Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015

doi: 10.1111/2041-210X.12382

APPLICATION

LEA: An R package for landscape and ecological association studies

Eric Frichot[1] and Olivier François[1]*

[1]*Centre National de la Recherche Scientifique, Université Joseph Fourier Grenoble 1, TIMC-IMAG UMR 5525, Grenoble 38042, France*

Summary

1 Based on population genomic and environmental data, genomewide ecological association studies aim at detecting allele frequencies that exhibit significant statistical association with ecological gradients. Ecological association studies can provide lists of genetic polymorphisms that are potentially involved in local adaptation to environmental conditions through natural selection.

2 Here, we present the R package LEA that enables users to run ecological association studies from the R command line. The package can perform analyses of population structure and genome scans for adaptive alleles from large genomic data sets. It derives advantages from R programming functionalities to adjust significance values for multiple testing issues and to visualize results.

3 This note also illustrates the main steps of ecological association studies and the typical use of LEA for analysing data sets based on R commands.

Key-words: control of false discoveries, ecological association studies, genome scans for signature of local adaptation, inference of population structure

Introduction

Local adaptation through natural selection is an important driver of evolutionary changes in natural populations (Darwin 1859; Williams 1966), and understanding the molecular bases of local adaptation is a fundamental step in evolution, molecular ecology, global change or conservation biology (Joost *et al.* 2007; Manel *et al.* 2010; Jay *et al.* 2012; Schoville *et al.* 2012).

Using landscape genomic data, signatures of local adaptation can be detected by identifying allele frequencies that exhibit significant association with ecological gradients linked to various selection pressures (Joost *et al.* 2007; Hancock *et al.* 2008; Fumagalli *et al.* 2011; Frichot *et al.* 2013). To achieve this goal, genomewide ecological association studies screen genomic data that consist of thousands of individual genotypes, including single nucleotide polymorphisms (SNPs) and other types of allelic data. Ecological factors encompass climatic variables such as temperature and precipitation data (Hancock *et al.* 2008; Manel *et al.* 2010), habitat descriptors such as elevation, or pathogen density (Fumagalli *et al.* 2011), which are sources of spatially varying selection.

Computer tools that implement ecological association tests include the programs sam (Joost *et al.* 2007), Bayenv (Coop *et al.* 2010) and LFMM (Frichot *et al.* 2013). The programs sam and Bayenv are based on generalized linear regression models, whereas the program LFMM uses linear mixed models. Bayenv and LFMM are based on Bayesian methods that perform corrections for confounding effects due to patterns of isolation-by-distance, population structure and genomic background. All these programs share the drawbacks of requiring pre- and post-treatments that include analysis of population structure, control of false discovery rates and visualization of results. A program allowing users to perform the aforementioned treatments within a unified interface is still missing.

In this study, we present an integrated framework for population genetic analyses and ecological association studies. We describe the R computer package LEA that runs large-scale ancestry analyses, performs genome scans for selection, provides methods for solving multiple testing issues and generates graphical outputs for the results. More specifically, the LEA toolbox contains population structure estimation methods such as principal component analysis (PCA) or non-negative matrix factorization algorithms (sNMF, Frichot *et al.* 2014), and association methods such as latent factor mixed models (LFMM, Frichot *et al.* 2013). In addition, LEA contains procedures for calibrating statistical models and for controlling false discovery rates. The details of all options of LEA are available in online documentation files and in a web tutorial.

Program implementation, materials and methods

Genomewide ecological association studies include two main steps. The first step consists of assessing population genetic structure from

*Correspondence author. E-mail: olivier.francois@imag.fr

the genomic data, and evaluating the factors that could influence the interpretation of results. The second step consists of testing association of allele frequencies with ecological gradients. This step includes correction for biases due to population structure and other – often unobserved – confounding factors. The R package `LEA` enables performing the two analytical steps within a unified framework based on factor models and on the R statistical program. The package optimizes algorithmic speed and memory allocation while preserving the flexibility of statistical analysis using R. Functions implemented in `LEA` call functions written in the C programming language. These functions are able to process massive genomic data from the R command line without loading the program memory. Thus, the strength of the `LEA` package is to allow its users to perform computer intensive analyses, while benefiting of the statistical and visualization methods available from R.

DATA FORMAT

The R package `LEA` can handle several classical formats for input files of genotypic matrices. More specifically, the package uses the `lfmm` and `geno` formats and provides functions to convert from `ped`, `vcf` and `ancestrymap` formats. While the `lfmm` and `geno` formats usually encode SNP data, those formats can also be used for coding amplification fragment length polymorphisms and microsatellite markers. In addition to genotypic matrices, `LEA` can also process allele frequency data when they are encoded in the `lfmm` formats. Ecological variables must be formatted in the `env` format used by the computer program `LFMM` (Frichot *et al*. 2013).

ANALYSIS OF POPULATION STRUCTURE

The R package `LEA` implements two classical approaches for the estimation of population genetic structure: principal component analysis (PCA) and admixture analysis (Pritchard, Stephens & Donnelly 2000; Patterson, Price & Reich 2006). The algorithms programmed in `LEA` are improved versions of PCA and admixture analysis able to process very large genotypic matrices efficiently.

The `LEA` function `pca` computes the scores of a PCA for a genotypic matrix and returns a scree plot for the eigenvalues of the sample covariance matrix. Using `pca`, an object of class `pcaProject` is created. This object contains a path to the files storing eigenvectors, eigenvalues and projections. The number of significant components can be evaluated using graphical methods based on the scree plot or computing Tracy–Widom tests with the `LEA` function `tracy.widom` (Patterson, Price & Reich 2006).

Similar to Bayesian clustering programs, `LEA` includes an R function to estimate individual admixture coefficients from the genotypic matrix (Pritchard, Stephens & Donnelly 2000; François & Durand 2010). Assuming K ancestral populations, the R function `snmf` provides least-squares estimates of ancestry proportions (Frichot *et al*. 2014). The `snmf` function also estimates an entropy criterion that evaluates the quality of fit of the statistical model to the data using a cross-validation technique. The entropy criterion can help choosing the number of ancestral populations that best explains the genotypic data (Alexander & Lange 2011; Frichot *et al*. 2014). The number of ancestral populations is closely linked to the number of principal components that explain variation in the genomic data. Both numbers can help determining the number of latent factors when correcting for confounding effects due to population structure in ecological association tests.

ECOLOGICAL ASSOCIATION TESTS

The R package `LEA` performs ecological association tests based on latent factor mixed models (LFMM, Frichot *et al*. 2013). Let G denote the genotypic matrix, storing allele frequencies for each individual at each locus, and let X denote a set of d ecological variables. LFMMs consider genotypic matrix entries as response variables in a linear regression model

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell} \qquad \text{eqn 1}$$

where μ_ℓ is a locus-specific effect, β_ℓ is a d-dimensional vector of regression coefficients, U_i contains K latent factors, and V_ℓ contains their corresponding loadings (i stands for an individual and ℓ for a locus). The residual terms, $\epsilon_{i\ell}$, are statistically independent Gaussian variables with mean zero and variance σ^2. In latent factor models, associations between ecological variables and allele frequencies can be tested while estimating unobserved latent factors that model confounding effects. In principle, the latent factors include levels of population structure due to shared demographic history or background genetic variation. After correction for confounding effects, significant association between allele frequencies and an observed ecological variable is often interpreted as evidence for selection at a particular locus.

The R package `LEA` implements an improved version of the LFMM estimation algorithm proposed by Frichot *et al*. (2013). The R function `lfmm` computes the posterior distribution of the regression coefficients corresponding to each ecological factor using a Gibbs sampler algorithm. The `lfmm` function allows users to perform multiple runs of the estimation algorithm for distinct values of K. It creates an object of class `lfmmProject` that contains the z-scores and P-values for locus-specific effects in each run. The P-values are obtained from the Student t-distribution using $n-d-1$ degrees of freedom and can be recalibrated using R commands.

LATENT FACTOR MIXED MODELS IN PRACTICE

A correct calibration of LFMM tests assumes that the test P-values have uniform distribution when the ecological variables have no effect on genetic variation Running LFMM with distinct numbers of latent factors is the way by which users could choose models that check this condition. LFMM association tests exhibit better performances for values close to the number of significant components in a PCA, or close to the number of clusters obtained from a clustering analysis (Frichot *et al*. 2013). We suggest that the values obtained from analyses using the R functions `pca` or `snmf` could define a range to explore when running `lfmm` analyses. Deciding the best values for the number of latent factors in LFMM can then be based on the analysis of the histograms of test P-values. For multiple runs using a same value of K, z-scores can be combined with the Stouffer or similar methods (Liptak 1958). To decide which test can be applied (and choose K, the number of latent factors), we use a genomic inflation factor, λ, for example defined by Devlin & Roeder (1999) as $\lambda = \text{median}(z^2)/0{\cdot}456$, where z is a vector that contains z-scores for all loci, and 0·456 corresponds to the median of the chi-square distribution. P-values are correctly calibrated when the inflation factor is close to one. We then modify the z-scores by dividing them by the square root of λ. With this method, standard algorithms implemented in R can be used to produce lists of candidate loci based on the control of the false discovery rate (Benjamini & Hochberg 1995).

SIMULATED AND BIOLOGICAL DATA

We considered simulated genotypes from populations that underwent a demographic range expansion 1000 generations ago (Frichot *et al.* 2015). In computer simulations using the program SPLATCHE (Currat, Ray & Excoffier 2004), a rectangular area was colonized from a unique source located south of the area. The simulations implemented a non-equilibrium stepping-stone model based on a rectangular array of demes. For each deme, the migration rate was equal to $m = 0{\cdot}4$, the expansion rate was equal to $r = 0{\cdot}4$, and the carrying capacity was equal to $C = 100$. We simulated genetic variation at 4500 neutral SNPs and at 500 adaptive SNPs. We sampled four individuals from each of the 165 demes. To simulate genetic variation at adaptive loci, we created an artificial ecological gradient that paralleled the main axis of expansion. We linked allele frequencies to the ecological gradient by using the Haldane transform (Haldane 1948). This transform reproduces clinal allele frequency patterns as expected under spatially varying selection intensities. In addition to our simulated data set, we considered genomic data from the model plant *Arabidopsis thaliana* genotyped at 205 406 SNPs (Atwell *et al.* 2010). We focused our example on the study of 49 accessions from Scandinavia and considered ecological gradients linked to temperature by extracting 11 variables from the WorldClim data base at each of the 49 sampling sites (annual mean temperature, mean diurnal range, temperature seasonality, etc). We summarized the 11 variables as a unique ecological factor by computing the first principal component of the temperature variables.

Ecological association studies using LEA

In this section, we illustrate the use of the R package LEA for analysing ecological genomic data from simulated populations and from Scandinavian populations of the plant species *Arabidopsis thaliana* (Atwell *et al.* 2010).

ANALYSIS OF SIMULATED DATA

We started our analysis of the data by evaluating population genetic structure with the R function `snmf`. For number of factors ranging from 1 to 10, we estimated ancestry coefficients for each individual in the sample, and we computed the cross-entropy criterion as follows

```
project.snmf = snmf("genotypes.geno",K=1:10,entropy=T)
```

The cross-entropy criterion decreased when the number of factors increased from 1 to 6. A minimum value was obtained when $K = 8$ clusters were considered, indicating that genetic contribution from 8 ancestral populations optimally predicts masked individual genotypes (Fig. 1a). Population structure was also assessed using principal component analysis using the LEA function `pca`. In agreement with NMF results, substantial drops in the distribution of the empirical covariance matrix eigenvalues were observed for components 1–6. Thus, the functions `pca` and `snmf` provided congruent evidence of complex population genetic structure in the data.

We continued our analysis by performing ecological association tests on the genotypic matrix. We used the R function `lfmm` to fit latent factors mixed models to the data and test association between loci and a simulated ecological gradient. Based on our analysis of population structure, we computed locus-specific z-scores and P-values for numbers of latent factors ranging between $K = 4$ and $K = 10$. For each value of K, the Gibbs sampler algorithm was run 10 times for a period of 5000 cycles following a burn-in period of 5000 cycles. The corresponding LEA command with $K = 6$ latent factors is

```
project.lfmm=lfmm(input.file="genotypes.lfmm",environment.file="gradients.env",
K=6,iterations=10000,burnin=5000,repetitions=10)
zs.table=z.scores(project.lfmm)
```

Each run took approximately 20 min of a 2·4 GHz Intel Xeon 64 bit computer processing unit. The created object `project` contained paths to external files recording the results of LFMM runs, and the function `z.scores` extracted z-scores from those external files. Using a standard R command, z-scores were combined using the median value

```
zs = apply(zs.table, MARGIN = 1, median)
```

and a genomic inflation factor was computed as follows

```
lambda=median(zs^2)/.456
```

The genomic inflation factor indicated that a good choice for the number of latent factors was $K = 6$ (Fig. 1b), and P-values were adjusted as follows

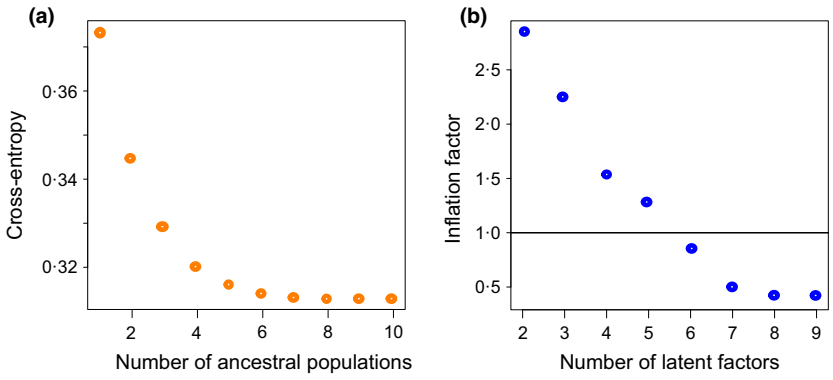

Fig. 1. Simulated data analysis. (a) Values of the cross-entropy criterion as a function of the number of factors in `snmf` runs. (b) Average values of the genomic inflation factor as a function of the number of latent factors in `lfmm` runs.

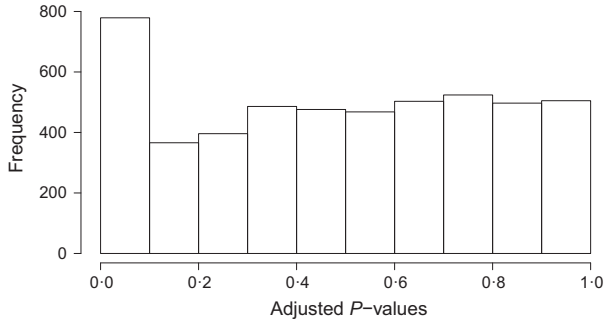

Fig. 2. Simulated data. Histogram of adjusted *P*-values obtained from `lfmm` runs using $K = 6$ factors (10 runs).

```
adj.p.values = pchisq(zs^2/lambda,
df=1,lower=F)
```

Figure 2 shows that , the adjusted *P*-values were correctly calibrated for $K = 6$ factors. To adjust *P*-values for multiple testing issues, we used the Benjamini–Hochberg procedure with expected levels of FDR equal to $q = 5\%$, 10%, 15% and 20%, respectively (Benjamini & Hochberg 1995). For an expected level of FDR equal to $q = 10\%$, a list of candidate loci is given by

```
L=length(adj.p.values)
q=0.1
w=which(sort(adj.p.values)
<q*(1:L)/L)
candidates=order(adj.p.values)[w]
```

For $K = 6$, the genomic inflation factor was equal to $\lambda = 0{\cdot}91$. The observed FDRs were equal to 4·9%, 8%, 10% and 13% for $q = 5\%$, $q = 10\%$, $q = 15\%$ and $q = 20\%$, respectively. These results suggest that values of the inflation factor less than 1 provide better calibration of LFMM tests than values greater than 1. In addition, the power to reject neutrality was equal to 70%, 85%, 91% and 94% for $q = 5\%$, $q = 10\%$, $q = 15\%$ and $q = 20\%$, respectively. For $K = 9$, the genomic inflation factor was equal to $\lambda = 0{\cdot}44$. The observed FDRs were equal to 7·7%, 11%, 15% and 19% for $q = 5\%$, $q = 10\%$, $q = 15\%$ and $q = 20\%$, respectively. The power to reject neutrality was equal to 81%, 91%, 96% and 99% for $q = 5\%$, $q = 10\%$, $q = 15\%$ and $q = 20\%$, respectively.

BIOLOGICAL EXAMPLE

We analysed *A. thaliana* population genetic data using the `LEA` functions `snmf` and `pca`. Using `snmf`, the cross-entropy criterion exhibited a minimum value for $K = 6$ factors (Fig. 3). Using `pca`, a break in the distribution of the eigenvalues was observed at the 6th eigenvalue. We performed ecological association tests using the `LEA` function `lfmm` with numbers of latent factors ranging from $K = 1$ to $K = 8$. The ecological gradient was derived from a linear combination of temperature variables. We ran the Gibbs sampler algorithm for a period of 5000 cycles following a burn-in period of 5000 cycles. The genomic inflation factor was closest to the value $\lambda = 1{\cdot}0$ for $K = 6$ latent factors. Using 6 latent factors and after controlling the FDR at the level $q = 5\%$, the program produced a list of 673 candidate SNPs, representing 0·3% of the total number of loci. We observed that 498 putatively adaptive SNPs were found in exomic sequences. Our list included SNPs in the chromosome 2 (AT2G27140, AT2G47940) and in the chromosome 5 (AT5G08000, AT5G07390) that were previously reported as being involved in biological processes related to heat stress and defence response. We performed a gene ontology enrichment analysis using the software `amiGO` in order to evaluate which molecular functions might be involved in adaptation to temperature gradients in *A. thaliana* (Carbon *et al.* 2009). We found significant enrichment in molecular functions linked to catalytic activity (catalysis of biochemical reaction at physiological temperatures, GO:0003824, $P = 1{\cdot}6\text{e-}8$) and hydrolase activity (G0:0016787, $P = 2{\cdot}7\text{e-}6$).

Discussion

Performing statistical analyses for genomewide ecological association studies requires several steps that include (i) assessment of confounding factors, (ii) corrections of statistical tests for biases generated by those factors and (iii) adjusting significance values for multiple testing issues. These steps are often conducted separately by using recently proposed approaches and by post-processing results with statistical programs. The main advantage of the R package `LEA` is to provide an approach for conducting all analytical steps from a unique interface. Users

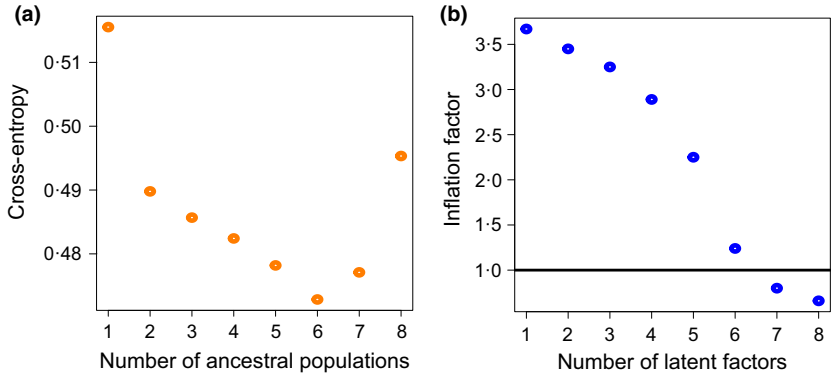

Fig. 3. *Arabidopsis thaliana* analysis. (a) Values of the cross-entropy criterion as a function of the number of factors used in `snmf` runs. (b) Average values of the genomic inflation factor as a function of the number of latent factors used in `lfmm` runs.

can benefit of the speed and efficiency of matrix factorization algorithms for analysing genomic data sets. In addition, they also benefit of many useful functionalities for visualization and analysis of the results obtained with those methods.

Our examples illustrated how traditional population structure analyses could be conducted from R, and how their results could be integrated in ecological association studies using latent factor models. PCA and clustering methods indeed provide useful information that help exploring the number of latent factors in LFMM analyses. Criteria that evaluate the quality of model predictions and the calibration of significance values were programmed in R using only a few language instructions. For example, model choice was based on the shape of *P*-value histograms evaluated though the genomic inflation factor. Computing the genomic inflation factor needed a single R language instruction, and *P*-values were corrected after running a simple R command. Calling the `pchisq` function, we applied FDR control procedures to generate lists of candidate loci, which was done using standard R functions as well. Our example suggested that evaluating the number of latent factors in latent factor models based on inflation factors and combining *P*-values from several runs lead to correct control of the FDR.

To conclude, the R package `LEA` provides an easy-to-use interface to ancestry estimation and genome scan programs for assessing association of allele frequencies to ecological gradients. The program combines the flexibility of the R environment and computer intensive programs that can process high volumes of genomic data.

INSTALLING THE R PACKAGE LEA

The `LEA` package can be installed from compressed `.zip` or `.tar.gz` files using the R command `install.packages`. These files are available from the `Bioconductor` resource repository http://www.bioconductor.org. Online documentations and tutorials are available from the authors' webpages.

Acknowledgments

This work was supported by a grant from la Région Rhône-Alpes to Eric Frichot and Olivier François. Olivier François acknowledges support from Grenoble INP and from the 'Agence Nationale de la Recherche' (project AFRICROP ANR-13-BSV7-0017).

Data accessibility

The *Arabidopsis thaliana* genotypes used in this study were publicly available from the Gregor Mendel Institute, Vienna, Austria (https://www.gmi.oeaw.ac.at/).

References

Alexander, D.H. & Lange, K. (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, **12**, 246.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., *et al*. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

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.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., *et al*. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

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

Currat, M., Ray, N. & Excoffier, L. (2004) SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Molecular Ecology Notes*, **4**, 139–142.

Darwin, C. (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London.

Devlin, B. & Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

François, O. & Durand, E. (2010) Spatially explicit bayesian clustering models in population genetics. *Molecular Ecology Resources*, **10**, 773–784.

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

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, **196**, 973–983.

Frichot, E., Schoville, S.D., de Villemereuil, P., Gaggiotti, O.E. & François, O. (2015) Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity*, doi: 10.1038/hdy.2015.7. In Press.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L. & Nielsen, R. (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, **7**, e1002355.

Haldane, J.B.S. (1948) The theory of a cline. *The Journal of Genetics*, **48**, 277–284.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G. & Di Rienzo, A. (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, 13.

Jay, F., Manel, S., Alvarez, N., Durand, E.Y., Thuiller, W., Holderegger, R., Taberlet, P. & François, O. (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, **21**, 2354–2368.

Joost, S., Bonin, A., Bruford, M.W., Després, L., Conord, C., Erhardt, G. & Taberlet, P. (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Liptak, T. (1958) On the combination of independent tests. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, **3**, 171–197.

Manel, S., Joost, S., Epperson, B.K., Holderegger, R., Storfer, A., Rosenberg, M.S., Scribner, K.T., Bonin, A. & Fortin, M.J. (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Patterson, N., Price, A.L. & Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, 20.

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

Schoville, S.D., Bonin, A., François, O., Lobreaux, S., Melodelima, C. & Manel, S. (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology and Systematics*, **43**, 23–43.

Williams, G.C. (1966) *Adaptation and Natural Selection, volume 1996*. Princeton University Press, Princeton, New Jersey.

Received 30 September 2014; accepted 24 March 2015
Handling Editor: Brian O'Meara