions provided during the redaction of this paper. We acknowledge funding from the IMAGE (Innovative Management of Animal Genetic Resources) project funded under the European Union's Horizon 2020 research and innovation program under grant agreement No 677353.

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 (pp. 1–62). Springer, Cham. https://doi.org/10.1007/13836_2017_2

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

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

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

Greenland, S., Mansournia, M. A., & Altman, D. G. (2016). Sparse data bias: a problem hiding in plain sight. *BMJ*, i1981. https://doi.org/10.1136/bmj.i1981

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

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

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 (pp. 3–12). Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-61779-870-2_1

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, 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

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

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

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

Open Source Geospatial Foundation. (2018). Geospatial Data Abstraction software Library. Retrieved from https://github.com/OSGeo

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 FST 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 Refrence, 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., & Treml, E. A. (2016). Navigating the currents of seascape genomics: how spatial analyses can augment population genomic studies. *Current Zoology*, *62*, doi: 10.1093/cz/zow067. https://doi.org/10.1093/cz/zow067

Ryan, W. B. F., Carbotte, S. M., Coplan, J. O., O'Hara, S., Melkonian, A., Arko, R., … Zemsky, 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/

Storey, J. D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *Annals of Statistics*, *31*(6), 2013–2035.

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., … Consortium, the N. (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., Milanesi, 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

Wenzel, M. A., Douglas, A., James, M. C., Redpath, S. M., & Piertney, 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

Data Accessibility Statement

All scripts and data used to perform this analysis are publicly available on Dryad (doi:10.5061/dryad.28sf38g).

Author Contributions

OS and SJ designed research, OS performed research, OS, EV and SJ analyzed the results and wrote the paper. All the authors undertook revisions, contributed intellectually to the development of this manuscript and approved the final manuscript.

553 ## Tables and Figures

554 **Table 1. Sampling design in landscape genomics studies.** A non-exhaustive list of landscape genomics studies, highlighting
555 different species and their related sampling strategies.

556 * Numbers from the Vincent et al. report (2013) concerning the *non-pooled* samples.

Study	Species	Sampling Design (D)	Sampling Locations (L)	Sample Size (S)
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, 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	Eurpean 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	Eurpeal Eel	Spatial and environmental representativeness	8 sites	179
Vincent et al., 2013	Atlantic Salmon	Spatial representativeness	26* rivers	641*

557

558

559

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) and the number of adaptive markers (5 to 20). The environmental variable of interest and the strength of selection is randomly sampled for every adaptive marker. Following this, the sampling strategy (here shown with red boxes) is set as a combination of design approach (geographic, environmental, hybrid or random), number of sampling locations (5, 10, 20, 40 or 50 locations) and sampling size (50, 100, 200, 400, 800 or 1600 samples). This results in the creation of a genotype matrix that undergoes a landscape genomics analysis. At the end of iterations, statistical power and false discovery rate are calculated to assess the performance of the sampling strategy.

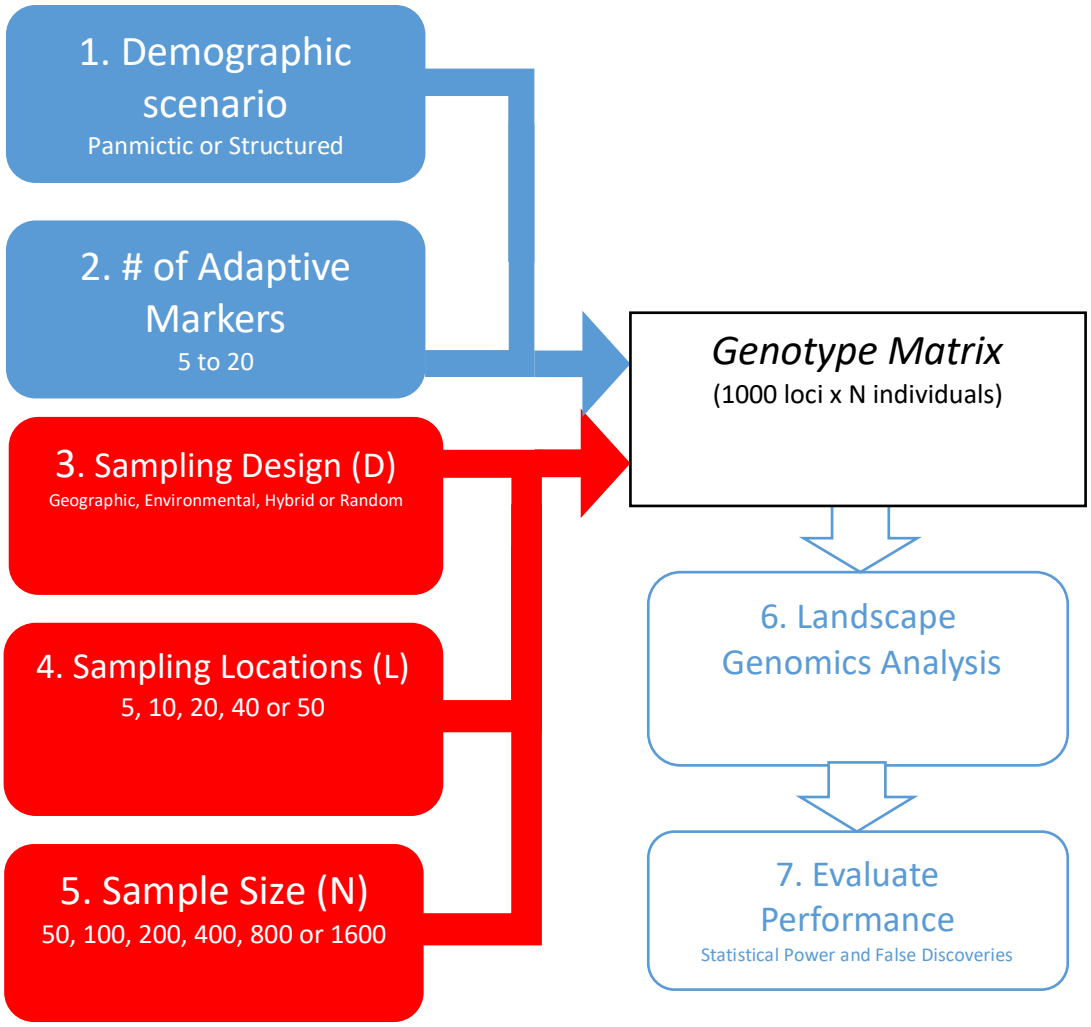

Figure 2. The three sampling design approaches accounting for landscape characteristics. The three maps illustrate how the eight sampling sites are chosen under three different sampling designs. Under a geographic strategy (A), sample location is selected using only geographic coordinates in order to 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.

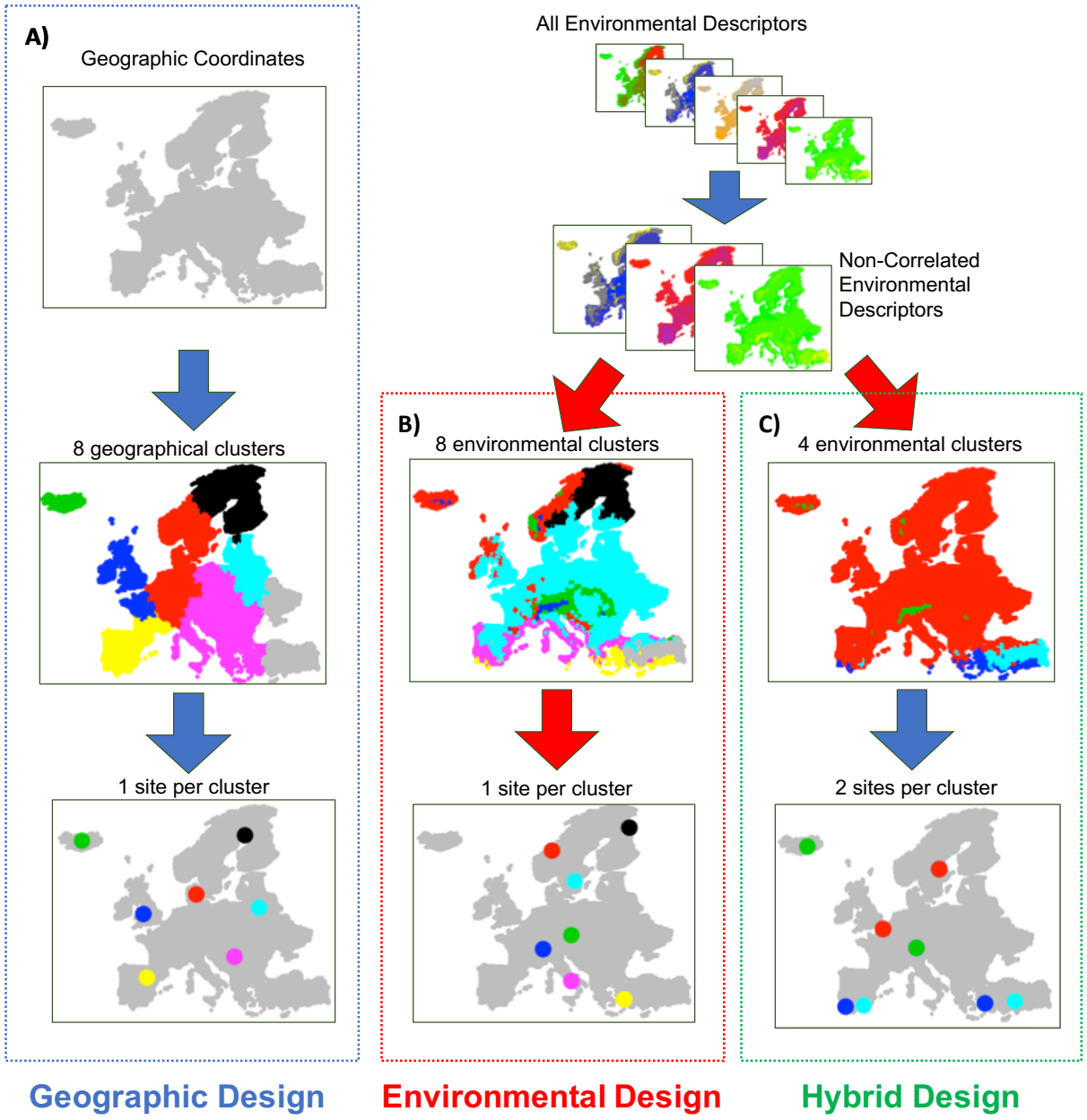

Table 2. Table of factors varying in the simulative approach. Two different demographic scenarios are possible, one in which there is no neutral genetic structure (panmictic population) and one in which there is a structured variation (structured population). We then used sampling strategies emulating those observed in real experiments. Three different sampling designs approaches accounting for landscape characteristics are proposed: one maximizing the spatial representativeness of samples (geographic), 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 employed were comparable to those from real experiment: 5 levels of number of sampling locations spanned from 5 to 50 sites, and 6 levels of sampling sizes (i.e. total number of samples) from 50 to 1600 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 simulation were computed.

Factor	**# levels**	**Levels**
Demographic Scenarios	2	Panmictic Population, Structured Population
Sampling Design (*D*)	4	Geographic, Environmental, Hybrid, Random
Sampling Locations (*L*)	5	5, 10, 20, 40, 50
Sampling Size (*N*)	6	50, 100, 200, 400, 800, 1600
Replicates	20	
Total	4800	

Table 3. Results of the Analysis of Variance (ANOVA). The ANOVA results for the two statistical rates a) Statistical Power (TPR) and b) False Discoveries (FDR) that measure the performances of the landscape genomics analyses at each iteration are shown. For both demographic scenarios (Panmictic or Structured), we show the p-values associated with the ANOVA for the rate of interest based on the three elements determining the sampling strategy: sampling design (D), sample size (N), number of sampling locations (L) and their interactions.

	a) True Positives Rate: TPR		**b) False Discovery Rate: FDR**	
	Panmictic	*Structured*	*Panmictic*	*Structured*
D	$<10^{-15}$	$<10^{-15}$	0.010	$<10^{-15}$
N	$<10^{-15}$	$<10^{-15}$	$<10^{-15}$	$<10^{-15}$
L	$<10^{-15}$	$<10^{-15}$	$<10^{-15}$	$<10^{-12}$
D x N	$<10^{-15}$	$<10^{-15}$	$<10^{-9}$	$<10^{-15}$
D x L	$<10^{-8}$	$<10^{-15}$	0.010	$<10^{-6}$
N x L	$<10^{-15}$	$<10^{-15}$	$<10^{-14}$	$<10^{-15}$

Figure 3. Performances of landscape genomics analysis under different sample size and sampling design approaches. The four plots display how statistical power (A, C) and false discoveries (B, D) of the landscape genomics analyses vary under different levels of sample size and sampling design approaches. The two distinct demographic scenarios under which analyses where run are shown (panmictic: A and B, structured: C and D). The white boxplots represent the distribution of the TPR or FDR under different sample sizes (50, 100, 200, 400, 800, 1600). Within each of the white boxes, the colored boxplots represent the distribution of TPR or FDR under different sampling design approach: environmental (red boxes) geographic (blue) hybrid (green) and random (yellow).

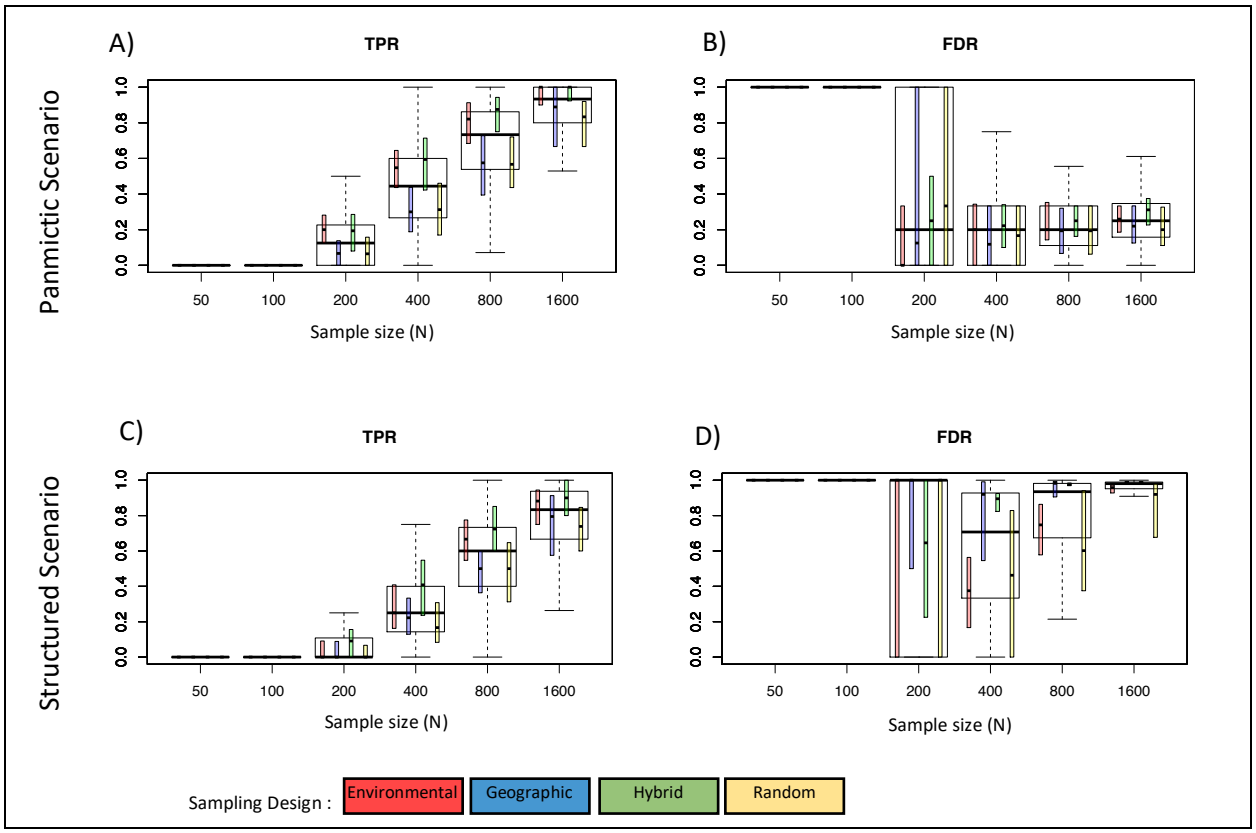
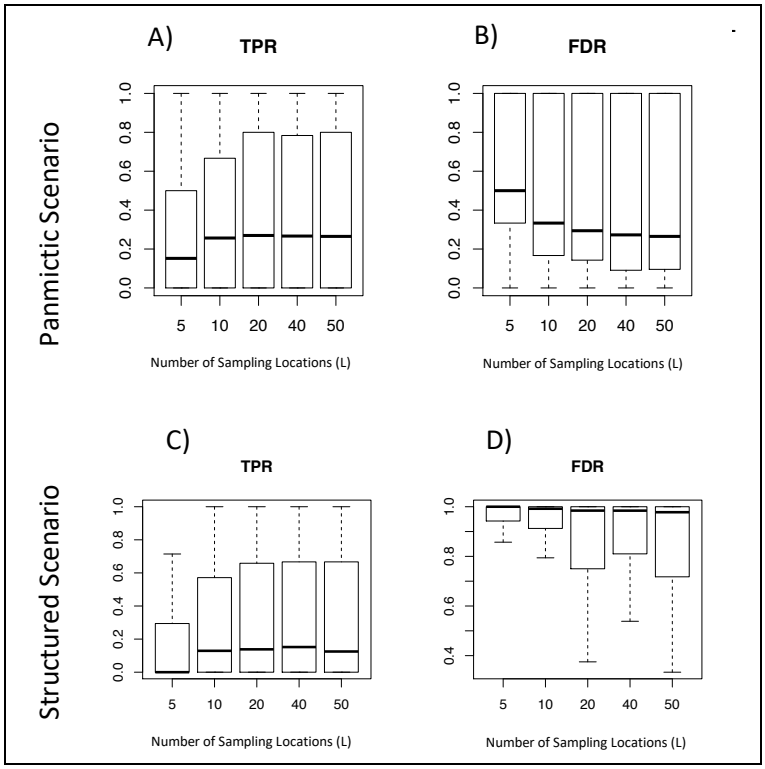
Figure 4. Performances of landscape genomics analysis under different sample size and sampling design approaches. The four plots display how statistical power (A, C) and false discoveries (B, D) of the landscape genomics analyses vary under different levels of sample size and sampling design approaches. The two distinct demographic scenarios under which analyses where run are shown (panmictic: A and B, structured: C and D). The white boxplots represent the distribution of the TPR or FDR under different sample sizes (50, 100, 200, 400, 800, 1600). Within each of the white boxes, the colored boxplots represent the distribution of TPR or FDR under different sampling design approach (environmental, geographic, hybrid and random).