MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 4348–4370 doi: 10.1111/mec.13322

INVITED REVIEWS AND SYNTHESES

A practical guide to environmental association analysis in landscape genomics

CHRISTIAN RELLSTAB,* FELIX GUGERLI,* ANDREW J. ECKERT,† ANGELA M. HANCOCK‡ and ROLF HOLDEREGGER*§

*WSL Swiss Federal Research Institute, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland, †Department of Biology, Virginia Commonwealth University, Richmond, VA 23284, USA, ‡Faculty of Molecular Biology, Max F. Perutz Laboratories and University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, §ETH Zürich, Institute of Integrative Biology, Universitätstrasse 16, 8092 Zürich, Switzerland

Abstract

Landscape genomics is an emerging research field that aims to identify the environmental factors that shape adaptive genetic variation and the gene variants that drive local adaptation. Its development has been facilitated by next-generation sequencing, which allows for screening thousands to millions of single nucleotide polymorphisms in many individuals and populations at reasonable costs. In parallel, data sets describing environmental factors have greatly improved and increasingly become publicly accessible. Accordingly, numerous analytical methods for environmental association studies have been developed. Environmental association analysis identifies genetic variants associated with particular environmental factors and has the potential to uncover adaptive patterns that are not discovered by traditional tests for the detection of outlier loci based on population genetic differentiation. We review methods for conducting environmental association analysis including categorical tests, logistic regressions, matrix correlations, general linear models and mixed effects models. We discuss the advantages and disadvantages of different approaches, provide a list of dedicated software packages and their specific properties, and stress the importance of incorporating neutral genetic structure in the analysis. We also touch on additional important aspects such as sampling design, environmental data preparation, pooled and reduced-representation sequencing, candidate-gene approaches, linearity of allele–environment associations and the combination of environmental association analyses with traditional outlier detection tests. We conclude by summarizing expected future directions in the field, such as the extension of statistical approaches, environmental association analysis for ecological gene annotation, and the need for replication and post hoc validation studies.

Keywords: adaptive genetic variation, ecological association, environmental correlation analysis, genetic–environment association, genotype–environment correlation, local adaptation, natural selection, neutral genetic structure, population genomics

Received 28 January 2015; revision received 10 July 2015; accepted 13 July 2015

The emergence of landscape genomics

Changing environmental conditions force organisms to be phenotypically plastic, migrate or adapt to avoid extinction. Local adaptation (Williams 1966; Kawecki & Ebert 2004; Savolainen *et al.* 2013) is the response to differential selective pressures among populations and habitats, acting on genetically controlled fitness differences among individuals. Hence, genes underlying heritable phenotypic variation are of great interest in evolution and ecology. To identify such genes, two types of approaches are currently used (Barrett & Hoekstra 2011). Top-down approaches, such as genomewide

Correspondence: Christian Rellstab, Fax: +41 44 739 2215; E-mail: christian.rellstab@wsl.ch

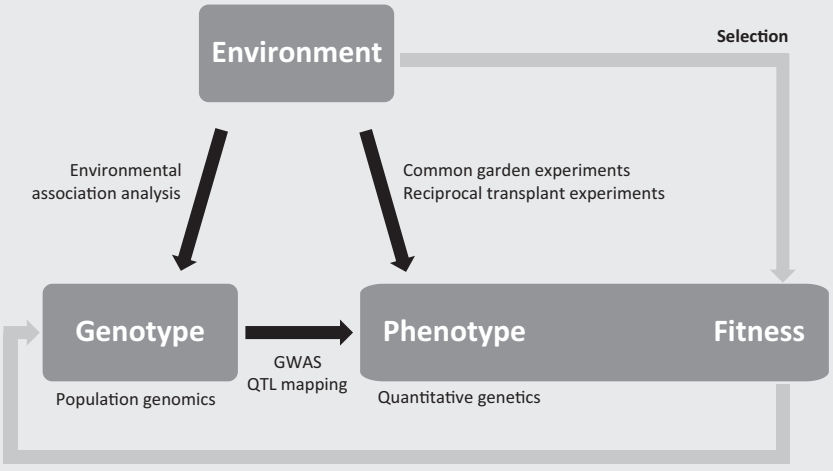
© 2015 John Wiley & Sons Ltd

association studies (GWAS, reviewed in Korte & Farlow 2013) and quantitative trait locus (QTL) mapping (reviewed in Stinchcombe & Hoekstra 2008), take advantage of phenotypic measurements and relate them to genotypic data (Box 1). Bottom-up approaches, such as population and landscape genomics, use genomic information to identify signatures of adaptive genetic variation and relate them to evolutionary processes and environmental variation. In population genomics, approaches based on identifying regions of high genetic differentiation among populations as compared to a neutral model are commonly used to detect positive selection (Luikart *et al.* 2003). Although they are frequently used, these outlier tests have drawbacks. First, in the case of positive selection, they are aimed at identifying loci that underwent selective sweeps of beneficial alleles. Adaptation to local conditions, however, can lead to subtle changes in allele frequencies that are hardly detected by outlier tests, for example in the case of polygenic additive effects (Pritchard & Di Rienzo 2010) or under high geneflow counteracting patterns of local adaptation (Kawecki & Ebert 2004). Second, outlier tests make the assumption that selection pressures differ among populations, but usually do not attempt to link to specific selection pressures that underlie adaptation. An approach that successfully integrates the environment, which is a major driving force behind natural selection, thus represents a valuable alternative to detect adaptive loci.

Some of the earliest examples of adaptation in natural populations come from observed concordances between phenotypic traits and environmental variation. Turesson (1922) was one of the first to consider the genotype as the relevant unit living in different habitats across the distribution of a species. Huxley (1938) reviewed several case studies of intraspecific variation in phenotypes across space. He coined the terms 'cline' to describe this phenomenon and 'ecocline' to describe the case where phenotypic variation is correlated with ecological factors. In recent years, with increasing

Box 1. Detecting signs of natural selection and genes involved in local adaptation

In the context of environmental, genetic, phenotypic and fitness variation, several approaches exist to uncover signs of natural selection and detect genes and environmental factors involved in local adaptation. The following simplified scheme presents some of these possibilities (modified from Sork *et al.* 2013). Boxes mark sources of variation that can be quantified, black arrows indicate the direction of the evolutionary process between cause and effect, and the grey arrow shows how selection acts on the different levels. Population genomics (reviewed in Hohenlohe *et al.* 2010b) and quantitative genetics (Stinchcombe & Hoekstra 2008) use genotypic and phenotypic information, respectively, alone to identify adaptive genetic variation. All other methods deal with the interaction of two of the different types of data. QTL (quantitative trait locus) mapping (Stinchcombe & Hoekstra 2008) and GWAS (genomewide association studies, Korte & Farlow 2013) are used to identify loci linked to specific phenotypes. Common garden and reciprocal transplant experiments (Savolainen *et al.* 2013) investigate the phenotypic and fitness differences of individuals originating from and living in different environments. Environmental association analysis (reviewed in this study) aims to correlate environment and genotypes. To our knowledge, only one methodological framework (Berg & Coop 2014) performs a joint analysis of all three aspects.



© 2015 John Wiley & Sons Ltd

availability of genetic data from diverse species, a popular approach seeks to identify genetic variants strongly associated with specific environmental conditions (see Mitton *et al.* 1977; for one of the earliest examples). This approach, referred to as environmental association analysis (EAA; Boxes 1 and 2) and also called genetic–environment analysis (e.g. Lotterhos & Whitlock 2015), has the potential to uncover patterns induced by adaptive processes that are not detected by traditional population genomic approaches, or to complement and support results of these. EAA is at the core of landscape genomics, an emerging research field that integrates tools from landscape genetics and population genomics to identify the environmental factors that have shaped present-day (adaptive) genetic variation and the gene variants that drive local adaptation (Holderegger *et al.* 2010; Manel *et al.* 2010a; Manel & Holderegger 2013; Sork *et al.* 2013). In practice, EAA is often used in concert with other population genomic tools such as outlier analysis (e.g. Fischer *et al.* 2013). It is thus difficult to draw a distinct line between these two approaches. As with many other areas of molecular ecology, the emergence of landscape genomics has been strongly facilitated by next-generation sequencing (NGS), which allows screening thousands to millions of single nucleotide polymorphisms (SNPs) across the entire genomes of many individuals and populations at reasonable costs. The data sets describing environmental characteristics (e.g. spatially explicit data on abiotic factors such as topography, climate, bedrock type, but also biotic factors such as dominant species or vegetation types) have also greatly improved and increasingly become publicly accessible, owing to versatile remote sensing techniques and database harmonization, respectively.

Numerous statistical methods for environmental association studies have recently been developed. However, no single widely accepted statistical approach has yet emerged. Accordingly, researchers often find it difficult to navigate the many possible avenues for EAA provided by recent innovation. Here, we present a practical guide to EAA, both for the landscape genomics community as well as for those freshly entering this research field. This article complements earlier conceptual reviews on landscape genomics (Holderegger *et al.* 2010; Manel *et al.* 2010a; Schoville *et al.* 2012; Joost *et al.* 2013; Manel & Holderegger 2013; Bragg *et al.* 2015) and comparisons of the statistical performance of selected methods (De Mita *et al.* 2013; Frichot *et al.* 2013; Jones *et al.* 2013; de Villemereuil *et al.* 2014; Lotterhos & Whitlock 2015) by focusing on the practical aspects of designing and analysing an environmental association study. First, we will introduce the basics of EAA by describing sampling designs and required data sets. Next, we present several methods, focusing on their optimal application, also referring to dedicated software packages and their specific properties. Subsequently, we touch on limitations and extensions of EAA and conclude by describing future directions and possible improvements in the field of landscape genomics. This review concentrates on SNPs as genetic markers, because they are currently the marker of choice and because they can often be functionally annotated. However, several environmental association methods can also be used with other, less commonly used marker types such as expressed sequence tag-derived simple sequence repeats (EST-SSR, e.g. Bradbury *et al.* 2013) or anonymous and dominant markers, such as amplified fragment length polymorphisms (AFLPs, e.g. Manel *et al.* 2012b).

Preparation of data

The basic goal of EAA is to test whether particular alleles or gene variants are significantly associated with any factor describing the environment in which they predominantly occur. For an environmental association study, two types of data are gathered, namely environmental factors and genetic polymorphisms, which should match in spatial resolution when establishing the sampling design. Processing environmental data includes data compilation (on-site measurement, data acquisition from existing sources), quality control, integration over time and/or space, and factor selection. Assessing genetic polymorphisms requires DNA extraction and sequencing or genotyping and is followed by bioinformatics, including quality control and data trimming. The two data components are then used in the actual EAA to assess evidence for allele–environment correlations. These steps are shown in Fig. 1 and detailed in the following sections.

Sampling design

When identifying sampling locations for an environmental association study, one intuitively thinks about sampling along environmental gradients. For instance, one could sample along a continental temperature or a local water salinity gradient. This design is appealing, but replication of gradients, also within evolutionary lineages, is important because multiple findings of the same candidate loci are a strong sign that they are true positives, and because replication reduces the confounding of population structure and covarying environmental factors. Usually, gradients of one particular environmental factor are the focus, but other environmental factors can be integrated into the analysis later. Another possibility is sampling in a categorical way, where researchers set up a 'quasi-experimental' design

© 2015 John Wiley & Sons Ltd

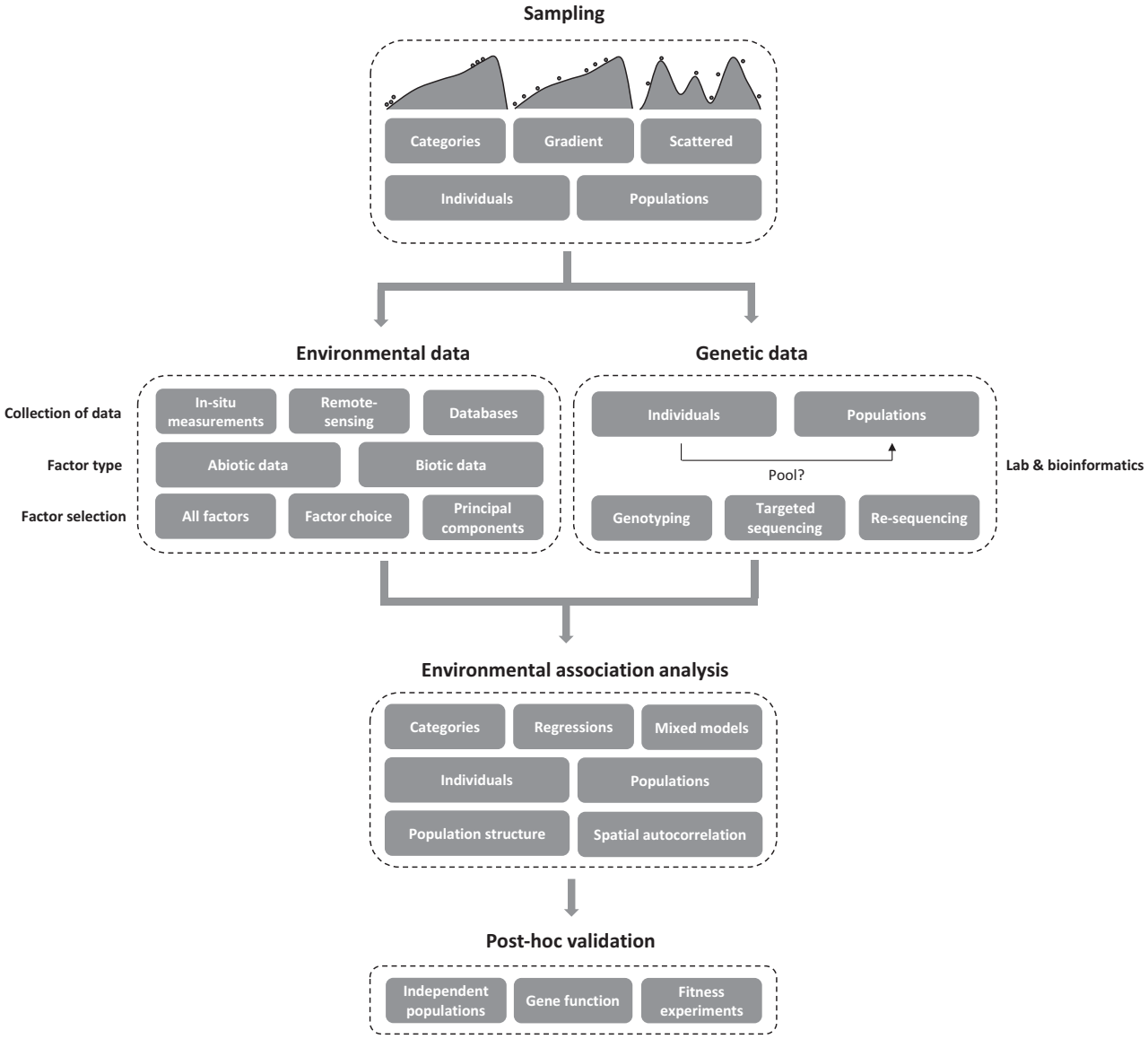


Fig. 1 A typical workflow in environmental association analysis (EAA). The three most important options per step are horizontally aligned. Genetic and environmental data are collected at the same sampling locations, processed separately and jointly analysed in EAA. The results can subsequently be validated with complementary approaches. All steps and options are described in detail in the manuscript.

with different 'treatments', for example low vs. high temperature or low vs. high salinity. Here, levels of a particular environmental factor are in focus. Categorical sampling seems attractive, but the range of subsequent adequate EAAs is limited, and one should clearly consider the number of replicates necessary for statistical significance testing. Researchers can also get a broad sample covering the entire environmental niche of a study species in a given area. Sampling locations would then be more widespread and scattered, or even randomly stratified (weighted random samples of representative subsets of sampling locations, Allaby 2009). This scattered sampling design leaves a lot of flexibility as a variety of different environmental association methods and environmental factors can potentially be used (Table 1), but it comes with the drawback of (often) lacking replication and clear hypotheses to test. In a review on sampling strategies in landscape genomics, Manel *et al*. (2012a) suggest to use model-based stratification and simulations to establish sampling designs (if sufficient biological and environmental knowledge is available) instead of applying classical ecological sampling designs like random sampling. The authors recommend choosing the climatic or biological

© 2015 John Wiley & Sons Ltd

Table 1 Overview of methods and software available for environmental association analysis in landscape genomics. Note that for some methods, other software or R packages are available

Method	Reference	Association type	Sampling design	Incorporation of neutral genetic structure	Incorporation of spatial autocorrelation	Individual/ population data	Mode for pooled data	Correction for sample size	Software/ R package
Categories		Categorical	Categorical	Possible	Possible	Both	Possible	Possible	Various statistical methods
Spatial analysis method (SAM)	Joost *et al.* (2007)	Logistic	Gradient/ scattered	Possible (in SAMβADA)	Possible (in SAMβADA)	Individual	No	No	SAM (Joost *et al.* 2008), SAMβADA (Stucki *et al.* submitted)
Multiple logistic regression		Logistic	Gradient/ scattered	Possible	Possible	Individual	No	No	R (R Development Core Team 2011)
Generalized estimating equations (GEEs)	Carl & Kuhn (2007), Poncet *et al.* (2010)	Logistic	Gradient/ scattered	No	Yes	Individual	No	No	GEEPACK (Yan & Fine 2004)
Partial Mantel test	Smouse *et al.* (1986)	Linear/ rank-linear	Gradient/ scattered	Yes	Possible	Both	No	No	ECODIST (Goslee & Urban 2007), VEGAN (Oksanen *et al.* 2013)
Multiple linear regression/ General linear models		Linear	Gradient/ scattered	Possible	Possible	Both	No	No	R (R Development Core Team 2011), TASSEL (Bradbury *et al.* 2007)
Canonical correlation analysis (CCA)	Legendre & Legendre (2012)	Linear	Gradient/ scattered	Possible	Possible	Both	No	No	VEGAN (Oksanen *et al.* 2013)
(Partial) redundancy analysis (RDA)	Legendre & Legendre (2012)	Linear	Gradient/ scattered	Possible	Possible	Both	No	No	VEGAN (Oksanen *et al.* 2013)

© 2015 John Wiley & Sons Ltd

© 2015 John Wiley & Sons Ltd

Table 1 *Continued*

Method	Reference	Association type	Sampling design	Incorporation of neutral genetic structure	Incorporation of spatial autocorrelation	Individual/ population data	Mode for pooled data	Correction for sample size	Software/ R package
BAYENV	Coop *et al.* (2010)	Linear/ rank-linear	Gradient/ scattered	Yes	No	Population	Yes (in BAYENV2)	Yes	BAYENV (Coop *et al.* 2010), BAYENV2 (Günther & Coop 2013)
Spatial generalized linear mixed model (SGLMM)	Guillot *et al.* (2014)	Linear	Gradient/ scattered	Yes	Yes	Both	No	Yes	GINLAND (Guillot *et al.* 2014)
Latent factor mixed models (LFMMs)	Frichot *et al.* (2013)	Linear	Gradient/ scattered	Yes	No	Both	No	No	LFMM (Frichot *et al.* 2013), LEA (Frichot & Francois 2015)
GWAS mixed models		Linear	Gradient/ scattered	Yes	No	Individual	No	No	EMMA (Kang *et al.* 2008), TASSEL (Bradbury *et al.* 2007), LME4 (Bates *et al.* 2014)
F_{ST}-based methods	de Villemereuil & Gaggiotti (in press)	Differentiation-based	Gradient/ scattered	Yes	No	Both	No	Yes	BAYESCENV (de Villemereuil & Gaggiotti in press)

space over topographic or geographic space when developing a stratified sampling design. Finally, an interesting approach suggested by Lotterhos & Whitlock (2015) is to sample scattered and random pairs of closely situated populations that exhibit substantial differences in environmental conditions while being within geneflow distance. These authors showed, using simulated data, that this sampling design has increased power in detecting true positives compared to random or transect designs, especially in models with weak selection. The reason for this is that the paired design maximizes the differences in adaptive environment while it minimizes the differences in neutral genetic structure. Importantly, landscape genomic studies should be performed over an appropriate geographic scale, which depends on the ecology of the organism (reviewed in, e.g. Anderson *et al.* 2010; Manel *et al.* 2010a; Richardson *et al.* 2014). A major issue is the mobility, dispersal capacity and migration rate of the species under study: for example, the relevant scale for mobile animals may be quite different to the scale for stationary plants. Moreover, researchers should be aware of potential mismatches in time between genomic and environmental data; there might be a time lag between the process causing the genetic pattern and the observed genetic response to it (Anderson *et al.* 2010).

Sampling can either be performed on the individual or population level. In studies that include only a single individual per sampling location, laboratory costs (but not costs for field sampling) are decreased, as only a comparatively low number of individuals has to be processed. Individual sampling limits the range of EAAs to individual-based approaches that can handle allele or locus genotype presences/absences or allele frequencies of 0/0.5/1 in the case of SNPs in a diploid species (Table 1 and Box 2, Figs C,E). In contrast, studies using population-based sampling can take advantage of population-based association approaches (Box 2, Figs A,B,D).

Environmental factors

Sources of environmental information. As many abiotic and biotic factors are potentially acting as selective pressures, it is crucial to consider those factors that are most pertinent for the question asked and most likely to provide high explanatory power. Because this knowledge is usually missing a priori, environmental association studies are often rather explorative.

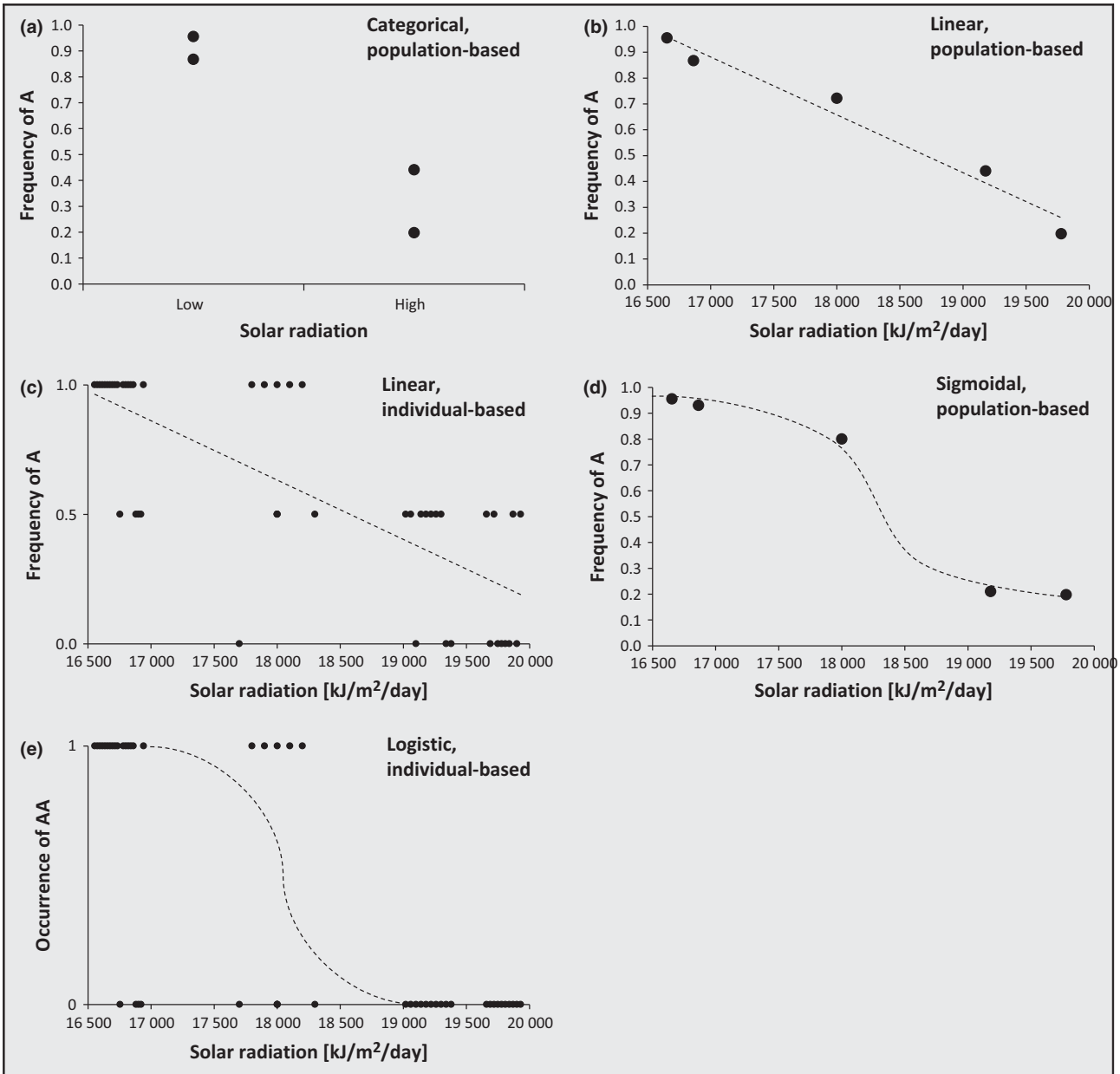
Abiotic data, in particular topo-climatic factors interpolated over large areas, are available from many public databases (Manel *et al.* 2010a; Thomassen *et al.* 2010). Limited to about 1-km^2 resolution is the ground-based interpolated WorldClim data (http://www.worldclim.org, Hijmans *et al.* 2005), where global climate layers for numerous factors, for recent as well as past and future periods, are freely available for analyses in a geographic information system (GIS) or in R (R Development Core Team 2011). Regional data sets based on ground-measured climate records, with higher resolution than the WorldClim data, are often available. Such climate data provide annual, seasonal, monthly or daily mean values as well as ranges and extremes. Increasingly, remote sensing supports data acquisition for large-scale environmental data, including elevation (http://glcf.umd.

Box 2. Response curves

The main goal of environmental association analysis (EAA) is to test whether a specific allele or locus genotype is associated with a specific environmental factor, while controlling for neutral genetic structure. However, depending on the genetic data available and the sampling design, different possibilities exist to detect such associations. Moreover, different response curves might be expected. This is illustrated by the following simplified examples. Imagine an adaptive SNP (locus X) with alleles A and G of a plant gene involved in response to light stimulus. In the simplest example, we sample four populations, two in each habitat with either low or high radiation. We genotype the locus in all individuals and calculate allele frequencies per population. If allele A at locus X is associated with low radiation, then we expect it to mainly occur in low-radiation populations, whereas allele G is mainly found in individuals of high-radiation populations (Fig. A). In the second case, we sampled five populations along a gradient in solar radiation. Here, an association (using linear regression) would look, for example, like in Fig. B, where the frequency of A in a population decreases when radiation increases. In a sampling design that includes scattered and geo-referenced single individuals from habitats with large differences in radiation, a significant association (using linear regression) should resemble Fig. C. Here, only three levels of allele frequencies (AA = 1, AG = 0.5 and GG = 0) are possible. If both alleles of locus X are mostly fixed for either high or low radiation, and intermediate frequencies are rare, we could expect a sigmoidal response of allele frequencies to the environmental gradient (Fig. D). Finally, in a logistic approach, one tests the association of the presence/absence of an allele or locus genotype, as, for example AA shown in Fig. E. Note that these simplified examples do not incorporate neutral genetic structure, which most of the more sophisticated environmental association methods do.

© 2015 John Wiley & Sons Ltd

Box 2 *Continued*



In essence, most EAA methods assume linear responses. Nevertheless, in the case of a selective sweep caused by strong directional selection (Barrett & Schluter 2008) leading to the fixation of alleles in two extreme environments, one might rather expect a sigmoidal (Fig. D) than a linear response curve (Fig. B) of allele frequencies along the environmental gradient, because the two alleles lead to clear-cut fitness differences in response to the environment (Richter-Boix *et al.* 2011). None of the currently used methods for EAA is able to deal with such sweep-like response curves of population allele frequencies. Logistic regression best meets the assumptions underlying a sigmoidal pattern, but has other drawbacks (see Future directions). However, Prunier *et al.* (2012) conceptualized adaptation along an environmental gradient in the following way. Each allele essentially displays a bell-shaped frequency curve on a part of the whole environmental gradient. In other words, it provides strong adaptation to and thus highest fitness at a particular position of the gradient. The additive effects of the bell-shaped frequency curves of all alleles together enable the species to live along the whole environmental gradient. If such additive effects are occurring, we expect a linear rather than a sigmoidal response curve between allele frequencies and environmental factors.

© 2015 John Wiley & Sons Ltd

edu), precipitation (http://trmm.gsfc.nasa.gov/data_dir/data.html) and vegetation indices (e.g. http://glcf.umd.edu, https://lpdaac.usgs.gov). The latter have been further developed through the use of light detection and ranging (LiDAR) technology, but such data are only available at regional (mostly national) scale so far. The primary limitations of current climate data sources are that they (i) often have a coarse spatial resolution, (ii) are integrated over a certain time period, (iii) represent spatial and temporal interpolations, and, consequently, (iv) ignore small-scale heterogeneity. Microscale conditions can therefore not be characterized in sufficient detail. Hence, researchers have started measuring data on site, for example by assessing soil conditions or using remote-sensing techniques (e.g. unmanned aerial vehicles, UAVs), but published examples are not available so far. While field-based measurements well reflect local site conditions in given years, they can fail in capturing long-term environmental conditions, which may often underlie adaptive response. An elegant, but restrictive way to make use of on-site measurements in EAA is to choose sampling locations where data recording has been performed over long periods.

For topographic data such as altitude, slope and aspect, detailed digital elevation or terrain models (DEMs/DTMs) are accessible at a worldwide scale and often at very high resolution (e.g. ASTER, http://asterweb.jpl.nasa.gov/gdem.asp). In this respect, new techniques, such as satellite- or plane-based LiDAR data assessment or UAVs help to improve the spatial resolution of DEMs to a few centimetres. From such high-resolution DEMs, microsite conditions may also be derived (Leempoel *et al.* in press). Furthermore, a wealth of other environmental data can possibly be considered, including geological factors, vegetation types, land cover, land use or species distributions, which might also serve as proxies for trophic interactions, prey availability or pathogen pressure (Gugerli *et al.* 2013).

Preparing environmental factors. A strategy that considers all environmental factors one-by-one provides a comprehensive analysis and reduces the risk of missing important loci and genes involved in local adaptation. However, using a large number of environmental factors rather than specific hypothesis tests increases the number of statistical tests, which needs to be considered in analyses to reduce the rate of false positives. In addition, many biotic and abiotic factors are highly correlated, for example altitude and temperature, or latitude and growing period. This leads to the testing of strongly interdependent models, increases variance in multivariate tests and makes estimates of the relative importance of different factors difficult. Including highly correlated factors may thus lead to the wrong conclusions if an understanding of the environmental drivers of local adaptation is desired.

One way to avoid collinearity is to reduce the number of factors (for a review of methods, see, e.g. Dormann *et al.* 2013). For example, Fischer *et al.* (2013) removed highly correlated factors (Pearson's $|r| \geq 0.8$) based on a pairwise correlation matrix. Another possibility is to select or remove highly correlated factors based on their contribution to the first few axes of a principal component analysis (PCA), keeping only those factors with the highest contribution to each axis (e.g. Manel *et al.* 2010b; Zulliger *et al.* 2013). A further option is to control for multicollinearity with the variance inflation factor, by iteratively removing the most highly correlated factor until the highest factor is lower than a certain threshold. Nevertheless, when reducing the number of factors, the removed factors should still be taken into consideration when interpreting the results. Imagine a sampling design with several alpine plant populations at high altitude. Solar radiation is removed in the process of factor reduction, because it shows a positive correlation with temperature (which is often the case at high altitude; Körner & Riedl 2012). If a gene known to be involved in response to radiation is associated with temperature, one might easily misclassify the selective pressure.

PCA (or related multivariate statistics) offers another possibility to condense a large number of environmental factors. This approach creates new synthetic environmental factors, consisting of groups of variables (e.g. Eckert *et al.* 2010a; Mosca *et al.* 2012; Nosil *et al.* 2012). While this simplification to a few derived factors makes statistical analysis easy, it can make the biological interpretation of the results difficult, notably if several factors strongly influence principal components. It is therefore only recommended to use PCA loadings as environmental factors when their interpretation is straightforward. PCA may also be problematic if the data show high levels of spatial autocorrelation (Thomassen *et al.* 2010).

Genomic data

Historically, after the use of isozymes (e.g. Mitton *et al.* 1977), dominant AFLPs were the marker of choice for EAA (Manel *et al.* 2010a), because they allowed testing up to hundreds of loci with a relatively simple and inexpensive laboratory protocol. The sequence of an AFLP marker and its flanking region, however, is commonly unknown unless additional sequencing efforts are made (e.g. Buehler *et al.* 2013; Zulliger *et al.* 2013). These anonymous markers have largely been replaced by SNPs, which are abundant across the entire genome, can easily be standardized

© 2015 John Wiley & Sons Ltd

among laboratories, and whose flanking sequences can be directly queried in public databases (Morin *et al.* 2004). In the coming years, whole-genome sequencing of all individuals will eventually become the standard in EAA and enable the association of millions of SNPs of known location and function. So far, we are aware of only one published environmental association study (Yoder *et al.* 2014) that used individually sequenced whole genomes. If such deep sequencing is not possible due to large sample sizes and genomes, researchers aiming for environmental association studies can reduce costs mainly by sequencing pooled samples (Pool-Seq) or by targeting a fraction of the genome (e.g. candidate-gene approach or genome complexity reduction). We detail these three options in the following sections.

Pool-Seq (reviewed in Schlötterer *et al.* 2014) is a cost-effective method of NGS, because the DNAs of several individuals are equimolarly pooled before sequencing (Futschik & Schlötterer 2010). This approach can lead to accurate SNP allele frequency estimates (reviewed in Rellstab *et al.* 2013) and population genomic parameters (Futschik & Schlötterer 2010; Schlötterer *et al.* 2014). As a drawback, individual multilocus genotypes and information on heterozygosity are inaccessible. As many environmental association approaches can handle population allele frequencies (Table 1), the use of whole-genome Pool-Seq is an attractive option, but only BAYENV2 (Günther & Coop 2013) yet accounts for the variance introduced by variation in sequencing coverage in Pool-Seq. Nevertheless, whole-genome Pool-Seq data have only rarely been used in EAA so far (but see Turner *et al.* 2010; Fabian *et al.* 2012; Fischer *et al.* 2013).

In a candidate-gene approach, genes or loci are characterized which have already been identified or known to potentially play an important role in local adaptation, or which are involved in a biological process related to the tested environmental factors. This is an especially appealing strategy for study species for which only limited genomic information is available. Information about biological processes can be retrieved, for example from homologous genes of species for which gene ontology (GO) databases exist (Primmer *et al.* 2013). For SNP genotyping, there are various high-throughput methods on the market (e.g. real-time PCR, KASP, Infinium, GoldenGate, pyrosequencing). Some genotyping technologies can also be used to accurately determine allele frequencies of population pools, for example Infinium (e.g. Bourret *et al.* 2013) or pyrosequencing (e.g. Gruber *et al.* 2002; Rellstab *et al.* 2011). To sequence genes or gene regions, targeted amplicon sequencing of individual or pooled samples using one of the NGS platforms is an attractive option (e.g. Homolka *et al.* 2012; Ho *et al.* 2014).

An alternative strategy to lower costs is complexity reduction of the genome. In exome capture (Bamshad *et al.* 2011), only the part of the genome is sequenced which hybridizes to probes covering exons. This approach requires at least partial knowledge about the transcriptome. In restriction-site associated DNA sequencing (RAD-Seq) and its variants (Puritz *et al.* 2014), the complexity of the genome is reduced using restriction enzymes, and the flanking regions of restriction sites are sequenced by NGS (Davey *et al.* 2011). This approach has successfully been applied to pooled population samples (Emerson *et al.* 2010). However, RAD-Seq identifies substantially fewer polymorphisms, from a few thousand to tens of thousands (e.g. Emerson *et al.* 2010; Hohenlohe *et al.* 2010a), as compared to millions of SNPs when using whole-genome Pool-Seq (e.g. Turner *et al.* 2010; Fabian *et al.* 2012; Fischer *et al.* 2013).

Incorporating neutral genetic structure

EAAs need to consider various types of autocorrelation, which arise from the mere historical relationships of individuals across the sites where they live. Consider two locations, where several individuals are sampled. The samples from the same location share a similar environment, which in turn is likely to differ from the other location. Likewise, individuals from one location tend to be more closely related to each other than to individuals from the second location. This concept can be expanded to any spatial scale and applies to both individual- and population-based sampling. If EAAs do not consider such dependencies, the identified associations might just be the consequence of the spatial arrangement and demographic history of the individuals or populations, and not a signature of local adaptation. It is therefore important to correct for neutral genetic structure in EAA. Alternatively (or additionally), some studies and methods (Table 1) include pure spatial autocorrelation in their approaches. Because spatial autocorrelation can serve as a proxy for neutral genetic structure, given isolation-by-distance patterns, a joint incorporation of both parameters (genetic and spatial structure) in such a situation is actually overly conservative. As spatio-environmental relationships are well covered in a recent review (Thomassen *et al.* 2010), we touch this issue only briefly and focus on how to deal with neutral genetic structure in EAA.

To account for the spatial signal in the data, one may just incorporate one or more spatial factors in regression-based models. A simple approach integrates either the

© 2015 John Wiley & Sons Ltd

geographic coordinates of, or the pairwise Euclidean distances between sampling locations into analysis (e.g. Guillot *et al.* 2014; Stucki *et al.* submitted). In a more elaborate strategy, Manel *et al.* (2010b) included Moran's eigenvector maps (MEMs, based on coordinates of the sampling locations, Borcard & Legendre 2002; Dray *et al.* 2006). MEMs represent environmental variation not specifically included in the model as well as pure spatial signals. Using generalized estimating equations (GEEs), Poncet *et al.* (2010) considered spatial autocorrelation of sampled individuals within populations. This concept assumes that individuals sampled within the same location share respective properties (habitat, kinship), whereas individuals sampled at any other site do not.

Neutral population genetic structure is defined as allele frequency differences among populations that have arisen due to neutral processes such as genetic drift, gene flow and mutation. The patterns of differences in allele frequencies among populations are the background against which loci contributing to local adaptation — a non-neutral process — are assessed in EAA. Neutral processes affect all loci across a genome, whereas non-neutral processes affect only a subset of loci. Corrections for neutral genetic structure are important in EAA, because neutral genetic structure can mimic patterns expected under non-neutral processes (Excoffier & Ray 2008; Excoffier *et al.* 2009; but see Vilhjalmsson & Nordborg 2013). For example, post-Pleistocene expansion by a species from a southern refugium may create clines of allele frequencies at neutral loci that are correlated with latitude, and any environmental factor related to latitude, resulting in false positives in EAA (but see Frichot *et al.* 2015). For instance, in *Picea sitchensis* along the western coast of North America (Holliday *et al.* 2010), demography created clines in allele frequencies that confounded tests of neutrality. Controlling for neutral genetic structure reduces the concern about this kind of false positives, because associations among SNPs and environmental factors are assessed after removing the confounding effects of neutral genetic structure (Sillanpää 2011).

Ideally, the subset of neutral markers used to estimate neutral genetic structure is known a priori. However, given that it is generally not possible to know which markers are neutral, a decision about how to best represent neutral genetic structure must be made. First, one can generate a large number of markers across the genome, and all these markers are used to estimate neutral genetic structure (e.g. Eckert *et al.* 2010a,b). This approach implicitly assumes that the number of loci affected by non-neutral processes in the data set is so small that their effects on global estimates of neutral genetic structure are negligible. Second, two sets of molecular markers can be created, where one set is used to estimate and control for neutral genetic structure and the other (often including all available markers) is used in EAA (e.g. Bourret *et al.* 2013). Typically, control markers are from sites in the genome thought to be neutral, such as nonoutliers, synonymous sites (coding for the same amino acid), or noncoding regions. They should be carefully matched against the focal loci with respect to heterozygosity, sample size, minor allele frequency, ascertainment scheme and location in the genome (e.g. in regions with similar levels of background selection, see Berg & Coop 2014; Tiffin & Ross-Ibarra 2014). Consequently, nuclear microsatellites are not the best choice for estimating neutral genetic structure in an EAA using SNPs, as they have very different properties (e.g. mutation rate, allelic diversity) than SNPs.

Traditional methods for estimating neutral genetic structure rely on estimating global or pairwise fixation indices among populations (see, e.g. Holsinger & Weir 2009). In EAA performed at the level of population allele frequencies, not only pairwise fixation indices (e.g. Fischer *et al.* 2013), but also population-specific fixation indices (*sensu* Foll & Gaggiotti 2006) can be used to control for neutral genetic structure. Another choice with which to describe population genetic structure in EAA is the estimation of kinship. Numerous estimators of kinship exist (Weir *et al.* 2006), which can yield substantially different results. Kinship is calculated in a pairwise fashion for all individuals in the data set and is used in subsequent analyses. Note, however, that association approaches using a kinship matrix were developed for GWAS of mostly inbred lines of model organisms. In natural populations, neutral genetic structure might substantially differ from these cases, eventually having unpredictable consequences on the kinship estimator. The use of kinship as an estimator for neutral genetic structure may therefore be inappropriate and remains to be tested. Other popular methods, at the level of individual samples, include matrix factorization methods (e.g. PCA, Patterson *et al.* 2006) and clustering algorithms like STRUCTURE (Pritchard *et al.* 2000). Matrix factorization methods produce scores for each individual on each synthetic component, which are used to control for neutral genetic structure in downstream analyses. In contrast, model-based clustering methods result in a Q-matrix, which describes the fraction of each individual's genome attributable to one of the inferred clusters, which is then used to control for neutral genetic structure in EAA.

Analysis of data

In the following, we introduce and discuss the most important and popular methods for EAA (for an over-

© 2015 John Wiley & Sons Ltd

view see Table 1 and Box 2), divided into five broadly defined categories. We recommend applying several environmental association approaches to compare results. This selection is not complete, there are further but less commonly applied methods described in the literature (see, e.g. Jones *et al.* 2013).

Testing categorical factors

Landscape genomics in its simplest form compares allele frequencies of individuals or populations from different types of environments (Box 2A), for example northern vs. southern or high- vs. low-altitude populations. In statistical terms, the different types of environment are introduced as categorical variables in parametric or nonparametric tests. Typically, a neutral genetic model is not implemented (but see, e.g. Foll *et al.* 2014), and all other environmental factors than the one defining the sampling design are ignored. The most prominent example for such an analysis comes from Turner *et al.* (2010), who performed Pool-Seq on four populations of *Arabidopsis lyrata*; two populations originated from serpentine and two from granitic soils. Across eight million SNPs, the authors detected several loci indicative of serpentine soil adaptation, because alleles at these loci were differentiated between soil types and were located in genes with functions associated with conditions characteristic of each soil type.

Logistic regressions

Logistic regressions test whether an environmental factor affects the presence or absence of an allele or single-locus genotype. Although mostly used for dominant markers such as AFLPs, which provide binomial information, logistic regression can also be applied to codominant markers such as SNPs. It is then necessary to prepare the data set in a format that describes the absence and presence of every allele or locus genotype. Because logistic regression can only take two states into account (the presence/absence of an allele or locus genotype), there is no clear way to deal with three or more genotypic states that occur in loci with heterozygous individuals. In this case, an EAA requires multiple analyses, two when using alleles and three when using single-locus genotypes in the case of a bi-allelic SNP. Sampling individuals from diverse habitats or along environmental gradients is ideally suited for this type of analysis.

The spatial analysis method (SAM; Joost *et al.* 2007) was the first implementation of logistic regression in EAA. This approach ignored neutral genetic structure, possibly leading to high false-positive rates under various demographic scenarios (De Mita *et al.* 2013; Frichot *et al.* 2013). Despite this, SAM has been intensively used in studies of local adaptation. For example, Quintela *et al.* (2014) combined SAM with the outlier locus detection approach BAYESCAN (Foll & Gaggiotti 2008) to identify AFLP markers and mitochondrial haplotypes associated with water temperature in the freshwater gastropod *Radix balthica*. Similarly, Nielsen *et al.* (2009) identified seven outlier SNPs that were related to temperature and/or salinity at spawning grounds of Atlantic cod (*Gadus morhua*).

Recently, an extended version of SAM, SAMβADA (Stucki *et al.* submitted; available on arXiv) was developed to overcome some of the limitations of SAM. The software now includes the possibility of multivariate analyses testing, enabling the introduction of neutral genetic structure as an additional factor. SAMβADA can further quantify the level of spatial autocorrelation of genotypes. According to tests performed by the authors, the software is substantially faster than BAYENV2 and LFMM with the univariate model (i.e. not including neutral genetic structure) and faster than BAYENV2 with a bivariate model. SAMβADA comes with a module that can split and remerge large data files. Hence, analyses can be run on different processors in parallel, potentially enabling genomewide analyses. Multiple logistic regressions to test several factors simultaneously including neutral genetic structure can also be performed in R using the generalized linear model function, as shown by Grivet *et al.* (2011) in a candidate-gene approach in two Mediterranean pine species. An alternative logistic approach is formalized in generalized estimating equations (GEEs, Carl & Kuhn 2007), an extension of generalized linear models with a logit-link and binomial error distribution that considers spatial autocorrelation within populations. It is an individual-based method best suited for sampling designs including many locations from a broad range of environmental conditions, and with a low number of samples per population. According to simulations, GEEs suffer from high false-positive rates under various demographic scenarios (De Mita *et al.* 2013).

Matrix correlations

In matrix correlations, one aims to test for correlation between matrices that express distances or dissimilarities between sampling units. A simple Mantel test estimates the strength of correlation (linear or rank linear) between two distance matrices (Mantel 1967) and computes a *P*-value for the correlation coefficient in a permutation procedure. As an extension, the partial Mantel test checks if there is a correlation between two distance matrices given a third matrix (Smouse *et al.* 1986). In EAA, partial Mantel tests can be used with individual or population data. The first matrix includes pairwise

© 2015 John Wiley & Sons Ltd

genetic distances or differentiation among individuals or populations at particular loci, the second matrix consists of environmental distances between sampling locations, and the third matrix can be used to control for genetic structure with neutral pairwise genetic distances. Hancock *et al.* (2011a) performed rank-linear partial Mantel tests using genomewide SNP data from Eurasian accessions of *Arabidopsis thaliana*, controlling for neutral genetic structure using a kinship matrix based on genomewide genetic variation. They found an enrichment of likely functional variants and could use the results to predict relative fitness in a common garden experiment. Fischer *et al.* (2013) used linear partial Mantel tests in their study of natural populations of *Arabidopsis halleri*, with pairwise whole-genome F_{ST} values of over 2 million SNPs as a measure of neutral genetic structure, to identify candidate SNPs for adaptation to five environmental factors.

The (partial) Mantel test has several nice features. For example, it can deal with distances and does not rely on any parametric assumptions. However, Mantel tests have been criticized (e.g. Oden & Sokal 1992; Guillot & Rousset 2013; but see Legendre & Fortin 2010). Guillot & Rousset (2013) showed that, if there is spatial autocorrelation in the two matrices, Mantel tests result in *P*-values that are not well calibrated, because the permutation procedure fails to produce a valid null hypothesis. One possible solution to overcome this problem is to ignore *P*-values and concentrate on effect sizes instead (i.e. the correlation coefficient *r*) when identifying top associations between loci and environmental factors. For example, Fischer *et al.* (2013) used the 99% quantile of 100 000 simulated *r*-values as a threshold for relevant environmental associations. Another solution is the use of the nonparametric extension of BAYENV2, which provides a robust alternative approach to (rank based) partial Mantel tests in cases where parametric assumptions are not met.

General linear models

General linear models are statistical models in which a response variable is modelled as a linear function of some set of explanatory variables. These models can account for neutral genetic structure and include statistical methods largely familiar to biologists.

Multiple linear regressions and univariate general linear models. Multiple linear regressions test linear effects of several environmental factors on population allele frequencies and thus enable including neutral genetic structure. For example, several studies (Manel *et al.* 2012b; Zulliger *et al.* 2013) investigated adaptive genetic variation for diverse alpine plant species and used multiple linear regressions including multiple environmental factors and MEMs to account for the effects of spatial structure and/or unobserved environmental variation. Both studies (Manel *et al.* 2012b; Zulliger *et al.* 2013) found that temperature and precipitation are the driving factors behind local adaptation in alpine plant species.

Some environmental association studies (e.g. Bradbury *et al.* 2013) have taken advantage of general linear models previously used in GWAS, in which the genotype is the explanatory variable and a phenotypic trait measure the response variable, while controlling for neutral genetic structure with a covariate, for example with the elements of the *Q*-matrix of STRUCTURE (Pritchard *et al.* 2000). In EAA, however, environment instead of phenotype is used as response variable. As the environment experienced by an organism is not caused by its genotype, this might seem conceptually counterintuitive. It is assumed, however, that environmental factors that are strongly correlated with heritable traits can replace them in statistical models. An example is illustrated by Eckert *et al.* (2009), who showed that a linear association between bud flush and mean annual temperature for Douglas fir (*Pseudotsuga menziesii*) can be described through an association of a SNP affecting bud flush with mean annual temperature. Such general linear models are implemented, for example in the software TASSEL (Bradbury *et al.* 2007) or can be performed using standard linear modelling in R.

Canonical correlations and multivariate linear regressions. The general linear model framework can be extended to models with multivariate response variables to account for the polygenic architecture of adaptive traits. The most popular method is canonical correlation analysis (CCA), which finds the linear combinations of two sets of variables – multiple loci and multiple environmental factors – that are maximally correlated (Legendre & Legendre 2012). The results are orthogonal sets of canonical variables that can be tested for significance. The loadings by loci and environmental factors indicate which loci respond to which environmental factors. However, users should be aware that strong patterns of multicollinearity could skew the results. Moreover, as CCA does not allow missing data, global deletion of samples or imputation of missing values is often required. Along this line, Mosca *et al.* (2012) used CCA to show how geographic factors shape the population genetic structure, based on several hundred SNPs, of four subalpine conifer tree species in the European Alps.

A useful approach to test hypotheses about specific environmental factors is redundancy analysis (RDA,

© 2015 John Wiley & Sons Ltd

Legendre & Legendre 2012). It allows for building and testing models of varying complexity, including those that condition results based on neutral genetic structure or spatial effects, referred to as partial RDA (pRDA). Significance of the model, each synthetic orthogonal axis and each explanatory variable can be tested using a permutation-based analysis of variance (Legendre & Legendre 2012). Lasky *et al.* (2012) used pRDA to assess correlations between multivariate climate and multivariate genetic variation in *A. thaliana* while controlling for spatial effects and identified putatively adaptive SNPs by looking at the contribution of each SNP to the first RDA axis. Using large sets of SNP loci, populations and environmental factors, Bourret *et al.* (2013) identified temperature and geological factors as drivers of local adaptation in Atlantic salmon (*Salmo salar*) with RDA. Many of the putatively adaptive genes showed growth-related functions.

Mixed effects models

The use of mixed effects models is powerful in EAA because they provide a unified statistical framework for controlling for the effects of neutral genetic structure. Here, allele frequencies of individuals or populations are treated as response variables, environmental factors are used as fixed factors, whereas neutral genetic structure is incorporated as a random factor. Approaches differ in how significance is tested, how neutral genetic structure is incorporated, and which type of genotype–environment association (linear/rank-linear/logistic) is assumed.

BAYENV. Coop *et al.* (2010) developed a Bayesian approach, BAYENV, to assess evidence for correlations between loci and environmental factors. For a given genetic variant, BAYENV tests whether a model that includes an environmental factor has an improved fit to the data compared to a null model that includes only neutral genetic structure, which is represented by a covariance matrix of estimated allele frequencies. BAYENV delivers Bayes factors for each locus–variable combination. One should note, however, that these factors may not be directly compared across environmental variables because of variable-specific value ranges. An advantage of BAYENV is that it allows for the incorporation of uncertainty of allele frequencies that arises from differences in sample sizes. It is not applicable to individual and scattered sampling designs. More recently, Günther & Coop (2013) published BAYENV2, which can be robustly applied to data from Pool-Seq and includes the option of nonparametric tests (Spearman rank correlation). Using Spearman rank correlation showed low detection power in two scenarios simulated by Lotterhos & Whitlock (2015). In cases where the data diverge from assumptions of linearity, however, the relative power of nonparametric tests should increase. Coop *et al.* (2010) emphasized that the fit of the null model may be imperfect, presumably due to complexities in demography that are not captured by the covariance matrix. Therefore, they suggested to additionally examine other evidence that the approach identifies true signals of selection, such as enrichment of likely functional variants (e.g. nonsynonymous substitutions) in the distribution tails of the resulting Bayes factors. A recent study by Blair *et al.* (2014) showed that the run-to-run variation of BAYENV (version 1) can be large. These authors thus advise to average Bayes factors among multiple runs to produce more stable and reliable results.

BAYENV was the first method specifically developed for EAA that controlled for neutral genetic structure. As a result, it has been used in several large-scale studies of candidate genes and for genomic data sets. Hancock *et al.* (2008) applied an early version of this approach to candidate loci for energy metabolism genotyped in a worldwide set of human populations. Subsequently, they used BAYENV with a human genomic data set to identify correlations using both continuous and categorical environmental factors (Hancock *et al.* 2010, 2011b). The studies identified enrichment of nonsynonymous SNPs, variants associated with disease traits and ecologically relevant sets of genes among the loci correlated with environmental factors. BAYENV has also been applied to studies of local adaptation in candidate genes in tree species, first by Eckert *et al.* (2010a) in loblolly pine (*Pinus taeda*) and later in different spruce (*Picea*) species (Chen *et al.* 2012; Prunier *et al.* 2012).

Using simulations, BAYENV was shown to detect a relatively low rate of false positives (De Mita *et al.* 2013) and to perform best under scenarios with weak hierarchical genetic structure (de Villemereuil *et al.* 2014). However, BAYENV is slow because it is computationally very intensive (De Mita *et al.* 2013; Stucki *et al.* submitted) and therefore less suited for analyses of a large number of genetic polymorphisms. A related method is GINLAND (Guillot *et al.* 2014), a spatial generalized mixed model (SGLMM) which uses a Markov chain Monte Carlo (MCMC)-free approach with shorter computing time. GINLAND also considers pure spatial autocorrelation based on a geographical distance matrix. To our knowledge, GINLAND has not yet been used in any empirical study.

Latent factor mixed models (LFMMs). In LFMMs (Frichot *et al.* 2013), neutral genetic structure is introduced as a random factor with the so-called latent factors, which

© 2015 John Wiley & Sons Ltd

are similar to principal components and calculated from all available markers. The advantage of this linear approach is that the effects of environmental factors and neutral genetic structure on allele frequencies are simultaneously estimated. Moreover, computing time is reasonably fast, making LFMM attractive for EAA with whole genomes or subsets of large random batches of SNPs in parallel. This approach surpasses the need for specifically formalizing neutral genetic structure, and it works without knowledge about which loci are putatively neutral, which is often not available in advance. LFMM computes *Z*-scores and *P*-values to quantify the strength of associations and which are also informative when compared among environmental factors. Before starting the final analysis, the number of latent factors (*K*) has to be chosen, either by an analysis of histograms of test *P*-values for different *K*-values (i.e. it should look similar to a uniform distribution), by performing a Tracy–Widom test on the eigenvalues of a PCA on the genetic data, or using programs such as STRUCTURE (Pritchard *et al.* 2000) to determine plausible values for *K*. As the stochastic algorithm of LFMM (MCMC) does not provide exact results, Frichot *et al.* (2013) recommend to perform multiple runs, use the median of the resulting *Z*-scores and adjust their *P*-values as described in the software manual. The software LFMM comes with two different interfaces, a graphical user interface and a command-line version. Only the latter can handle population allele frequencies. LFMM is therefore suited for both population based and scattered, individual-based sampling designs.

Frichot *et al.* (2013) found that LFMM has low rates of false positives and negatives and that it performs slightly better than BAYENV in detecting weak selection. de Villemereuil *et al.* (2014) showed that LFMM provides the best compromise between detection power and error rates in situations with complex hierarchical neutral genetic structure and polygenic selection. Finally, Lotterhos & Whitlock (2015) showed that LFMM is quite robust to a variety of sampling designs and underlying demographic models. LFMM has been used in several recent empirical studies. For example, Zueva *et al.* (2014) investigated pathogen- and environment-driven selection in populations of Atlantic salmon. They identified around 900 of the 4631 tested SNPs to be associated with one of the five environmental factors considered, including parasite-induced mortality as a measure for pathogen-driven selection. De Kort *et al.* (2014) found strong associations between temperature and 15 outlier SNPs in black alder (*Alnus glutinosa*) and showed, with additional evidence from a common garden experiment, that temperature is the main driver of local adaptation in this drought-sensitive tree species.

GWAS mixed models. Mixed models have been a standard approach for some time for the discovery of genotype–phenotype associations (Korte & Farlow 2013). As in the general linear models described above, environmental association studies have taken advantage of computationally efficient GWAS methods by replacing the response variable phenotype by environment. Kang *et al.* (2008) developed an efficient mixed-model association (EMMA) method that includes a simple identity-by-state allele sharing kinship matrix to control for neutral genetic background. EMMA was used to associate the RegMap panel SNPs (Horton *et al.* 2012) in *A. thaliana* to cold- and moisture-related climatic factors (Lasky *et al.* 2014). Genes with genetically variable expression responses to abiotic stress were enriched by SNPs strongly associated with climate. It is important to note that EMMA is optimized to test associations of only one allele with climate. Allowing heterozygous genotypes of outbred individuals is possible, but complex and computationally intensive (Kang *et al.* 2008). Moreover, the use of a kinship matrix to describe neutral genetic structure of populations may be inappropriate. Similarly, a linear mixed-model method is implemented in the software TASSEL (Bradbury *et al.* 2007). For example, Yoder *et al.* (2014) tested for associations of nearly 2 million SNPs to three climatic factors in 202 inbred accessions of barrel clover (*Medicago truncatula*). They identified more than 20 genes that were associated with climate and have a function in response to abiotic factors and pathogens in homologs of *A. thaliana*. GWAS mixed models are designed for individual rather than population sampling, making them best suited for analyses with samples continuously distributed across a study region.

Limitations and extensions of environmental association analysis

The main hurdle for EEAs (and notably also of population genomic approaches, De Mita *et al.* 2013; Lotterhos & Whitlock 2014) is that they might result in high rates of false positives (De Mita *et al.* 2013; Lotterhos & Whitlock 2014; de Villemereuil *et al.* 2014; Frichot *et al.* 2015), which are significant associations that are actually not casual. The main reason is that geographic and demographic processes can lead to patterns that mimic those observed as a consequence of selection. In fact, de Villemereuil *et al.* (2014) found high rates of false discovery in some scenarios with complex, hierarchical structure and polygenic selection. Fortunately, applying analyses that control for neutral genetic structure can mitigate this problem. De Mita *et al.* (2013) simulated different demographic, selective and mating type scenarios and found false-positive rates of up to 40% (logistic regres-

© 2015 John Wiley & Sons Ltd

sion) and 50% (GEE) for approaches not specifically correcting for neutral genetic structure, but only 20% for BAYENV, which corrects for structure. Depending on the combination of approach and scenario, power and error rates differed greatly in this study. Similarly, Frichot *et al.* (2013) reported low false-positive rates (0–7%) for methods that correct for neutral genetic structure. Unfortunately, some demographic scenarios may be particularly challenging for EAA. For example, scenarios in which the range expansion of a species creates a cline in allele frequencies along an environmental gradient (Keller *et al.* 2009; Novembre & Di Rienzo 2009) or in which individuals/populations are under strong isolation by distance (Lotterhos & Whitlock 2015) are hard to deal with in EAA (but see, Frichot *et al.* 2015). False positives can also arise due to the failure to account for multiple testing, which is needed when a large number of loci and environmental factors are included in the analysis. We strongly recommend to control for false-discovery rate (FDR) using the algorithms described by Benjamini & Hochberg (1995) and Storey & Tibshirani (2003). FDR (unlike, e.g. classical Bonferroni correction) does not depend on the number of tests and aims to accurately estimate the proportion of false discoveries among positive findings. A third cause of false positives is that it can be difficult to distinguish between correlated environmental selective pressures. More specifically, observed correlations with a specific environmental factor can be due to adaptation to covarying factors that were not included in the analyses or excluded in the process of factor reduction. In these cases, it is the association, not the locus, that represents a false positive. In other words, the detected locus might actually play a role in local adaptation, but is linked to a different factor. For example, the presence of an allele may be correlated with high temperature, but is actually involved in defence against pathogens whose development, survival and transmission is sensitive to temperature (Harvell *et al.* 2002). Moreover, correlations among loci (i.e. linkage disequilibrium between an adaptive locus and other variants) can result in a spurious signal of correlation at linked variants (hitchhiking, Strasburg *et al.* 2012). Finally, false positives can also derive from coincidental outlier values of environmental factors and allele frequencies. A simple way to deal with these cases is to avoid populations with extreme environmental values already in the sampling design, or to use rank-based, nonparametric statistics such as BAYENV2 or rank-linear partial Mantel tests. In any case, landscape genomic studies should carefully consider the issue of false positives, keeping in mind that applying stricter thresholds to possibly account for this issue will result in lower power to detect true positives and will inflate the rate of false negatives.

As for most biological studies, the results of EAAs are restricted to the sampled populations and environmental conditions. Therefore, several studies (e.g. Poncet *et al.* 2010; Prunier *et al.* 2012; Buehler *et al.* 2013) have considered geographical subsets that were analysed separately to detect more general patterns. Overlap among identified loci of adaptive relevance of such population subsets is, however, often minimal. For example, Poncet *et al.* (2010) found 61 and 21 climate-related AFLP loci in populations of the alpine rockcress (*Arabis alpina*) from the French and Swiss Alps, respectively. Only four of these loci were found in both regions. This result implies the presence of false positives (in the case of the SNPs that were only identified in one region) or to geographically restricted patterns of adaptation.

Combined approaches and downstream analyses

Given the issues discussed in the preceding section, it is desirable to combine EAA with other approaches in order to reduce the rate of false positives and to assess the relevance of findings. In this section, we list a selection of such integrative approaches (for more ideas, see, e.g. Pardo-Diaz *et al.* 2015) and exemplify them with respective empirical studies.

Combination with tests for outlier locus detection. Instead of opposing EAA and outlier detection methods, one could combine them to obtain more information from the data. For example, one could first perform an outlier test using, for example BAYESCAN (Foll & Gaggiotti 2008), FDIST and derivates (Beaumont & Nichols 1996; Beaumont & Balding 2004), FLK (Bonhomme *et al.* 2010), or ARLEQUIN (Excoffier & Lischer 2010) and use only the resulting outlier loci in subsequent EAA. For example, Fischer *et al.* (2013) used POPOOLATION (Kofler *et al.* 2011) to select the most extremely differentiated SNPs of *A. halleri* and subsequently correlated the resulting outlier loci to topo-climatic factors using partial Mantel tests. Selection processes that lead to small shifts in allele frequencies, however, are not likely to be detected with this strategy, and the overlap among different methods can be small (de Villemereuil *et al.* 2014). Alternatively, one could perform multiple analyses in parallel using the entire set of loci, and then discuss the results by comparing the two lists of putatively adaptive loci (e.g. Quintela *et al.* 2014). Finally, in EAAs using a categorical sampling design, one could perform outlier tests among groups of individuals that are defined by the environment (e.g. Buehler *et al.* 2013; Roda *et al.* 2013), while appropriately dealing with neutral genetic structure. Buehler *et al.* (2013) used DFDIST (Beaumont & Balding 2004) in *A. alpina* to identify one

© 2015 John Wiley & Sons Ltd

outlier AFLP marker that exhibited particularly high genetic differentiation among three contrasting habitat types. Foll *et al.* (2014) recently presented a flexible hierarchical extension of the BAYESCAN approach (Foll & Gaggiotti 2008), which allows for the simultaneous analysis of populations living in different environments in several distinct regions. It includes a convergent (parallel) evolution model that directly identifies candidate loci in replicated pairs of populations instead of using intersecting sets of candidate loci.

Gene function and gene ontology analyses. Recent technological and scientific advances have not only resulted in the availability of reference genomes for numerous species, but also led to the establishment of public databases where annotated genes are described in detail. For several model species, large parts of their genomes are now annotated, although not with the same level of reliability (Primmer *et al.* 2013). Most studies on evolutionary and molecular ecology, however, focus on nonmodel species. While draft genomes for nonmodel species are emerging (Ekblom & Galindo 2011), they still often lack annotation (Primmer *et al.* 2013). Fortunately, in most cases, annotation from related model organisms can be transferred to less well-studied species by identifying homologous sequences, assuming that they have the same function in both model and study species.

Gene ontology (GO) databases describe the biological process, molecular function and cellular component of a gene in a standardized, species-neutral vocabulary (Primmer *et al.* 2013). They therefore enable linking EAA with gene function. Many EAA studies rely on GO databases in one or the other way, not only in the planning phase (e.g. for selecting candidate loci), but also for downstream analyses. In most cases, researchers try to verify the biological function of a gene post hoc. In the best case, gene function appears reasonable in the context of the associated environmental factor (e.g. Eckert *et al.* 2009). This inference increases evidence that a given association is not purely coincidental. An additional option for EAA are GO enrichment tests (e.g. Fischer *et al.* 2013), which examine whether certain gene functions are over- or under-represented in a set of genes (e.g. those associated with an environmental variable).

Nonsynonymous vs. synonymous substitutions. Not all nucleotide substitutions lead to changes in the encoded amino acid. Usually, the third nucleotide of a codon is silent (synonymous, i.e. the derived codon codes for the same amino acid) and therefore thought to be selectively neutral. Annotation of investigated polymorphisms can therefore be applied to interpret the results obtained from EAA. This is only feasible if a reference genome of the investigated or a closely related species is available. The occurrence of nonsynonymous (amino acid changing) SNPs, especially if it also concerns SNPs significantly related to environmental factors, can increase evidence for relevance in adaptation. If many substitutions are present, one can calculate the ratio of nonsynonymous to synonymous variants within the distribution tail of the EAA and compare this to the ratio in nonsignificant loci. For example, Hancock *et al.* (2011a) looked at the top 1% of SNPs associated with climate in *A. thaliana* and found an enrichment of nonsynonymous compared to synonymous and nongenic substitutions.

Post hoc validation in independent data sets. Replicated patterns of local adaptation can derive from the spread of an adaptive allele to multiple geographic locations or by repeated and parallel adaptation (discussed, e.g. in Schmidt *et al.* 2008; Nosil *et al.* 2009; Prunier *et al.* 2012; Buehler *et al.* 2014; Tiffin & Ross-Ibarra 2014). However, studies using an independent data set to test the generality of adaptive loci are rare. Buehler *et al.* (2014), using 30 independent populations of *A. alpina,* did not find the same association of an AFLP outlier locus as identified previously (Buehler *et al.* 2013). In contrast, 15 previously identified AFLP loci of the gastropod *Littorina saxatilis* exhibiting signs of selection were distributed in the same clinal manner on two independent shores along the Atlantic coast in England (Grahame *et al.* 2006). Although such a validation step represents a useful addition to EAA, successful validation in an independent data set is not necessarily expected. This is because locus-specific selection is crucially dependent on the local genomic context and local environmental conditions, and genotype-by-environment interactions may modulate selection patterns in an unpredictable way (Schmidt *et al.* 2008), leading to geographically restricted local adaptation. However, finding recurrent patters in independent data sets greatly improves evidence for the generality of adaptive patterns detected.

Experimental validation. Direct proof that a genetic variant actually leads to a fitness advantage in a local environment can only be obtained experimentally (Barrett & Hoekstra 2011; Savolainen *et al.* 2013). Compelling support for EAA (or GWAS) findings is to employ a common garden experiment, in which genotyped individuals coming from different habitats share the same natural or manipulated environment(s) and are measured for fitness-related phenotypic traits (e.g. Fournier-Level *et al.* 2011; Hancock *et al.* 2011a; De Kort *et al.* 2014; Yoder *et al.* 2014). To this end, Hancock *et al.* (2011a) identified climate associations in *A. thaliana*

© 2015 John Wiley & Sons Ltd

accessions from across Eurasia and found that the identified SNPs could be used to predict rank fitness in a common garden. Conversely, Fournier-Level *et al.* (2011) grew hundreds of inbred *A. thaliana* lines derived from natural populations across their native distribution and planted them in four European field sites (common gardens) that spanned the species' native range. Alleles that were associated with higher fitness in particular common gardens were more frequent in the respective environment the plant originated from. In theory, only in reciprocal transplant experiments, it is possible to test whether the fitness of 'home' populations is actually higher than that of 'away' populations (Kawecki & Ebert 2004). Although reciprocal transplant experiments have been carried out repeatedly in the past (e.g. see Savolainen *et al.* 2013), they have mostly been conducted at the phenotypic level and have rarely taken advantage of genomic information. In the context of EAA, reciprocal transplant experiments are the perfect addition to check for fitness advantages of given alleles associated with particular environments. We are not aware of a study that has validated identified associations with this often laborious approach. While transplant and common garden experiments with genetic variants might be feasible in the case of processes of monogenic adaptation, they could be challenging for polygenic adaptation. One should also bear in mind that the potentially different genetic backgrounds of populations included in experiments can interfere with the detection of the adaptive signal (Holderegger *et al.* 2008). Finally, it should be noted that even if a fitness advantage is not detected in the above-described experiments, it does not mean it does not exist, as the results and interpretation of the experiment is bound to the experimental conditions (site conditions, duration, age of individuals, interactions with pathogens, etc.) and the measure of fitness.

Future directions

Options for further developing the field of EAA are manifold and involve theoretical, methodological and statistical issues, some of which we highlight in this section. One of the major challenges in EAA is the computational bottleneck resulting from increasingly large data sets, made possible by decreasing sequencing costs. Effort therefore will likely be put into developing faster algorithms that adequately control for neutral genetic structure.

Another issue to be improved in EAA is the lack of methods that can deal with nonlinear responses at the level of population allele frequencies. In a scenario that does not assume gradual changes in allele frequencies along an environmental gradient, but in contrast rather corresponds to a sweep model where beneficial alleles are mostly fixed, one would expect extreme allele frequencies. Accordingly, only a narrow range of environmental conditions should exhibit intermediate allele frequencies (Box 2D). Therefore, we suggest that effort should be put into the development of alternative population-based models. SAM and SAMβADA, which use logistic regression with a sigmoid-like response curve, deal with individual-based presence/absence data and not allele frequencies per population. Hence, they are not applicable to Pool-Seq. In addition, the interpretation of loci with heterozygous individuals is difficult. One possibility could be sigmoidal curve fitting with the challenge of incorporating neutral genetic structure. A promising development in this respect is integrated in BAYESCENV (de Villemereuil & Gaggiotti in press), which extends the outlier detection software BAYESCAN (Foll & Gaggiotti 2008) by implementing an additional model that includes information about the environment. It is based on genetic and environmental distance and can also detect patterns of allele frequency that are not linearly dependent on environmental factors. Along the same line, although one of the main arguments for using EAA is the detection of polygenic adaptation, most methods presented here only test single-locus effects. There is a clear lack of approaches that test associations with multivariate response variables (Sork *et al.* 2013), with the exception of multivariate analyses like CCA and RDA described above.

Analogs to the GO databases, which mostly stem from a cellular perspective, have recently been advocated based on an ecological association ontology that includes findings from ecological and evolutionary studies (Landry & Aubin-Horth 2007; Pavey *et al.* 2012). This so-called ecological gene annotation would complement existing GOs, introduce a vocabulary used by evolutionary ecologists, present links to ecological and evolutionary studies and decrease the proportion of genes with unknown function. Landscape genomics could nicely contribute to, and also greatly profit from such an additional source of information (for discussion, see also Primmer *et al.* 2013).

Finally, there is need for more post hoc validation regarding function and adaptive generality of the alleles, loci or genomic regions identified in EAA. Many studies, including most of those described in this review, perform EAA, present a list of interesting loci, compare it to GO databases and stop there, that is half way to the goal of identifying those genes that are functionally involved in local adaptation of natural populations. Instead, studies should go further and test their findings using, for example independent populations, knockout mutants, common garden and reciprocal

© 2015 John Wiley & Sons Ltd

transplant experiments. The effort of such follow-ups should, however, not be underestimated.

Conclusions

The recent advances in sequencing technologies have opened the door for analyses of the genetic basis underlying local adaptation. Landscape genomics is an emerging research area, focused on understanding the role of the environment in genetic adaptation and identifying putatively adaptive loci. EAAs are the main tools of this research field. In this review, we described several strategies to test for associations between genotypes and environment. We strongly recommend controlling for neutral genetic structure in the analyses, carefully applying and comparing complementary statistical approaches, using a large number of populations and/or individuals, and looking at as large as possible fraction of the genome. It is clear that for many researchers, especially for those with limited financial budgets or those working on nonmodel organisms, some of these goals are hard to reach. Luckily, sophisticated and increasingly inexpensive genomic technologies are emerging. It is important to note that landscape genomics, however, has yet several relevant issues to solve. Besides the potential for further extending current statistical models for EAA, the main concern is the lack of post hoc testing of fitness advantages of putatively beneficial genetic variants in specific environments. After the gold rush in finding loci that are associated with important environmental drivers, it is now time for the field to take the next step and show that promising candidates are linked to fitness-related variation and thus relevant for adaptation.

Acknowledgements

We thank Eric Frichot, Torsten Günther, Sylvie Stucki, Gilles Guillot, Dorena Nagel for practical advice when testing methods, Martin Fischer, Alex Widmer, Kentaro Shimizu for ideas and discussion, and four anonymous reviewers for valuable comments on earlier versions of this manuscript. This study was supported by the Swiss National Science Foundation (projects CRSI33_127155 to RH, 31003A_152664/1 and CR32I3_149741/1 to FG), a Marie Curie Grant (PCIG10-GA-2011-304301) and an MFPL Vienna International Postdoctoral Fellowship to AMH, and the Federal Office for the Environment (project QuercAdapt to FG).

References

Allaby M (2009) *Dictionary of Zoology*. Oxford University Press, Oxford.

Anderson CD, Epperson BK, Fortin M-J *et al.* (2010) Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**, 3565–3575.

Bamshad MJ, Ng SB, Bigham AW *et al.* (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, **12**, 745–755.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature Reviews Genetics*, **12**, 767–780.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.

Bates D, Maechler M, Bolker B, Walker S (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7. http://CRAN.R-project.org/package=lme4

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B-Biological Sciences*, **263**, 1619–1626.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289–300.

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. *PLoS Genetics*, **10**, e1004412.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Human Genomics*, **8**, 1. doi: 10.1186/1479-7364-8-1.

Bonhomme M, Chevalet C, Servin B *et al.* (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics*, **186**, 241–262.

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Bourret V, Dionne M, Kent MP, Lien S, Bernatchez L (2013) Landscape genomics in Atlantic salmon (*Salmo salar*): searching for gene-environment interactions driving local adaptation. *Evolution*, **67**, 3469–3487.

Bradbury PJ, Zhang Z, Kroon DE *et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Bradbury D, Smithson A, Krauss SL (2013) Signatures of diversifying selection at EST-SSR loci and association with climate in natural *Eucalyptus* populations. *Molecular Ecology*, **22**, 5112–5129.

Bragg JG, Supple MA, Andrew RL, Borevitz JO (2015) Genomic variation across landscapes: insights and applications. *New Phytologist*, **207**, 953–967.

Buehler D, Poncet BN, Holderegger R *et al.* (2013) An outlier locus relevant in habitat-mediated selection in an alpine plant across independent regional replicates. *Evolutionary Ecology*, **27**, 285–300.

Buehler D, Holderegger R, Brodbeck S, Schnyder E, Gugerli F (2014) Validation of outlier loci through replication in independent data sets: a test on *Arabis alpina*. *Ecology and Evolution*, **4**, 4296–4306.

Carl G, Kuhn I (2007) Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modelling*, **207**, 159–170.

Chen J, Kallman T, Ma X *et al.* (2012) Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce *(Picea abies)*. *Genetics*, **191**, 865–881.

© 2015 John Wiley & Sons Ltd

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

De Kort H, Vandepitte K, Bruun HH *et al.* (2014) Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. *Molecular Ecology*, **23**, 4709–4721.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Dormann CF, Elith J, Bacher S *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, **196**, 483–493.

Eckert AJ, Bower AD, Wegrzyn JL *et al.* (2009) Association genetics of coastal Douglas fir (*Pseudotsuga menziesu* var. *menziesii*, Pinaceae) I. Cold-hardiness related traits. *Genetics*, **182**, 1289–1302.

Eckert AJ, Bower AD, Gonzalez-Martinez SC *et al.* (2010a) Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology*, **19**, 3789–3805.

Eckert AJ, van Heerwaarden J, Wegrzyn JL *et al.* (2010b) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969–982.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*, **23**, 347–351.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Fabian DK, Kapun M, Nolte V *et al.* (2012) Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology*, **21**, 4748–4769.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Molecular Ecology*, **22**, 5594–5607.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, **174**, 875–891.

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

Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L (2014) Widespread signals of convergent adaptation to high altitude in Asia and America. *American Journal of Human Genetics*, **95**, 394–407.

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **333**, 86–89.

Frichot E, Francois O (2015) LEA: An R Package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Schoville SD, Bouchard G, Francois O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, Francois O (2015) Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity*, **115**, 22–28.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1–19.

Grahame JW, Wilding CS, Butlin RK (2006) Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. *Evolution*, **60**, 268–278.

Grivet D, Sebastiani F, Alia R *et al.* (2011) Molecular footprints of local adaptation in two Mediterranean conifers. *Molecular Biology and Evolution*, **28**, 101–116.

Gruber J, Colligan P, Wolford J (2002) Estimation of single nucleotide polymorphism allele frequency in DNA pools by using pyrosequencing. *Human Genetics*, **110**, 395–401.

Gugerli F, Brandl R, Castagneyrol B *et al.* (2013) Community genetics in the time of next-generation molecular technologies. *Molecular Ecology*, **22**, 3198–3207.

Guillot G, Rousset F (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution*, **4**, 336–344.

Guillot G, Vitalis R, Rouzic Al, Gautier M (2014) Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics*, **8**, 145–155.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Hancock AM, Witonsky DB, Gordon AS *et al.* (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, e32.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A (2010) Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 2459–2468.

Hancock AM, Brachi B, Faure N *et al.* (2011a) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **333**, 83–86.

Hancock AM, Witonsky DB, Alkorta-Aranburu G *et al.* (2011b) Adaptations to climate-mediated selective pressures in humans. *PLoS Genetics*, **7**, e1001375.

Harvell CD, Mitchell CE, Ward JR *et al.* (2002) Climate warming and disease risks for terrestrial and marine biota. *Science*, **296**, 2158–2162.

© 2015 John Wiley & Sons Ltd

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Ho T, Cardle L, Xu X *et al.* (2014) Genome-Tagged Amplification (GTA): a PCR-based method to prepare sample-tagged amplicons from hundreds of individuals for next generation sequencing. *Molecular Breeding*, **34**, 977–988.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010a) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Phillips PC, Cresko WA (2010b) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences*, **171**, 1059–1071.

Holderegger R, Herrmann D, Poncet B *et al.* (2008) Land ahead: using genome scans to identify molecular markers of adaptive relevance. *Plant Ecology & Diversity*, **1**, 273–283.

Holderegger R, Buehler D, Gugerli F, Manel S (2010) Landscape genetics of plants. *Trends in Plant Science*, **15**, 675–683.

Holliday JA, Yuen M, Ritland K, Aitken SN (2010) Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, **19**, 3857–3864.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Homolka A, Eder T, Kopecky D *et al.* (2012) Allele discovery of ten candidate drought-response genes in Austrian oak using a systematically informatics approach based on 454 amplicon sequencing. *BMC Research Notes*, **5**, 175.

Horton MW, Hancock AM, Huang YS *et al.* (2012) Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics*, **44**, 212–216.

Huxley J (1938) Clines: an auxiliary taxonomic principle. *Nature*, **142**, 219–220.

Jones MR, Forester BR, Teufel AI *et al.* (2013) Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution*, **67**, 3455–3468.

Joost S, Bonin A, Bruford MW *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources*, **8**, 957–960.

Joost S, Vuilleumier S, Jensen JD *et al.* (2013) Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Molecular Ecology*, **22**, 3659–3665.

Kang HM, Zaitlen NA, Wade CM *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kawecki TJ, Ebert D (2004) Conceptual issues in local adaptation. *Ecology Letters*, **7**, 1225–1241.

Keller SR, Sowell DR, Neiman M, Wolfe LM, Taylor DR (2009) Adaptation and colonization history affect the evolution of clines in two introduced species. *New Phytologist*, **183**, 678–690.

Kofler R, Pandey RV, Schloetterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**, 3435–3436.

Körner C, Riedl S (2012) *Alpine Treelines: Functional Ecology of the Global High Elevation Tree Limits*. Springer, Berlin.

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, **9**, 29.

Landry CR, Aubin-Horth N (2007) Ecological annotation of genes and genomes through ecological genomics. *Molecular Ecology*, **16**, 4419–4421.

Lasky JR, Des Marais DL, McKay JK *et al.* (2012) Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Molecular Ecology*, **21**, 5512–5529.

Lasky JR, Des Marais DL, Lowry DB *et al.* (2014) Natural variation in abiotic stress responsive gene expression and local adaptation to climate in *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **31**, 2283–2296.

Leempoel K, Parisod C, Geiser C *et al.* (in press) Very high resolution digital elevation models: are multi-scale derived variables ecologically relevant? *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.12427.

Legendre P, Fortin MJ (2010) Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, **10**, 831–844.

Legendre P, Legendre L (2012) *Numerical Ecology*. Elsevier, Amsterdam.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Manel S, Holderegger R (2013) Ten years of landscape genetics. *Trends in Ecology & Evolution*, **28**, 614–621.

Manel S, Joost S, Epperson BK *et al.* (2010a) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010b) Common factors drive adaptive genetic variation at different spatial scales in *Arabis alpina*. *Molecular Ecology*, **19**, 3824–3835.

Manel S, Albert CH, Yoccoz NG (2012a) Sampling in landscape genomics. In: *Data Production and Analysis in Population Genomics* (eds Pompanon F, Bonin A), pp. 3–12. Humana Press, New York.

Manel S, Gugerli F, Thuiller W *et al.* (2012b) Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology*, **21**, 3729–3738.

Mantel N (1967) Detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Mitton JB, Linhart YB, Hamrick JL, Beckman JS (1977) Observations on genetic structure and mating system of ponderosa pine in Colorado front range. *Theoretical and Applied Genetics*, **51**, 5–13.

© 2015 John Wiley & Sons Ltd

Morin PA, Luikart G, Wayne RK, the SNP workshop group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208–216.

Mosca E, Eckert AJ, Di Pierro EA *et al.* (2012) The geographical and environmental determinants of genetic diversity for four alpine conifers of the European Alps. *Molecular Ecology*, **21**, 5530–5545.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Nosil P, Gompert Z, Farkas TE *et al.* (2012) Genomic consequences of multiple speciation processes in a stick insect. *Proceedings of the Royal Society B-Biological Sciences*, **279**, 5058–5065.

Novembre J, Di Rienzo A (2009) Spatial patterns of variation due to natural selection in humans. *Nature Reviews Genetics*, **10**, 745–755.

Oden NL, Sokal RR (1992) An investigation of three-matrix permutation tests. *Journal of Classification*, **9**, 275–290.

Oksanen J, Blanchet FG, Kindt R *et al.* (2013) Vegan: community ecology package. R package version 2.0-8. http://CRAN.R-project.org/package=vegan

Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci of adaptive evolution. *Methods in Ecology and Evolution*, **6**, 445–464.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology & Evolution*, **27**, 673–678.

Poncet BN, Herrmann D, Gugerli F *et al.* (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology*, **19**, 2896–2907.

Primmer CR, Papakostas S, Leder EH, Davis MJ, Ragan MA (2013) Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Molecular Ecology*, **22**, 3216–3241.

Pritchard JK, Di Rienzo A (2010) Adaptation — not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Prunier J, Gerardi S, Laroche J, Beaulieu J, Bousquet J (2012) Parallel and lineage-specific molecular adaptation to climate in boreal black spruce. *Molecular Ecology*, **21**, 4270–4286.

Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

Quintela M, Johansson MP, Kristjansson BK, Barreiro R, Laurila A (2014) AFLPs and mitochondrial haplotypes reveal local adaptation to extreme thermal environments in a freshwater gastropod. *PLoS One*, **9**, e101821.

R Development Core Team (2011) R: a language and environment for statistical computing. http://www.R-project.org

Rellstab C, Louhi K-R, Karvonen A, Jokela J (2011) Analysis of trematode parasite communities in fish eye lenses by pyrosequencing of naturally pooled DNA. *Infection, Genetics and Evolution*, **11**, 1276–1286.

Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS One*, **8**, e80422.

Richardson JL, Urban MC, Bolnick DI, Skelly DK (2014) Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology & Evolution*, **29**, 165–176.

Richter-Boix A, Quintela M, Segelbacher G, Laurila A (2011) Genetic analysis of differentiation among breeding ponds reveals a candidate gene for local adaptation in *Rana arvalis*. *Molecular Ecology*, **20**, 1582–1600.

Roda F, Ambrose L, Walter GM *et al.* (2013) Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. *Molecular Ecology*, **22**, 2941–2952.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nature Review Genetics*, **15**, 749–763.

Schmidt PS, Serrao EA, Pearson GA *et al.* (2008) Ecological genetics in the north Atlantic: environmental gradients and adaptation at specific loci. *Ecology*, **89**, S91–S107.

Schoville SD, Bonin A, Francois O *et al.* (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 23–43.

Sillanpää MJ (2011) Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, **106**, 511–519.

Smouse PE, Long JC, Sokal RR (1986) Multiple-regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627–632.

Sork VL, Aitken SN, Dyer RJ *et al.* (2013) Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes*, **9**, 901–911.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Strasburg JL, Sherman NA, Wright KM *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 364–373.

Stucki S, Orozco-terWengel P, Bruford MW *et al.* (submitted) High performance computation of landscape genomic models integrating local indices of spatial association. arXiv preprint:1405.7658.

Thomassen HA, Cheviron ZA, Freedman AH *et al.* (2010) Spatial modelling and landscape-level approaches for visualizing intra-specific variation. *Molecular Ecology*, **19**, 3532–3548.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.

© 2015 John Wiley & Sons Ltd

Turesson G (1922) The genotypical response of the plant species to the habitat. *Hereditas*, **3**, 211–350.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*, **42**, 260–263.

Vilhjalmsson BJ, Nordborg M (2013) The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, **14**, 1–2.

de Villemereuil P, Gaggiotti OE (in press) A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.12418.

de Villemereuil P, Frichot E, Bazin E, Francois O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Weir BS, Anderson AD, Hepler AB (2006) Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, **7**, 771–780.

Williams GC (1966) *Adaptation and Natural Selection*. Princeton University Press, Princeton, New Jersey.

Yan J, Fine J (2004) Estimating equations for association structures. *Statistics in Medicine*, **23**, 859–874.

Yoder JB, Stanton-Geddes J, Zhou P *et al.* (2014) Genomic signature of adaptation to climate in *Medicago truncatula*. *Genetics*, **196**, 1263–1275.

Zueva KJ, Lumme J, Veselov AE *et al.* (2014) Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? *PLoS One*, **9**, e91672.

Zulliger D, Schnyder E, Gugerli F (2013) Are adaptive loci transferable across genomes of related species? Outlier and environmental association analyses in Alpine Brassicaceae species. *Molecular Ecology*, **22**, 1626–1639.

C.R. and R.H. conceived the study. All authors designed the focus, structure and content of the review. C.R. wrote the manuscript, with substantial contributions from F.G., A.J.E., A.M.H. and R.H.

© 2015 John Wiley & Sons Ltd