








Review

Re-thinking the environment in landscape genomics

Benjamin Dauphin ^{1,*} Christian Rellstab ¹ Rafael O. Wüest ¹ Dirk N. Karger ¹
Rolf Holderegger ^{1,2} Felix Gugerli ^{1,5} and Stéphanie Manel ^{1,3,4,5}

Detecting the extrinsic selective pressures shaping genomic variation is critical for a better understanding of adaptation and for forecasting evolutionary responses of natural populations to changing environmental conditions. With increasing availability of geo-referenced environmental data, landscape genomics provides unprecedented insights into how genomic variation and underlying gene functions affect traits potentially under selection. Yet, the robustness of genotype–environment associations used in landscape genomics remains tempered due to various limitations, including the characteristics of environmental data used, sampling designs employed, and statistical frameworks applied. Here, we argue that using complementary or new environmental data sources and well-informed sampling designs may help improve the detection of selective pressures underlying patterns of local adaptation in various organisms and environments.

A growing research field with moderate explanatory power?

The environment, and in particular its spatio-temporal changes, exerts strong selection pressures on fitness-related phenotypic traits. If these traits are genetically controlled, natural selection leads to locally adapted populations. Thus, environmental drivers leave a specific signature in the genomes of species and populations. This enables the inference of local adaptation without directly measuring fitness traits, but instead determining the effect of **environmental factors** (see [Glossary](#)) on allele frequencies using **landscape genomic** methods such as **genotype–environment associations (GEA)** [1]. The research field of landscape genomics contributes to the understanding of the genomic mechanisms underlying local adaptation, while detecting the environmental factors driving it. Landscape genomic inference can be expanded to assess the possible maladaptation of populations to environmental change (**genomic offset** [2,3]). This knowledge is particularly valuable in the context of human-induced environmental alterations, such as climate or land-use change, and is therefore relevant for nature conservation and ecosystem management [4,5]. In fact, the concept of genomic offset can inform assisted gene flow and migration strategies that are now at the heart of conservation programmes to support climate-threatened populations and strengthen ecosystem resilience [6].

The field of landscape genomics emerged about 15 years ago with the publication of the first conceptual approach specifically designed for GEA analysis [7]. Thereafter, the advent of next-generation sequencing techniques and large geo-referenced environmental datasets has dramatically increased the quality and quantity of both genomic resources and environmental factors [8]. Consequently, the field of landscape genomics has grown rapidly. According to our literature survey ($n = 278$; [Figure 1A](#); see the supplemental information online for methodological details), more than 50% of all landscape genomic articles published during the survey period 2007–2021 were issued after 2017. A majority of studies initially focused on plants (especially trees), likely due to their sedentariness. This trend has become more balanced in recent years,

Highlights

The increasing availability of new, high-quality geo-referenced environmental data(bases) is stimulating landscape genomic studies of terrestrial and aquatic organisms.

Environmental data (e.g., climate, soil, and topography) are now available at multiple spatial and temporal scales and, together with environmentally and genetically informed sampling designs, enable us to capture selection pressures at high resolution in various organisms.

Statistical advances in genotype–environment association methods now allow testing the response of population genomic variation to complex environments, using nonredundant and informative environmental predictors.

Our understanding of the environmental constraints underlying local adaptation of living organisms has provided insights into the potential responses of populations to environmental changes such as global warming. This understanding is a key component of well-informed biodiversity conservation programmes.

¹Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland

²Institute of Integrative Biology (IBZ), ETH, Zurich, 8092 Zurich, Switzerland

³CEFE, University of Montpellier, CNRS,

EPHE-PSL University, IRD, 34000 Montpellier, France

⁴Institut Universitaire de France,

Paris, France

⁵These co-last authors contributed equally to this work.

*Correspondence: benjamin.dauphin@wsl.ch (B. Dauphin).



resulting in 51% and 47% of all studies (2007–2021) examining plant and animals, respectively (Figure 1A). Landscape genomics has also expanded to the aquatic world (termed seascape or riverscape genomics [9,10] as subfields of landscape genomics), among which marine studies have become particularly frequent in the past 2 years. This development is mirrored by the three most studied ecosystems [forest (32.7%), marine (13.3%), and agriculture (10.8%); Figure 1A], which together account for 56.8% of all studies.

A landscape genomic study typically relies on: (i) an appropriate sampling design to incorporate intraspecific genetic diversity while capturing relevant environmental differences, (ii) geo-referenced environmental data that accurately describe the putative selective pressures of interest acting on populations or individuals, (iii) high-quality genome-wide data, and (iv) appropriate statistical tools to correlate response (genomics, Y) and predictor (environmental, X) variables, while accounting for confounding effects such as neutral genetic structure [11]. As a result, manifold decisions need to be made at an early stage of a study (Figure 2). While there are numerous overviews on methodological issues and genomic approaches (e.g., [12–14]), there are hitherto no general guidelines on the use of environmental data in landscape genomics. Here, we fill this gap and review the types and applications of environmental data in landscape genomics and highlight promising avenues for better characterising environmental factors capturing selection pressures from complex and heterogeneous habitats, with the aim of improving the robustness of landscape genomic outcomes.

Notably, our literature survey revealed that only 36% of all studies reported values of explanatory power of the presented GEA models. Always taking the best model of each study, the median explanatory power across all studies resulted in an R^2 of 0.38 ($n = 93$, standard deviation = 0.29) with values ranging from 0.01 to 0.98 and an overall trend towards low R^2 values in studies with small sample sizes (Figure 1B). Although the maximum explanatory power achievable with such models remains unknown, these values indicate that landscape genomic analyses have revealed only moderate explanatory power yet. Hence, we argue that better informed decisions, for example, based on prior knowledge of environmental data (i.e., choice, type, source, and scale of environmental factors as predictors; Figure 2), should guide the sampling design and anticipate statistical limitations, which may enhance the confidence in future landscape genomic studies. To this end, we highlight trends in the literature and present four main avenues that can be pursued to improve GEA models.

Environmental data as explanatory variables

Researchers have three main data sources to describe the environmental conditions at sampling locations: **in situ measurements**, **remote sensing**, or **spatial interpolation** (Figure 2), the latter being clearly the most widely used to date (e.g., WorldClim; [15]). The open access and user-friendly design of large interpolation-based, geo-referenced **environmental databases** have greatly improved the characterisation and exploration of ecological gradients relevant to identify selective pressures in natural populations. Their convenience is that no additional environmental data has to be collected in the field. The availability of such big data has been facilitated by initiatives led by international consortia to acquire, model, store, and share environmental data on a global scale. For example, the World Ocean Database [16], powered by over 20 000 datasets, centralises a coordinated effort of uniformly curated data for oceanographic, climate, and environmental research. With such resources, ambitious studies have tackled the genomics of adaptation over continental-scale areas (e.g., [17]; Figure 1C), leaving sampling effort and the cost of genomic analyses as the main bottleneck. Nevertheless, further efforts in data acquisition and interpolation modelling are still needed to better grasp environmental variation, particularly in poorly sampled areas of freshwater and marine ecosystems. In parallel, highly accurate pocket-sized devices have been developed to measure biophysical properties at sampling locations,

Glossary

Backward stepwise selection: a variable selection procedure that starts with all variables in the first model tested, then removes one variable at a time and inspects the improvement of the model fit.

Cross-validation: a model validation technique to assess the robustness of the statistical prediction using a resampled random subsample of the dataset.

Environmental database: a digital infrastructure that collects, synthesises, and stores geo-referenced environmental data that can be shared and reused by a broad community.

Environmental factor: a quantification of the natural features of the habitat and climate that are expected to potentially evoke adaptive response in populations.

Forward stepwise selection: a variable selection procedure that starts with a single variable in the first model tested, then adds one variable at a time and inspects the improvement of the model fit.

Genotype–environment association (GEA): also known as environmental association analysis, an approach that seeks associations between environmental and genetic variation based on correlative statistical approaches to identify the molecular basis of local adaptation and the environmental variables that drive it.

Genomic offset: a concept to quantify the difference between the current genomic composition and the one required to cope with a change in environmental conditions in a set of putatively adaptive loci.

In situ measurements: an on-site assessment of environmental conditions to characterise the local abiotic or biotic habitat. Site-specific measurements can be carried out on the ground or from airborne devices via remote sensing.

Landscape genomics: a research field that studies the interactions between adaptive genetic variation and environmental conditions in natural populations.

Remote sensing: the acquisition of environmental characteristics through capturing specific reflections of radiation usually emitted from air-borne vehicles (satellites, planes, drones).

Spatial autocorrelation: a pattern of spatial covariation in which adjacent observations have more similar data values than more distant observations.

allowing *in situ* measurements with high **spatial resolution** and **temporal grain**. These are powerful tools for studying the local habitats of living organisms, especially for those with short generation times and strong population dynamics associated with selection pressures. However, *in situ* measurements often represent short snapshots of environmental conditions and may inadequately reflect the long-term selection pressure within an evolutionary timescale.

Nearly 90% of landscape genomic studies focused on abiotic factors as potential selective pressures, largely neglecting biotic factors that may be of adaptive relevance (Figure 1D; but see [18,19]). Yet, biotic and abiotic factors often operate on different spatial and temporal scales and affect different genomic features, thus providing distinct insights into how selective forces may act in natural populations [20]. For instance, mean ecological indicator values can be derived from floristic compositions at multiple sampling locations to ascribe not only long-term humidity, light availability, soil organic matter content, or pH conditions [21], but also biotic interactions.

Of the studies employing abiotic factors, the vast majority (91.4%; Figure 1D) considered climate factors in GEA analyses, followed by topography (35.3%) and soil (11.5%) factors. This focus reflects the desire to gain knowledge on genetic patterns and molecular mechanisms of adaptation to local (micro-) climatic conditions and how they may respond to future climate change. However, recent studies frequently supplemented climate data with other environmental data (Figure 1D). For instance, Yadav *et al.* [22] investigated local adaptation of two grasshopper species endemic to the Australian Alps and found significant GEAs, particularly with long-term precipitation seasonality, number of frost days, terrain ruggedness, and soil pH conditions. This study illustrates the benefit of incorporating diverse environmental data conditioned on hypotheses to assess the multifactorial nature of local adaptation.

While soil plays a key role in the establishment and persistence of many organisms, including animals, fungi, and plants, few studies have focused on below-ground environmental properties in GEA analyses because appropriate data were largely missing. Yet, great efforts have been made to generate comprehensive datasets of interpolated soil descriptors at decent spatial resolution to complement *in situ* measurements. For instance, a global prediction of bacterial diversity has been established to provide the first reference atlas of dominant bacteria on Earth [23], based on random forest modelling using interpolated soil data at high spatial resolution (i.e., 250 m [24]). Similarly, soil variables were used to explain the global abundance of nematodes with links to soil fertility and functioning [25]. These chemical and physical soil factors are usually released with **cross-validation** scores and uncertainty maps, which help users to perceive the level of confidence and, thereby, the number of observations that support the target geographical areas (e.g., <https://soilgrids.org>). Recently, Lembrechts *et al.* [26] highlighted the difference between *in situ* soil temperature measurements and atmospheric air temperature (up to 10°C in some areas, mean $3.0 \pm 2.1^\circ\text{C}$) at a global scale, in particular in cold and dry biomes. These authors advocated the need to collect soil data in yet unsampled geographical areas to improve the quality and density of environmental data, essential for spatial interpolation. Remote sensing techniques can also help improve the modelling of soil characteristics over time, such as surface soil moisture [27].

Spatial and temporal scales to capture selective pressures

Generation time and population dynamics influence the ability to detect patterns of local adaptation (Figure 3A,B). For example, forest tree species are known to have long generation times and produce large numbers of seeds, resulting in dense seedling layers subjected to strong selection and intraspecific competition for light and nutrients [28]. Characterising sampling site conditions at the time of juvenile tree establishment, considered as the period of strongest selection

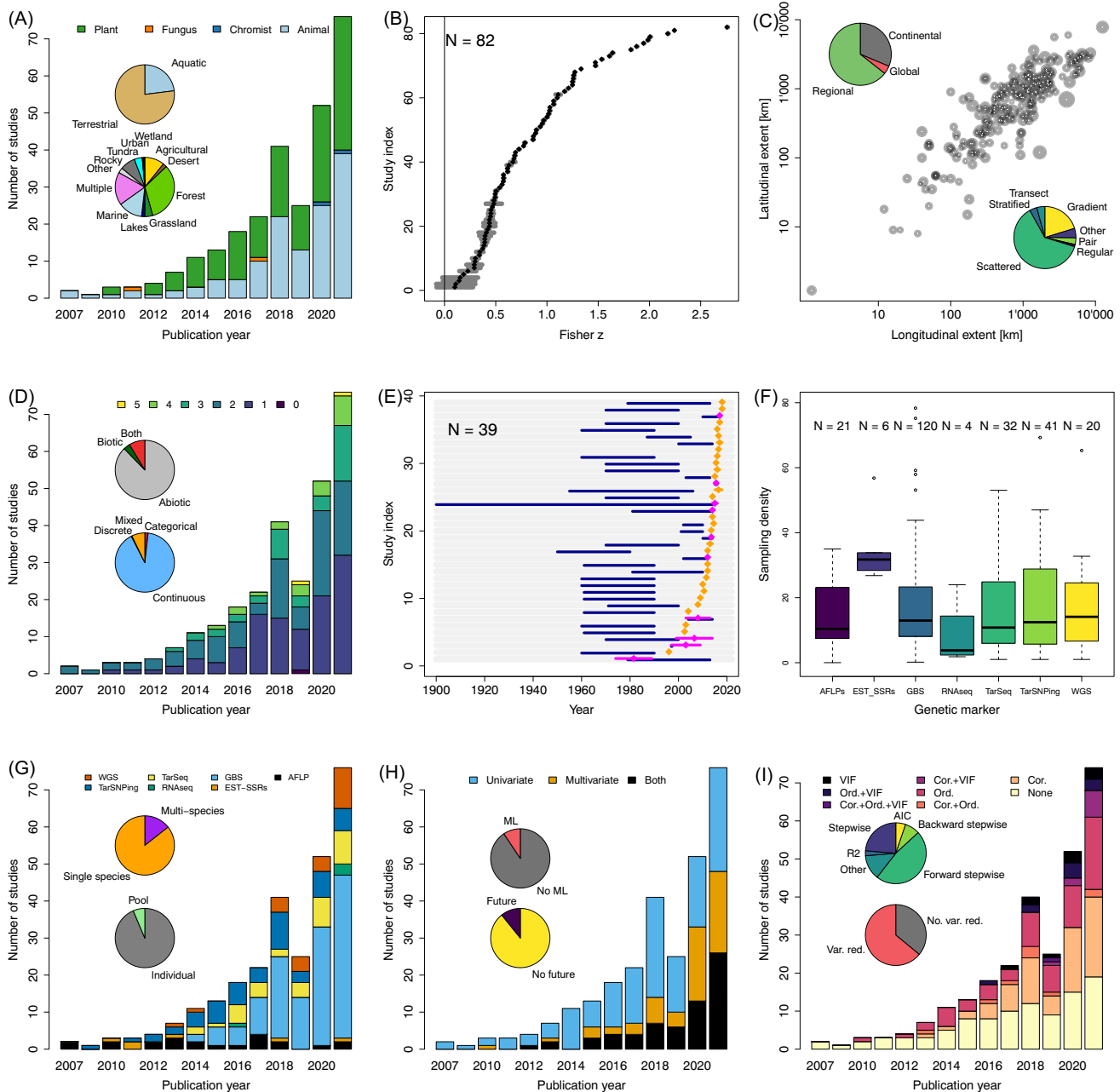
Spatial extent: the geographic area covered by the study design.

Spatial interpolation: a statistical procedure using geo-referenced locations with known values to estimate values at other locations in-between for a given environmental variable.

Spatial resolution: the size of the grain or grid cell of a raster map that is used to describe environmental conditions.

Temporal extent: the time period covered by observations in an environmental dataset.

Temporal grain: the frequency of observations in an environmental dataset.



Trends in Ecology & Evolution

Figure 1. The use of environmental data for landscape genomic studies. Results from a literature survey (search period 2007–2021) on the topics listed in the supplemental information online. (A) Yearly numbers of studies per kingdom. Pie charts indicate the aquatic and terrestrial environments represented and the different ecosystems involved. (B) Distribution of the Fisher's z-transformed correlations based on the strongest explanatory power (i.e., R^2) and sample sizes reported in landscape genomic studies. The horizontal grey bars show the 95% confidence intervals. Studies with confidence intervals not overlapping with the 0-line denote a significant effect of the environment on allele frequencies. Large horizontal confidence intervals relate to small sample sizes. (C) Spatial extent of landscape genomic studies with circle radius reflecting sample size (i.e., number of individuals). Pie charts summarise the geographical coverage of studies and the types of sampling designs used. (D) Yearly numbers of environmental data sources used per study. Pie charts indicate categories and types of variables used. (E) Temporal extent (horizontal blue lines) of climate data used, ranked by sampling year (diamonds). Sampling years that overlap the temporal extent are represented in pink; sampling years that do not fall within the temporal extent are shown in orange. (F) Sampling density (individuals/sampling locations) grouped according to genetic marker type. All Tukey-HSD pairwise comparisons are

(Figure legend continued at the bottom of the next page.)

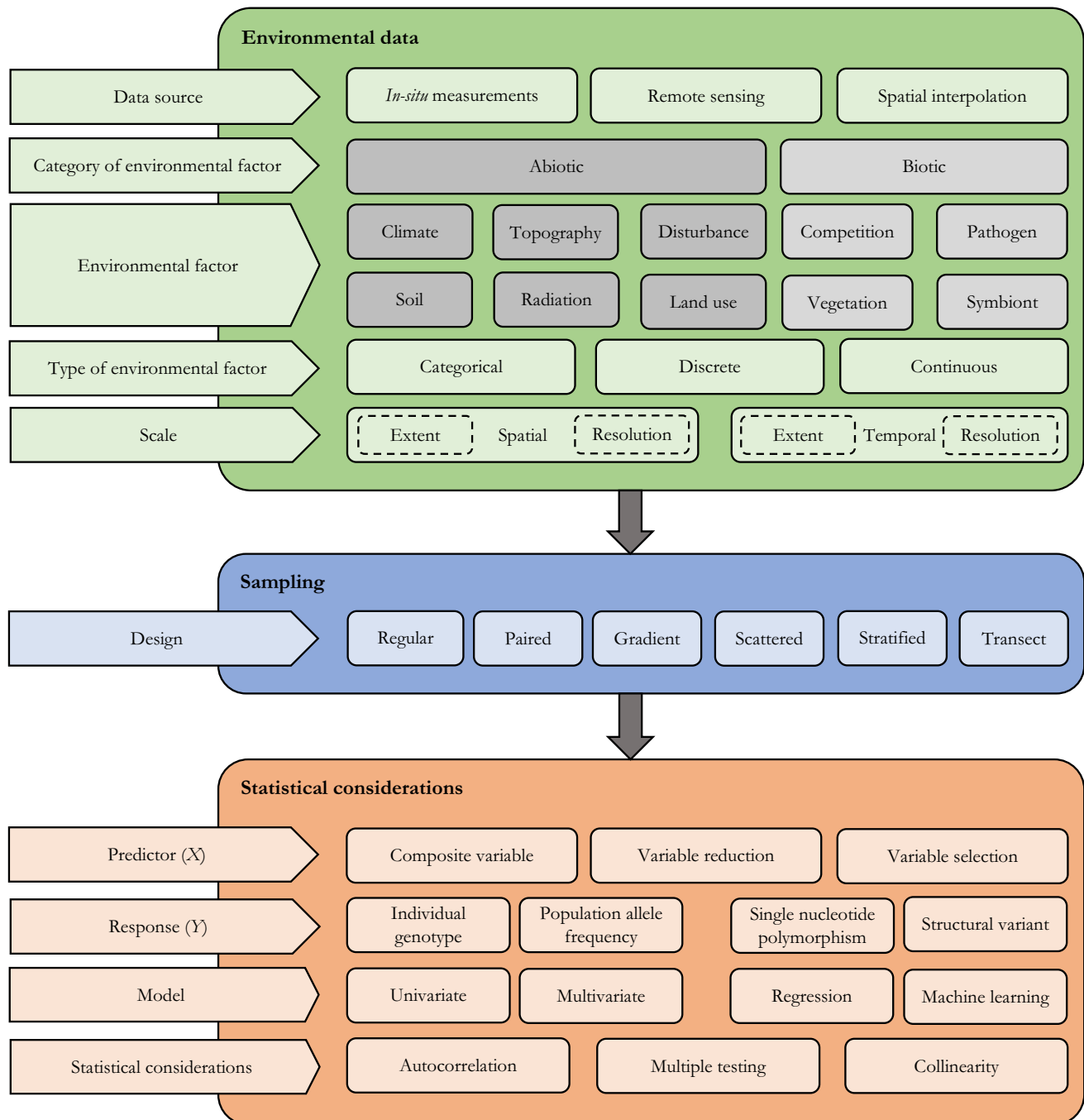
(sometimes more than 100 years ago), helps to capture the adaptation-relevant environmental variation used in GEA, as shown in Swiss stone pine (*Pinus cembra*) [29]. This is achieved by tuning the temporal grain as well as the **temporal extent** of environmental data to target the desired time period (Figure 3A). In a study on the breeding range of the yellow warbler (*Setophaga petechia*) [5], a migratory bird with a short generation time, the authors found significant GEAs with local precipitation conditions averaged over 1970–2000, highlighting the impact of climate-induced selection on the birds' breeding habitat. Interestingly, based on the same terrestrial bioclimatological variables (reference period 1970–2000), populations of the migratory Arctic charr (*Salvelinus alpinus*) were found to be adapted to their local environmental conditions irrespective of breeding range and sampling years and with individuals spanning several age classes (2005–2017) [4]. This time lag between the reference period of environmental predictors and sampling years (Figure 1E), intentional or not, is of primary importance because natural selection can vary across years and operate with antagonistic effects on allele frequencies in natural populations [30]. Hence, we advocate optimising the match between the temporal scales of the environmental data and the selection periods to ensure that they optimally reflect the anticipated evolutionary responses to which the sampled individuals have been exposed.

With numerous geo-referenced environmental databases and knowledge of a species' life history at hand, it is possible to explore and test new adaptation-focused hypotheses by going beyond the often-chosen strategy of using conveniently available factors such as those related to bioclimatology or geography. Although there is probably no single data source that meets all needs in terms of spatial and temporal extent and grain, a set of complementary datasets can be used to capture environmental heterogeneity in space and time (Figure 3). For instance, when studying selection based on long-term climatic means versus minimum/maximum values or extreme climate events, the gradual change in average climate conditions (e.g., sum of annual precipitation for 1970–2010) could be teamed with hourly or daily data (e.g., precipitation during the growing season 2018) to assess local and singular pulses that are strong enough to induce rapid evolution within populations. In addition, as the temporal grain size commonly increases going back in time due to the diminishing number of observations, a trade-off has to be made between incorporating environmental conditions in the early life stages of long-lived organisms and the accuracy of the predictors tested in GEA. This appears to be particularly challenging for extreme events such as droughts, although there has recently been substantial improvement combining multiple independent datasets (e.g., [31]). At what spatial scale selection operates seems an impractical question, as selection acts on phenotypes of single individuals and therefore has no explicit spatial scale. Instead, when planning a landscape genomic study, the focus should be on the spatial scale at which local adaptation is detectable for certain taxonomic groups and life histories, which is intimately linked to sampling strategies and the match between genetic and environmental data outlined in the next section.

Matching sampling designs for genetic and environmental data

There is growing awareness of applying environmentally and genetically informed sampling designs [12]. Sampling for landscape genomic studies generally follows one of the following three strategies: (i) uninformed, (ii) environment-informed, (iii) and environment- and genetics-

non-significant. (G) Yearly numbers of studies by genetic marker type. Pie charts represent the taxonomic breadth of the studies and the types of sequencing approach. (H) Yearly numbers of studies per genotype–environment association (GEA) model. Pie charts indicate the proportion of studies using machine learning (ML) approaches and predictive assessment (genomic offset) under future environmental conditions. (I) Yearly numbers and proportion of studies using variable reduction and selection to minimise the number of environmental factors as explanatory variables; correlation (Cor.), ordination (Ord.), and variance inflation factor (VIF). Pie charts show the variable selection approaches used and the proportion of studies applying one or more variable reduction steps.



Trends in Ecology & Evolution

Figure 2. Overview of the workflow and important decision steps in the use of environmental data in landscape genomic studies. Note that for each step, the suggestions are not exhaustive, but only common examples are given. Steps and options are described in more detail in the main text.

informed designs (Box 1), using, or not, *a priori* knowledge of environmental and/or genetic variation. While the alternatives within the sampling strategy (i) do not rely on environmental or genetic data, regular, scattered, or transect sampling designs are often used (67.6% of the

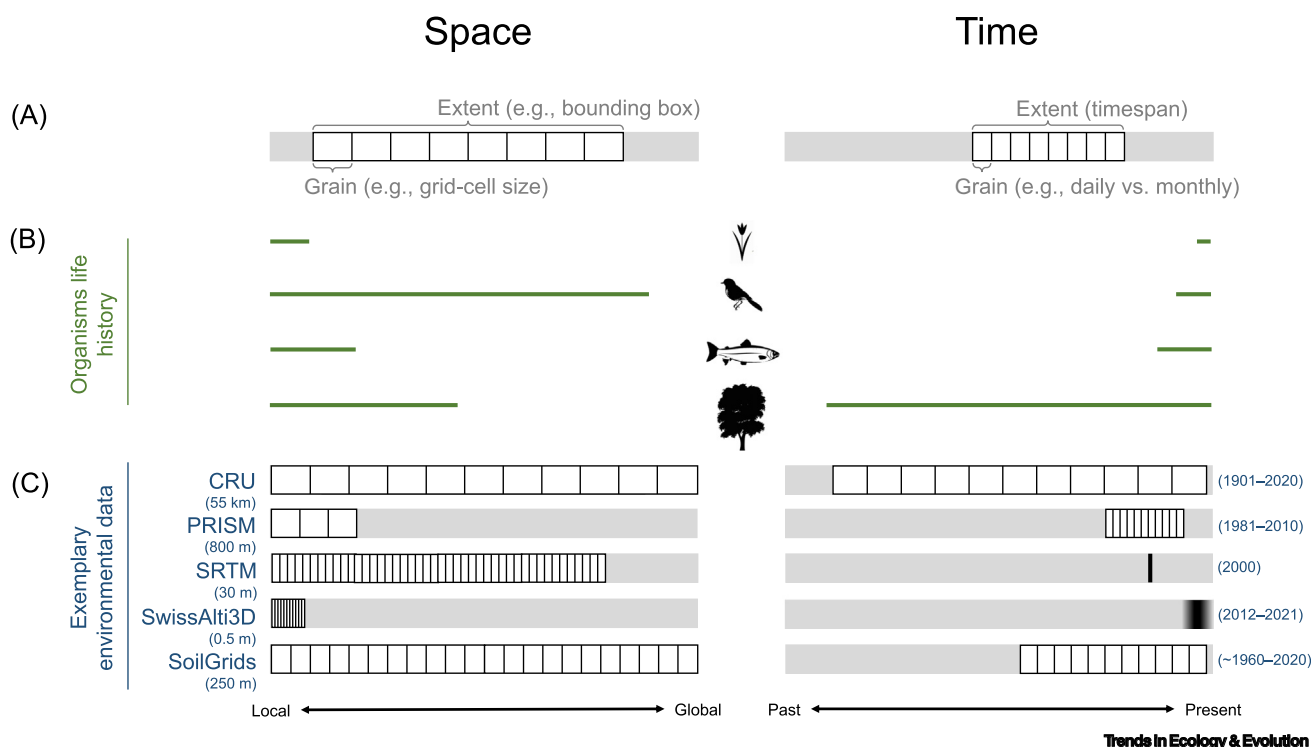


Figure 3. Illustration of spatial and temporal scales with key concepts, organisms' characteristics, and available environmental data. (A) Description of spatial and temporal scales used in landscape genomics, characterised by extent and grain. (B) Spatial and temporal characteristics of exemplary organisms' life history (annual plant, short-lived migratory bird, mid-lived fish, long-lived tree) that are affected by selection pressures. Green lines for each organism symbolise the spatial extent (occurrence, dispersal/migration) and life span, respectively. (C) Examples of environmental data based on different spatial and temporal scales (length of white horizontal bar) and resolution (density of cell subdivision within the bar) according to available data sources: Abbreviations: CRU, Climate Research Unit; PRISM, Parameter-elevation Regressions on Independent Slopes Model; SRTM, Shuttle Radar Topography Mission; SwissAlti3D, Swiss digital elevation model; SoilGrids, Global Gridded Soil Information.

studies surveyed; Figure 1C) as they are assumed to adequately capture both genetic diversity and environmental heterogeneity (see Figure 1A in Box 1). In contrast, strategy (ii) incorporates knowledge from environmental data beforehand to establish a sampling design. Considering paired populations (e.g., wet/dry or low/high elevation) or sampling along known ecological gradient(s) strives for maximising environmental variation covered by sampling locations, while keeping a good representation of a species' distribution range (23.7%; Figures 1C and see Figure 1B in Box 1). However, ecological gradients can be correlated with patterns of demographic history (e.g., [32]; see later), a phenomenon that should be accounted for in sampling and downstream data analysis. In addition, although sampling of multiple paired populations is known to have high statistical power while limiting false positive associations [11], it is rarely implemented in the field (3.4%, but see [33]; Figure 1C), likely due to the limited knowledge of sites with a similar environmental contrast. Yet, such simulation-supported sampling greatly mitigates the putative confounding effect of demographic history by selecting closely related populations growing in consistently contrasting environments [11]. Another drawback is that many of the available environmental factors (e.g., climate data) do not have the spatial resolution to capture the environmental contrast between closely situated locations. Finally, the most informed sampling strategy (iii) takes benefit of environmental data and available genetic information to set up a suitable sampling strategy (see Figure 1C in Box 1). Thus, the sampling design is optimised by incorporating both environmental variation (e.g., climatic zones) and genetic variation (e.g., genetic clusters, evolutionary lineages), usually resulting in stratified sampling for both

Box 1. Sampling designs used in landscape genomic studies

Landscape genomic studies are generally conducted using one of the following six sampling designs, which may or may not incorporate prior knowledge of environmental and/or genetic variation from populations studied (Figure 1). Definitions of these main sampling designs considered in this review, with their possible variants and examples, are next.

Uninformed design

Regular: sampling individuals or populations according to a regular grid.

Scattered: sampling individuals or populations in an unstructured way (e.g., randomly or depending on available sampling locations).

Transect: sampling individuals or populations along regular and linear Euclidian distances (e.g., every 20 km along a geographical axis).

Environment-informed

Paired: sampling individuals in paired populations with contrasting environmental conditions (e.g., cold and warm habitats). This is a special case of categorical sampling and pairs are usually replicated several times.

Gradient: sampling individuals or populations along an environmental gradient (e.g., elevation/latitude, atmospheric temperature, soil pH conditions).

Environment- and genetics-informed

Stratified: sampling evenly (often randomly) within clearly defined subsets of individuals or populations (strata; i.e., an equivalent number of samples per environmental cluster and genetic group). Stratification variation can also be conducted with environmental or genetic information only (e.g., with even numbers of populations from biogeographical regions or a fair representation of previously characterised genetic lineages).

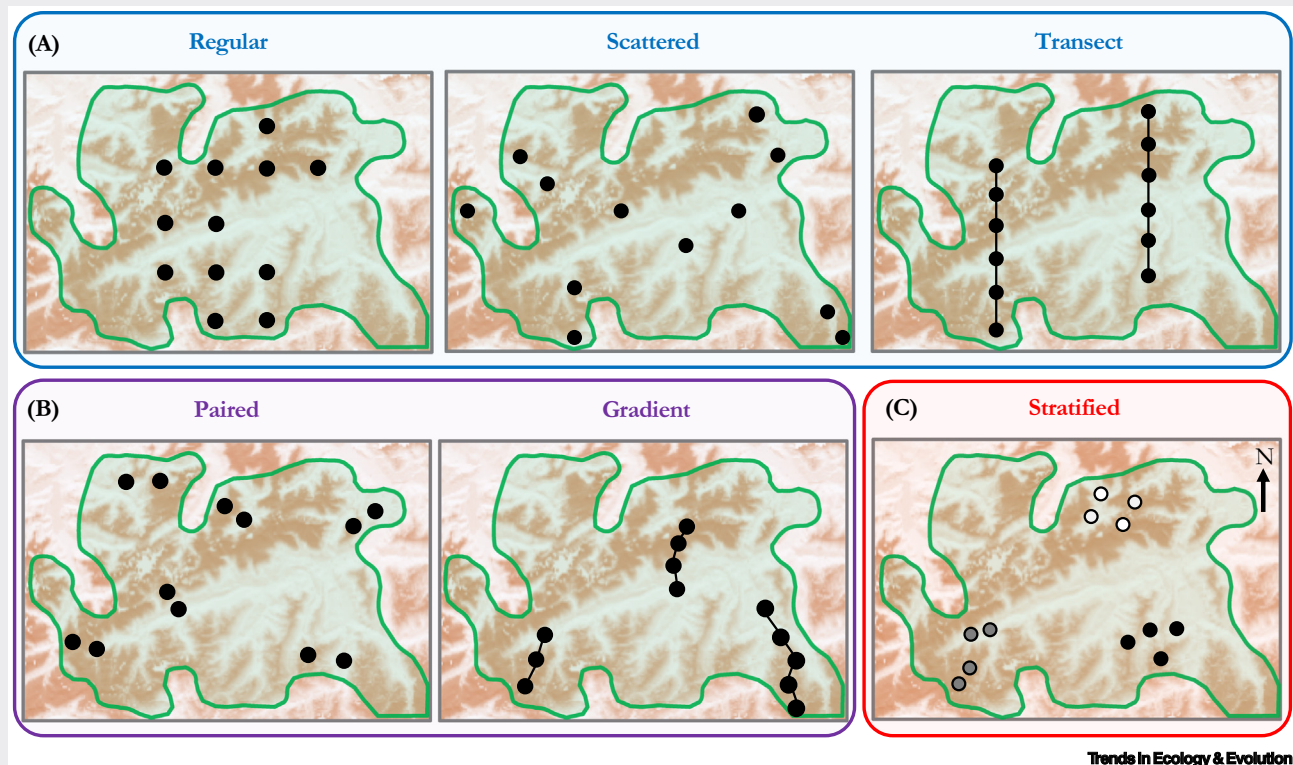


Figure 1. Illustration of main sampling strategies used in landscape genomic studies. Three main strategies summarise the different levels of prior knowledge used in the sampling designs, all of which require occurrence data in the complete or partial species range. The background map symbolises topography with dark and light brown colours for high and low elevations as environmental variable, respectively. Green boundary represents the species' (regional) range limit. (A) The sampling strategy relies on neither environmental nor genetic data. Regular, scattered, or transect sampling approaches are commonly used, as they are assumed to adequately capture both genetic and environmental variation based on the geographical distribution of sampling locations. (B) Environmental data, but no genetic information, is considered. Sampling in closely situated populations experiencing contrasting habitat conditions or along environmental gradients maximise environmental variation covered by sampling locations, while keeping a good representation of a species' distribution range. Finally, (C) takes advantage of environmental data and available genetic information to set up a sampling strategy. The sampling design is optimised by stratifying both environmental variation (e.g., climatic zones) and genetic diversity (e.g., genetic clusters, evolutionary lineages). Circles represent sampling locations, with shadings referring to *a priori* known neutral genetic differences.

components (see Figure 1C in Box 1), with a balanced representation of individuals sharing the same demographic background. In conclusion, we anticipate that choosing a well-informed sampling design such as strategy (iii) may contribute to increased robustness of GEA outcomes.

Sample size per study varied over several orders of magnitude, ranging from 1 to 1193 locations and from only 10 to as many as 17 300 individuals in total, with a median value of 267 individuals falling within the 200–400 units recommended by simulations [12]. Not surprisingly, the number of locations and the **spatial extent** of a study (longitudinal extent \times latitudinal extent) showed a significant positive correlation (Pearson's $r = 0.383$; $P < 0.001$; Figure 1C), but not the number of individuals and spatial extent ($r = 0.083$, $P = 0.222$). Sampling density (i.e., number of individuals/locations) was similar across most types of genetic markers, with a median of less than 20 individuals (Figure 1F).

About two-thirds of the reviewed studies were conducted regionally (64.4%; mean latitudinal extent = 488 km; mean longitudinal extent = 537 km) and only 31.3% and 4.3% of the studies were conducted at the continental or even global (i.e., more than two continents) scale, respectively. Pooled sequencing (DNAs of individuals usually pooled at the population level), which allows to increase the number of individuals sequenced while keeping costs affordable, has rarely been employed (6.4% of studies; Figure 1G). To date, genotyping by sequencing has been the most often used technique to generate genetic markers for GEA (48.2%), followed by targeted SNP genotyping and targeted sequencing in 18.7% and 12.6% of the studies, respectively (Figure 1G). Whole-genome sequencing was employed only in 9.0% of the studies, suggesting that further efforts are required to generate genome-wide resources for non-model organisms and further reduce sequencing costs per base.

Statistical considerations in association models

There are multifaceted benefits and drawbacks of the statistical models mostly used to test genomic responses to environmental factors (Box 2), and the models' robustness may be improved by optimally incorporating adaptation-relevant environmental variation and accounting for neutral genetic structure. A notable change in the practice of landscape genomics over time is that researchers model the multivariate genetic response to capture the polygenic nature of local adaptation. For example, redundancy analysis (RDA; Figure 1H) [34] is a multivariate model (for Y and X) that identifies covarying allele frequencies associated with several environmental factors (Box 2). Such a model is particularly appropriate in the usual case of polygenic responses occurring in complex environments [35]. Another recent advance in GEA is the use of machine learning approaches (i.e., Gradient Forest) to account for the nonlinear genomic responses to environmental predictors [36,37], while testing for a set of environmental predictors in a single model (see Table 1 in Box 2).

An important requirement for GEA analysis is to remove the confounding effect of demographic history, as the genetic signature is expected to be different between neutral loci and those under selection. However, neutral evolutionary processes can mimic patterns of local adaptation in the genome (e.g., when allele frequencies resulting from range shifts coincide with ecological gradients of adaptive relevance) [32]. Earlier GEA studies relied on **spatial autocorrelation** to account for population demography and limit false-discovery rate [38]. For instance, the spatial locations of samples were transformed with Moran's eigenvector maps and then used as additional explanatory variables to the environmental predictors in statistical analysis [39]. However, with the development of large genomic datasets (Figure 1F), we recommend incorporating neutral genetic structure directly as a covariate in statistical models, for example, using the residuals after linear regression on overall structure (partial RDA [40]) or as random factors

Box 2. Genotype–environment association models

Different statistical frameworks are used to investigate associations between genetic and environmental variation while accounting for the confounding effect of population demography based on neutral genetic structure (Table I). The response variable Y refers to genetic variation assessed at the individual or population level, either as individual genotypes or translated into allele frequencies. The predictor variable X denotes environmental variation using original or composite variable(s) (Figure 2 and Table I).

Univariate (Y) models

Linear model: the response variable (Y , allele frequency) is linearly associated to the environmental predictor (X). Confounding factors can be added as predictors (e.g., neutral genetic structure).

Generalized linear model: logistic regression (allele frequency) or Poisson regression (allele or genotype counts) are based on a linear association between the logit or the log of Y and environmental factors, respectively. Confounding factors can be added as predictors (e.g., neutral genetic structure).

Mixed effect model: an extension of linear models to allow for both fixed and random effects. Here, fixed effects are environmental factors and random effects denote, for example, genetic distances or latent factors, as in latent factor mixed model (LFMM) [49]. The non-independence of the random effects and response variable calls for randomisation tests.

Bayesian hierarchical model: this Bayesian method (i) calculates the posterior distribution (null model) from the empirical allele frequencies, and (ii) tests the regression between population allele frequencies and environmental factors against the null model. Two main implementations for GEA have been developed so far, Bayenv2 [42] and BayPass2 [50], both of which control for population genetic structure using a covariance matrix.

F -model: assumes that all populations share a common pool of migrants, while their effective size and migration rate are population-specific. Population structure at each locus is described by local F_{ST} estimates, which measure genetic differentiation between each local population and the migrant pool [51]. The effect of environmental differentiation on each locus can then be tested with a local adaptation model, as implemented in BayeScEnv [52].

Gradient forest: an extension of the random forest machine learning approach that models the nonlinear change of allele frequency at each locus along an environmental gradient. The gradient forest algorithm assesses allele-specific turnover functions that identify major tipping point environmental conditions [53]. In its current implementation, this approach has no explicit factor accounting for neutral population structure.

Multivariate (Y) models

Redundancy analysis (RDA): a constrained ordination that models linear relationships between genetic variation and environmental factors, from which associations are interpreted in a principal component analysis (PCA). To include neutral population structure or any other confounding factor, partial RDA (pRDA [54]) should be applied (Table I). Note that incorporating the population genetic structure may inhibit the detection of loci under selection [54].

Table I. Overview of the most commonly used statistical methods for testing the genetic response to environmental factors

Genetic response (Y)	Environmental predictor (X)	Statistical method	Neutral genetic structure	Implementation ^a	Refs
Univariate	Univariate	F -model	β	BayeScEnv	[52]
		Bayesian hierarchical model	Covariance matrix	Bayenv2 ^b	[42,55]
				BayPass2 ^b	[41,50]
	Multivariate ^c	Logistic regression model ^d	Principal component(s)	SamBada ^e	[56,57]
		Latent factor mixed model	Latent (random) factors(s)	LEA3 ^e	[43,49,58]
		Gradient Forest	Not yet implemented	GradientForest ^e	[37]
Multivariate	Multivariate ^c	Redundancy analysis	Principal component(s)	Vegan ^e	[40,54,59]

^aSoftware or R packages are mentioned under the implementation column.

^bImplementations that can specifically account for pooled data and correct for pool size.

^cNonlinearity can be added in models by adding polynomial factors (e.g., quadratic terms).

^dSuitable GEA model for haploid data.

^eR packages.

characterised by a covariance matrix (Bayenv2 and BayPass2 [41,42]; Box 2) or latent factor(s) (LEA3 [43]; Box 2). These covariates are best assessed with a set of independent neutral loci or with the full genomic dataset available that presumably captures the overall population structure, reducing false-positives under various demographic scenarios [11]. Ideally, a panel of unlinked neutral SNPs should be targeted from genomic sites in noncoding regions [44].

Considering several environmental predictors of adaptive genomic response in multiple regression models is likely to increase the robustness of landscape genomic analyses [45]. However, the presence of many environmental factors as predictors in classical models (e.g., regression, RDA) requires a preliminary reduction of the predictors to avoid collinearity among them, which can lead to unstable model fit and biased interpretation of results (e.g., regression coefficients or predictor significance) [46]. This reduction step can be based on various principles, for example, pairwise correlation, variation inflation factor, or principal component analysis (PCA; Figure 1). Excluding correlated factors may require additional information, such as expert knowledge of putatively selective predictors or species information to retain the most relevant factor(s). Although PCA is enticing, its first synthetic orthogonal axes (i.e., principal components, PCs) of environmental factors may not necessarily represent the ecological drivers of divergent selection, nor may they adequately capture complex environments, as the PCs may be dominated by factors that exhibit the greatest variation and covary the most. When possible, we recommend using a reduced set of uncorrelated environmental factors in GEA or to consider advanced forms of regression models based on regularization procedures that have shown to be robust to multicollinearity [3].

Although rarely used according to our literature review (14.0%; Figure 1), if the chosen analytical framework allows, variable selection should be an important step to inform on the explanatory power of predictors. Variable selection uses **backward stepwise selection** or **forward stepwise selection** of environmental factors to retain the most informative and largely independent predictors in a model, thus avoiding model overfitting. The choice of the best model, and thus of the variables retained, is usually based on Akaike's information criterion, the Bayesian information criterion, or adjusted R^2 . However, stepwise regression procedures in landscape genomics suffer from omitting cross-validation using an independent dataset to assess possible biases in parameter estimation [47] when moving from explanatory to predictive applications (but see [48]).

Concluding remarks: perspectives in landscape genomic research

We emphasise here a few aspects in which increasingly available environmental data will play a key role in the future: (i) recent and future changes in selective pressure as a consequence of direct human activity, (ii) soil factors as integrative descriptors of local site conditions, and (iii) effects of climate change on selection processes relevant for biodiversity conservation.

Across the globe, human settlements are growing in area and density, evoking formerly unknown local selection pressures (e.g., microclimate, artificial habitat, disturbance [60]). Likewise, agriculture and fisheries are expanding in space and increasing in intensity of use [61,62]. Under such conditions, organisms are deemed to migrate to the remaining, less affected habitats, or to adapt rapidly (see Outstanding questions). High-resolution environmental data for urban areas are assessed owing to our interest in microclimatic effects of buildings and streets on human well-being [63]. In turn, agricultural practice increasingly relies on digital data (e.g., for pest control and optimised nutrient and water supply) [64]. Using automatic identification system messages, it is now possible to track industrial fishing vessels [65]. Hence, these big data on urban, agricultural, or commercial fishing pressures allow researchers to study rapid evolution under recent environmental changes. Such studies could focus on paired comparisons (e.g., [66]), either in space (urban vs. rural, intensive vs. extensive agriculture, outside vs. inside marine reserves) or in time (before vs. after urbanisation or agricultural/fishery intensification). However, these recent and marked environmental changes may represent unstable conditions in a way that compromises the detection of evolutionary responses using GEA, and concern only organisms with a short-term genomic response. Nevertheless, landscape genomics may benefit from the increased interest in

Outstanding questions

What are the main environmental drivers of genetic adaptation and at which spatial and temporal scales can we detect them? What is the contribution of long- and short-term (e.g., extreme events) environmental changes that leave discernible signatures in the genome? Shall we put more emphasis on extreme events such as heat waves, floods, or droughts to capture selective pressures underlying rapid evolutionary processes? New developments in remote sensing techniques and machine learning might enable such detailed environmental characterisation and the detection of patterns for landscape genomic analyses.

How can spatially heterogeneous soil characteristics best describe selective pressures at small spatial scale? How do chemical and physical soil properties and communities of soil microbiota impact individuals and populations of above- and below-ground terrestrial organisms? In particular, soil nutrient content, soil temperature, and water availability should exert strong selection on natural populations.

To what extent do biotic factors drive patterns of local adaptation in terrestrial and aquatic organisms? A possible avenue for organisms depending on soil as a substrate would be to describe bacterial and fungal communities and include their characteristics as explanatory factors. Which composite community predictors should be selected for GEA analysis? Possible options are abundance data, derived ecological indicators such as competition or vegetation indices, or diversity measures. Moreover, how common is co-adaptation between symbiotic partners and under which environmental conditions does it take place?

How can we use knowledge of adaptive genomic variation for conservation purposes? Are existing environmental factors accurate enough to support conservation decisions? Genomic offset analyses based on well-justified environmental factors represent an interesting tool for conservation measures, but need additional validation and have to be combined with complementary observations such as demographic trends.

anthropogenically altered environments and provide genetic-informed contributions to evaluating the fate of species inhabiting these environments.

Clearly, the set of environmental factors considered in GEA should be broadened in future studies to capture as many of the selective pressures as possible, but still in relation to plausible adaptation mechanisms for a given species. A trade-off has to be found between environmental factors chosen on the basis of hypotheses (expert opinion) and those derived from a human-free mind-set (all possible factors), with the risk of missing important variables having the strongest association(s) with genetic variation in the former case. Besides biotic factors that have hitherto been largely ignored (see Outstanding questions), abiotic factors have commonly been restricted to climate and topography. Particularly for plants and underground-dwelling species, soil properties represent an under-explored dimension of the environment. Physical, chemical, and biotic soil properties impose prominent selective pressures on individuals, populations, and communities. While heavy metal or salt tolerance has been widely studied in relation to plant [67–69] and fungi [70,71] adaptation, many other aspects still await consideration. Among these are acidification (which may indirectly affect toxic ion concentrations in soils), soil compaction (change in aeration), soil water content (drought stress related to precipitation, evapo-transpiration, and terrain), and biotic activity (micro- and macro-organisms). We see great prospects for interpolated soil factors or, even better, measured *in situ*, to complement the topo-climatic factors hitherto favoured.

In nature conservation, it is commonly advocated to optimise ecological similarity between source and target habitats in translocation efforts [6]. With increased robustness of landscape genomic approaches, we see vast opportunities for deepening our understanding of how particular genotypes may respond to a given environment in their present or new location (e.g., climate change), both now and under predicted conditions in the future [3]. Incorporating more accurate and species-specific comprehensive environmental datasets will allow us to better grasp polygenic responses and, hence, improve our knowledge of selective pressures to support conservation decisions. As such, biodiversity conservation at the intraspecific level will specifically address the adaptive fraction of genetic diversity, a key topic for conservation practitioners [72].

Acknowledgments

S.M. was supported by the WSL visiting fellowship. We thank two anonymous reviewers and the Editor for valuable input on earlier versions of this manuscript.

Declaration of interests

The authors declare no competing interests.

Supplemental information

Supplemental information associated with this article can be found online <https://doi.org/10.1016/j.tree.2022.10.010>.

References

1. Rellstab, C. *et al.* (2015) A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370
2. Capblancq, T. *et al.* (2020) Genomic prediction of (mal)adaptation across current and future climatic landscapes. *Annu. Rev. Ecol. Evol. Syst.* 51, 245–269
3. Rellstab, C. *et al.* (2021) Prospects and limitations of genomic offset in conservation management. *Evol. Appl.* 14, 1202–1212
4. Layton, K.K.S. *et al.* (2021) Genomic evidence of past and future climate-linked loss in a migratory Arctic fish. *Nat. Clim. Chang.* 11, 158–165
5. Bay, R.A. *et al.* (2018) Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science* (1979) 359, 83–86
6. Hoffmann, A.A. *et al.* (2021) Genetic mixing for population management: from genetic rescue to provenancing. *Evol. Appl.* 14, 634–652
7. Joost, S. *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol. Ecol.* 16, 3955–3969
8. Xia, J. *et al.* (2020) Research challenges and opportunities for using big data in global change biology. *Glob. Chang. Biol.* 26, 6040–6061

9. Liggins, L. *et al.* (2020) Seascape genomics: contextualizing adaptive and neutral genomic variation in the ocean environment. In *Population Genomics: Marine Organisms* (1st edn) (Oleksiak, M.F. and Rajora, O.P., eds), pp. 171–218, Springer
10. Grummer, J.A. *et al.* (2019) Aquatic landscape genomics and environmental effects on genetic variation. *Trends Ecol. Evol.* 34, 641–654
11. Lotterhos, K.E. and Whitlock, M.C. (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol. Ecol.* 24, 1031–1046
12. Selmoni, O. *et al.* (2020) Sampling strategy optimization to increase statistical power in landscape genomics: a simulation-based approach. *Mol. Ecol. Resour.* 20, 154–169
13. Wold, J. *et al.* (2021) Expanding the conservation genomics toolbox: incorporating structural variants to enhance genomic studies for species of conservation concern. *Mol. Ecol.* 30, 5949–5965
14. Bourgeois, Y.X.C. and Warren, B.H. (2021) An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Mol. Ecol.* 30, 6036–6071
15. Fick, S.E. and Hijmans, R.J. (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315
16. Boyer, T.P. *et al.* (2018) World Ocean Database. In *NOAA Atlas NESDIS (87)* (Mishonov, A.V., ed.), pp. 1–207, NOAA
17. Babin, C. *et al.* (2017) RAD-Seq reveals patterns of additive polygenic variation caused by spatially-varying selection in the American eel (*Anguilla rostrata*). *Genome Biol. Evol.* 9, 2974–2986
18. Maselko, J. *et al.* (2020) Long-lived marine species may be resilient to environmental variability through a temporal portfolio effect. *Ecol. Evol.* 10, 6435–6448
19. Fraik, A.K. *et al.* (2020) Disease swamps molecular signatures of genetic-environmental associations to abiotic factors in Tasmanian devil (*Sarcophilus harrisii*) populations. *Evolution (N Y)* 74, 1392–1408
20. Descombes, P. *et al.* (2020) Spatial modelling of ecological indicator values improves predictions of plant distributions in complex landscapes. *Ecography* 43, 1448–1463
21. Leempoel, K. *et al.* (2018) Multiscale landscape genomic models to detect signatures of selection in the alpine plant *Biscutella laevigata*. *Ecol. Evol.* 8, 1794–1806
22. Yadav, S. *et al.* (2021) Microgeographical adaptation corresponds to elevational distributions of congeneric montane grasshoppers. *Mol. Ecol.* 30, 481–498
23. Delgado-Baquerizo, M. *et al.* (2018) A global atlas of the dominant bacteria found in soil. *Science* 325, 320–325
24. Hengl, T. *et al.* (2017) SoilGrids250m: global gridded soil information based on machine learning. *PLoS One* 12, e0169748
25. van den Hoogen, J. *et al.* (2019) Soil nematode abundance and functional group composition at a global scale. *Nature* 572, 194–198
26. Lembrechts, J.J. *et al.* (2022) Global maps of soil temperature. *Glob. Chang. Biol.* 28, 3110–3144
27. Chen, Y. *et al.* (2021) An improved global remote-sensing-based surface soil moisture (RSSM) dataset covering 2003–2018. *Earth Syst. Sci. Data* 13, 1–31
28. Petit, R.J. and Hampe, A. (2006) Some evolutionary consequences of being a tree. *Annu. Rev. Ecol. Syst.* 37, 187–214
29. Dauphin, B. *et al.* (2021) Genomic vulnerability to rapid climate warming in a tree species with a long generation time. *Glob. Chang. Biol.* 27, 1–15
30. Troth, A. *et al.* (2018) Selective trade-offs maintain alleles underpinning complex trait variation in plants. *Science* 361, 475–478
31. Cook, E.R. *et al.* (2015) Old World megadroughts and pluvials during the Common Era. *Sci. Adv.* 1, 1–22
32. Yeaman, S. *et al.* (2016) Convergent local adaptation to climate in distantly related conifers. *Science* 353, 23–26
33. Buehler, D. *et al.* (2013) An outlier locus relevant in habitat-mediated selection in an alpine plant across independent regional replicates. *Evol. Ecol.* 27, 285–300
34. Forester, B.R. *et al.* (2018) Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Mol. Ecol.* 27, 2215–2233
35. Yeaman, S. (2022) Evolution of polygenic traits under global vs local adaptation. *Genetics* 220, 1–15
36. Yu, Y. *et al.* (2022) Using landscape genomics to delineate seed and breeding zones for lodgepole pine. *New Phytol.* 235, 1653–1664
37. Fitzpatrick, M.C. *et al.* (2021) Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. *Mol. Ecol. Resour.* 21, 2749–2765
38. Manel, S. *et al.* (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Mol. Ecol.* 19, 3760–3772
39. Wagner, H.H. *et al.* (2017) Spatial detection of outlier loci with Moran eigenvector maps. *Mol. Ecol. Resour.* 17, 1122–1135
40. Capblancq, T. and Forester, B.R. (2021) Redundancy analysis: a Swiss Army Knife for landscape genomics. *Methods Ecol. Evol.* 12, 2298–2309
41. Olazcuaga, L. *et al.* (2021) A whole-genome scan for association with invasion success in the fruit fly *Drosophila suzukii* using contrasts of allele frequencies corrected for population structure. *Mol. Biol. Evol.* 37, 2369–2385
42. Günther, T. and Coop, G. (2013) Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220
43. Caye, K. *et al.* (2019) LFMM 2: fast and accurate inference of gene–environment associations in genome-wide studies. *Mol. Biol. Evol.* 36, 852–860
44. Lotterhos, K.E. (2019) The effect of neutral recombination variation on genome scans for selection. *G3 Genes Genomes Genetics* 9, 1851–1867
45. Manel, S. *et al.* (2010) Common factors drive adaptive genetic variation at different spatial scales in *Arabidopsis alpina*. *Mol. Ecol.* 19, 3824–3835
46. Dormann, C.F. *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 27–46
47. Whittingham, M.J. *et al.* (2006) Why do we still use stepwise modelling in ecology and behaviour? *J. Anim. Ecol.* 75, 1182–1189
48. Manel, S. *et al.* (2018) Predicting genotype environmental range from genome–environment associations. *Mol. Ecol.* 27, 2823–2833
49. Frichot, E. *et al.* (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699
50. Gautier, M. (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* 201, 1555–1579
51. Foll, M. and Gaggiotti, O.E. (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180, 977–993
52. de Villemereuil, P. and Gaggiotti, O.E. (2015) A new F_{ST} -based method to uncover local adaptation using environmental variables. *Methods Ecol. Evol.* 6, 1248–1258
53. Fitzpatrick, M.C. and Keller, S.R. (2015) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecol. Lett.* 18, 1–16
54. Forester, B.R. *et al.* (2016) Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol. Ecol.* 25, 104–120
55. Coop, G. *et al.* (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185, 1411–1423
56. Duruz, S. *et al.* (2019) Rapid identification and interpretation of gene–environment associations using the new R.SamBada landscape genomics pipeline. *Mol. Ecol. Resour.* 19, 1355–1365
57. Stucki, S. *et al.* (2017) High performance computation of landscape genomic models including local indicators of spatial association. *Mol. Ecol. Resour.* 17, 1072–1089
58. Gain, C. and François, O. (2021) LEA3: factor models in population genetics and ecological genomics with R. *Mol. Ecol. Resour.* 21, 2738–2748
59. Oksanen, J. *et al.* (2022) Package vegan: Community ecology package. Available at: <https://cran.r-project.org/web/packages/vegan/index.html>
60. Santangelo, J.S. *et al.* (2022) Global urban environmental change drives adaptation in white clover. *Science* 375, 1275–1281
61. Cinner, J.E. *et al.* (2020) Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. *Science* 368, 307–311

62. Zabel, F. *et al.* (2019) Global impacts of future cropland expansion and intensification on agricultural markets and biodiversity. *Nat. Commun.* 10, 1–10
63. Toparlak, Y. *et al.* (2017) A review on the CFD analysis of urban microclimate. *Renew. Sustain. Energ. Rev.* 80, 1613–1640
64. Faye, E. *et al.* (2016) A toolbox for studying thermal heterogeneity across spatial scales: from unmanned aerial vehicle imagery to landscape metrics. *Methods Ecol. Evol.* 7, 437–446
65. Kroodsma, D.A. *et al.* (2018) Tracking the global footprint of fisheries. *Science* 359, 904–908
66. Harris, S.E. *et al.* (2013) Signatures of rapid evolution in urban and rural transcriptomes of white-footed mice (*Peromyscus leucopus*) in the New York metropolitan area. *PLoS One* 8, 1–19
67. Krämer, U. (2010) Metal hyperaccumulation in plants. *Annu. Rev. Plant Biol.* 61, 517–534
68. Rahman, M.M. *et al.* (2021) Adaptive mechanisms of halophytes and their potential in improving salinity tolerance in plants. *Int. J. Mol. Sci.* 22, 1–28
69. Sailer, C. *et al.* (2018) Transmembrane transport and stress response genes play an important role in adaptation of *Arabidopsis halleri* to metalliferous soils. *Sci. Rep.* 8, 1–13
70. Bazzicalupo, A.L. *et al.* (2020) Fungal heavy metal adaptation through single nucleotide polymorphisms and copy-number variation. *Mol. Ecol.* 29, 4157–4169
71. Bazzicalupo, A.L. *et al.* (2020) Gene copy number variation does not reflect structure or environmental selection in two recently diverged California populations of *Suillus brevipes*. *G3 Genes Genomes Genetics* 10, 4591–4597
72. Pärli, R. *et al.* (2021) Developing a monitoring program of genetic diversity: what do stakeholders say? *Conserv. Genet.* 22, 673–684