Molecular Ecology (2007) **16**, 3955–3969 doi: 10.1111/j.1365-294X.2007.03442.x

A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation

S. JOOST,*† A. BONIN,‡ M. W. BRUFORD,§ L. DESPRÉS,‡ C. CONORD,‡ G. ERHARDT¶ and P. TABERLET‡**

Istituto di Zootecnica, Università Cattolica del S.Cuore, via E. Parmense 84, 29100 Piacenza, Italy, †Laboratoire de Systèmes d'Information Géographique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Bâtiment GC, Station 18, 1015 Lausanne, Switzerland, ‡Laboratoire d'Ecologie Alpine, CNRS-UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 09, France; §Cardiff School of Biosciences, Main Building, Museum Avenue, Cardiff CF10 3TL, UK, ¶Department of Animal Breeding and Genetics, Justus-Liebig-University of Giessen, Ludwigstrasse 21B, 35390 Giessen, Germany

Abstract

The detection of adaptive loci in the genome is essential as it gives the possibility of understanding what proportion of a genome or which genes are being shaped by natural selection. Several statistical methods have been developed which make use of molecular data to reveal genomic regions under selection. In this paper, we propose an approach to address this issue from the environmental angle, in order to complement results obtained by population genetics. We introduce a new method to detect signatures of natural selection based on the application of spatial analysis, with the contribution of geographical information systems (GIS), environmental variables and molecular data. Multiple univariate logistic regressions were carried out to test for association between allelic frequencies at marker loci and environmental variables. This spatial analysis method (SAM) is similar to current population genomics approaches since it is designed to scan hundreds of markers to assess a putative association with hundreds of environmental variables. Here, by application to studies of pine weevils and breeds of sheep we demonstrate a strong correspondence between SAM results and those obtained using population genetics approaches. Statistical signals were found that associate loci with environmental parameters, and these loci behave atypically in comparison with the theoretical distribution for neutral loci. The contribution of this new tool is not only to permit the identification of loci under selection but also to establish hypotheses about ecological factors that could exert the selection pressure responsible. In the future, such an approach may accelerate the process of hunting for functional genes at the population level.

Keywords: AFLP, GIS, landscape genomics, local adaptation, microsatellites, natural selection, spatial analysis

Received 21 March 2007; revision accepted 4 June 2007

Introduction

Uncovering the genetic basis of adaptation to different environments represents a goal of central importance in evolutionary biology (Storz 2005). The detection of signatures of natural selection within the genomes of organisms is key, since it may allow a greater understanding of what proportion of a genome or which genes are being shaped by ongoing natural selection. Such studies are also of fundamental biological interest because they can reveal the very nature of adaptation and speciation (MacCallum & Hill 2006). In general, regions of the genome that are under selection are likely to be of functional importance, and inferences regarding selection may provide important information (Nielsen 2005). The discovery of such genomic regions is the keystone of promising applications in

Correspondence: Dr Stéphane Joost, Fax: +41 21 691 77 86; E-mail: stephane.joost@a3.epfl.ch
**and the ECONOGENE Consortium (http://www.econogene.eu)

health, conservation biology and selective breeding (Luikart *et al*. 2003).

Various methods have been developed to reveal genomic regions that are likely to be the target of natural selection (Vasemägi & Primmer 2005). Some of them belong to the 'candidate-gene' approach and take a particular locus as a starting point and assess whether it has been affected by selection (Phillips 2005; Wright & Gaut 2005). This can be done, for example, by using neutrality tests based on DNA sequence variation (Ford 2002; Nielsen 2005). Another category of methods aims at identifying quantitative trait loci (QTL) involved in the expression of adaptive traits. Usually, such QTL are detected by measuring statistical associations between phenotypic values and genotypes at molecular markers in a mapping population (Mackay 2001; Erickson *et al*. 2004).

Although these two kinds of approaches have proved helpful in many cases, their application is nevertheless often limited to relatively few well-studied species (Phillips 2005) or to populations that have been studied long enough to allow environmental variance to be adequately incorporated (e.g. Wilson *et al*. 2006). They usually require a priori information that may not be easily accessible for nonmodel organisms, such as phenotypic or family data (QTL detection) or information about the sequence and/or function of the studied gene (Ford 2002; Erickson *et al*. 2004). Fortunately, alternative strategies are now possible with the development of population genomics, which enables genome selection studies in the absence of prior knowledge about the selectively advantageous gene or phenotype (Storz 2005). Population genomics relies on the principle that loci across the genome are influenced by genome-wide evolutionary forces (e.g. genetic drift, gene flow), whereas locus-specific forces, such as selection, imprint a particular pattern of variability on linked loci only (Luikart *et al*. 2003). By comparing the genetic diversity of many loci across the genome, it is then possible to reveal loci displaying an atypical variation pattern, which are likely to be linked to those genomic regions affected by selection (Black *et al*. 2001). Therefore, in contrast to candidate-gene-based methods, strategies making use of population genomics do not focus on a few loci only, but rather depict the effect of selection over the whole genome (Storz 2005).

However, the population genomics approach still makes use of genetic models in order to identify those adaptive genes, thus resting on hypotheses that are not always verifiable, like the Hardy–Weinberg equilibrium. Moreover, this strategy makes it possible to detect possible adaptive loci, but this still remains difficult and above all it is often not possible to link them up with specific selection pressures (environmental for example).

In this paper, we propose an approach to address the issue from an environmental perspective in order to complement results obtained by population genetic models. We introduce a new method to detect signatures of natural selection based on the application of spatial analysis. With the contribution of geographical information systems (GIS), environmental data, molecular data and multiple univariate logistic regressions, we test for association between the allelic frequencies at molecular markers and data from various environmental variables. This spatial analysis method (SAM) is totally different from the population genomics approach, although both are designed to scan hundreds of molecular markers and have a common goal that is to identify loci likely to be under natural selection.

We applied this approach to two case studies. The first concerns large pine weevil (*Hylobius abietis*) populations at 20 infested European sites (Fig. 1). This major pest of conifer plantations is widespread throughout European managed forests, where its life cycle and activity vary according to location (reviewed in Day *et al*. 2004). The combination of a large geographical range together with large populations, where genetic drift is likely to be minor (Conord *et al*. 2006), makes it a potentially good candidate to detect signatures of natural selection throughout various environmental gradients at the continental scale. In this case, 10 environmental variables were exploited in order to look for signatures of natural selection within the genome of the pine weevil, which was analysed using 83 polymorphic amplified fragment length polymorphism (AFLP) markers (with band frequency varying between 0.09 and 0.86).

The second case study focuses on breeds of sheep (*Ovis aries*) sampled in the context of the European ECONOGENE Project, whose goal was to address the conservation of sheep and goat genetic resources in marginal agrosystems in Europe (http://www.econogene.eu). Fifty-seven sheep breeds originating from European and Middle Eastern countries were analysed (Fig. 2). As agriculture spread from the Middle East via southeastern Europe to the rest of Europe in the Neolithic (Ryder 1983), numerous sheep breeds were developed. This was partly due to selection by man but also to climatic and other environmental variables, since Europe is a geographically complex continent with a particularly wide variety of landscapes and climates: highland regions of the Alps contrast to lowland plains of Poland, Germany or the Netherlands, and the oceanic wet and cold climate of northern Europe to the dry Mediterranean. Further, there is no palaeoevidence for wild sheep having been indigenous to Europe during the Pleistocene (Clutton-Brock 1999), hence rapid adaptation to hostile environments is expected to have shaped at least some of the genomic diversity in marginal European sheep breeds. We thus expected that evidence for divergent selection should be detected in European sheep, both as a response to human-mediated selection and due to environmental selection pressures on marginal habitat. Here, genomic diversity was estimated using 31 microsatellite markers (mean expected heterozygosity 0.72).

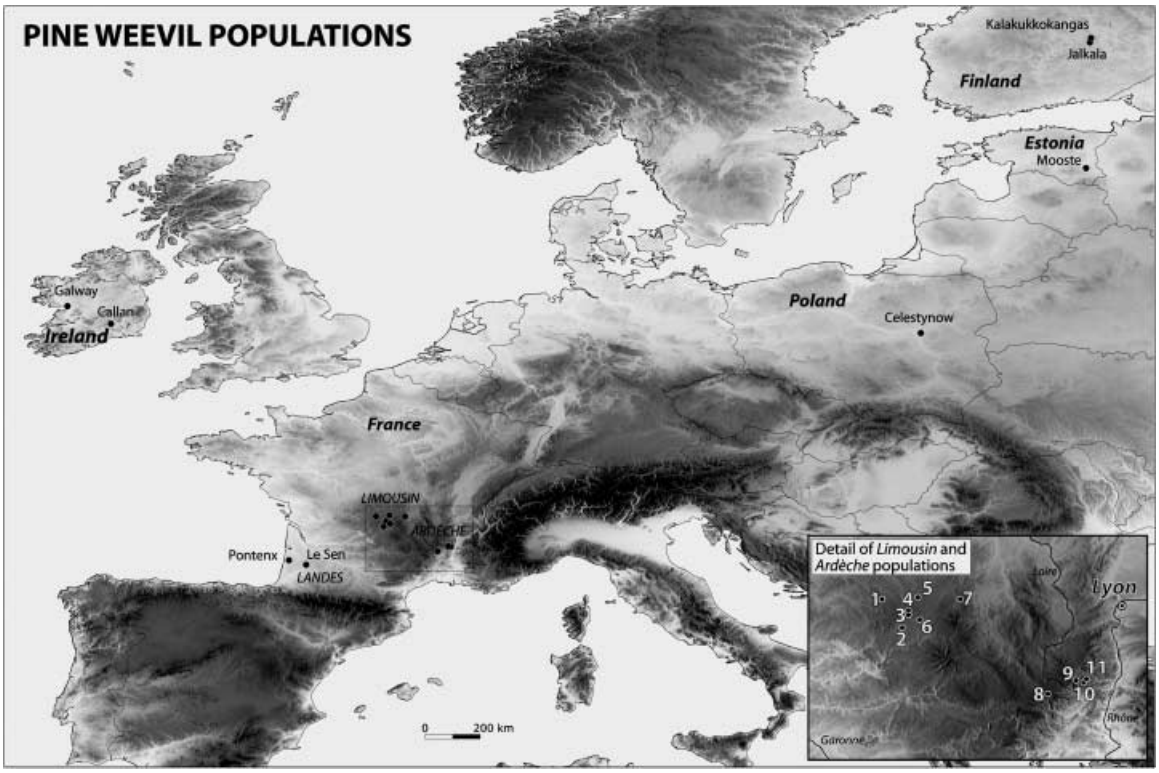

Fig. 1 Spatial distribution of the studied pine weevil populations in Europe (black dots). *Limousin* populations are: (1) Royère, (2) Maussac, (3) Bellechassagne, (4) Annouillards, (5) Basville, (6) Ebraly and (7) Pontgibaud; and *Ardèche* populations are: (8) St Etienne de Lugdares, (9) Lachamp Raphaël, (10) Les Quatre Vios and (11) Mézilhac. [Topography: SRTM30 NASA]

Both case studies (*Hylobius abietis* and *Ovis aries*) were also analysed using the population genomics methodology described in Beaumont & Nichols (1996) in order to compare the results provided by the two approaches.

Materials and methods

Spatial Analysis Method

SAM is based on one of six concepts of spatial analysis distinguished by Goodchild (1996). In order to connect genetic information with geo-environmental data (i.e. information characterizing animal or plant organisms with properties of their surroundings), we utilize *spatial coincidence* analysis. This approach associates information levels and is able to compare across them thanks to their common geographical coordinates. Thus SAM requires a geo-referenced data set comprising one or more environmental variables describing the sampling location (for instance mean monthly precipitation), and a geo-referenced molecular marker data set for the study population(s).

Logistic regression is then used to provide a measure of the association between the frequency of molecular markers or AFLP bands and the environmental parameters at each site. AFLP data are ideal for logistic regression because they provide binomial information. However, in the case of microsatellites, it is necessary to encode the data: each allele is set to '1' if it occurs in a given individual, and to '0' if not. Then the association is tested between each allele and each environmental parameter. Logistic regression is used to assess the significance of the models constituted by all possible [marker ↔ environmental variable] pairs, and to highlight the markers implicated in the most significant models as potential candidates for linkage to genomic regions involved in adaptation. To this end, the significance of coefficients calculated by the logistic regression function is evaluated by statistical tests addressing the question of whether a model including an environmental variable is more informative about the response variable than a model with a constant only. In logistic regression, the comparison of observed with predicted values is based on the log-likelihood function. Following Hosmer & Lemeshow (2000), we used the likelihood ratio (G) and Wald tests to determine the significance of the models:

(a) The likelihood ratio or G statistic is

$$G = -2 \ln \frac{L}{L'}$$

where L is the likelihood of the initial model (with a constant only) and L' is the likelihood of the new model including

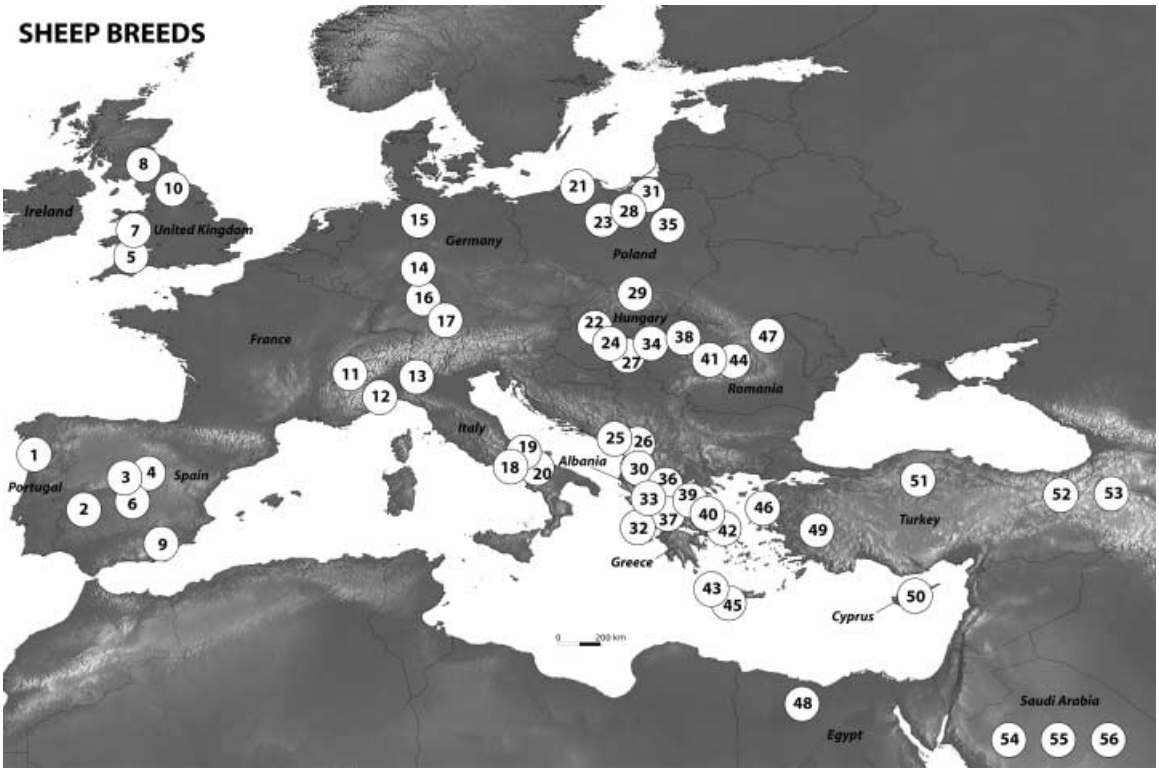

Fig. 2 Spatial distribution of sheep breeds through the ECONOGENE study area. 1. Churra Braganzana, 2. Spanish Merino I and II, 3. Rubia del Molar, 4. Colmenarena, 5. Exmoor Horn, 6. Manchega, 7. Welsh Mountain, 8. Scottish Blackface, 9. Segurena, 10. Swaledale, 11. Thones et Marthod, 12. delle Langhe, 13. Bergamasca, 14. German Merino (Merinolandschaf), 15. German Grey Heath, 16. Rhoensheep, 17. White/Brown Mountain, 18. Laticauda, 19. Gentile di Puglia, 20. Altamurana, 21. Pomeranian, 22. Cikta, 23. Polish Merino, 24. Hungarian Merino, 25. Shkodrane, 26. Bardhoka, 27. Hungarian Tsigaia, 28. Kameniec, 29. Polish Mountain (Gorska), 30. Ruda, 31. Wrzosowska, 32. Keffaleneas, 33. Orino, 34. Magyar Racka, 35. Zelazna, 36. Kalarritiko, 37. Karagouniko, 38. Transylvanian Merino, 39. Pilioritiko, 40. Skopelos, 41. Turcana, 42. Kymi, 43. Sfakia, 44. Romanian Tsigaia, 45. Anogeiano, 46. Lesvos, 47. Black Karakul, 48. Ossimi, 49. Daglic, 50. Cypriot fat-tailed, 51. Karayaka, 52. Akkaraman, 53. Morkaraman, 54. Naemi* (Saudi Arabia), 55. Heri* (Saudi Arabia) and 56. Najdi* (Saudi Arabia) **breeds from Saudi Arabia are not accurately located*. [Topography: SRTM30 NASA]

the examined variable. If added parameters are equal to zero, this statistic follows a chi-squared distribution, where the degrees of freedom equal the number of added parameters (Hosmer & Lemeshow 2000).

(b) The Wald statistic is

$$W = \frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}$$

where β is the maximum likelihood for parameter $_i$; $\hat{\beta}_i$ is the maximum likelihood estimate of the parameter β_i, and $\sigma(\hat{\beta}_i)$ an estimate of its standard error. Under the null hypothesis, the resulting ratio follows a normal distribution. The method to assess the variance is in accordance with the theory of maximum likelihood (Hosmer & Lemeshow 2000). For both tests, the null hypothesis is that the model with the examined variable does not explain the observed distribution better than a model with a constant only.

A model is considered significant only if both tests reject the corresponding null hypothesis. Indeed, contradictory observations were found in the literature about their reliability. Hauck & Donner (1977) consider that the Wald test behaves in an aberrant manner and often fails to reject the null hypothesis. Agresti (1990) and Tu & Zhou (1999) state that the likelihood ratio test outperforms the Wald statistic, while the performance of the latter is satisfactory when the size of the samples is large. Whereas Conte & de Maio (2003) stipulate that the Wald test outperforms the others and is very effective: in case of large logit coefficients the standard error is inflated. This lowers the Wald statistic and leads to type II errors, i.e. false negatives (Menard 2002).

Given the fact that molecular data sets may contain many markers and that many different environmental parameters are likely to describe a sampling site, many univariate models have to be run simultaneously in order to detect markers likely to be under natural selection. It is recognized that when one wishes to test several hypotheses at a common significance level α simultaneously, the generalized type I error probability (the probability of rejecting at least one of the hypotheses being tested that is in fact true) is typically much higher than α (multiple

hypotheses testing). There are a large number of multiple testing procedures available. We chose to apply the Bonferroni correction (Shaffer 1995). This correction divides the desired significance threshold α by the number of comparisons (the number of models simultaneously processed). Although this correction is known to be very conservative (Garcia 2004; Narum 2006), we chose to efficiently limit the number of significant models in order to restrict the analysis to robust candidate associations. Moreover, the application developed easily allows adaptation of the confidence threshold and thus one can gradually take more models into consideration.

The processing of the numerous resulting models has been automated within the SAM program developed in Matlab®, and makes use of the GLMfit (generalized linear model fitting, MacCullagh & Nelder 1989) function. The procedure manages the number of models to be calculated (the user has to indicate the number of markers and the number of environmental variables to be processed), solves the likelihood equations allowing the maximum likelihood estimates of the parameters to be determined, calculates the *P*-values associated with both *G* and Wald statistical tests for each model, generates graphs with response curves for each model, and exports tables in text format to be imported in any spreadsheet or statistical software. To analyse the results, an Excel macro was developed in Visual Basic to set up dynamic tables designed to automatically process the large amount of results provided by the SAM program. It allows the user to set an initial confidence level and to progressively adapt it in order to identify the most significant models. Both SAM and the Excel macro can be obtained on demand.

The method highlights molecular markers involved in significant models as defined by both statistical tests, and it also provides the list of markers revealed by one test only. Indeed, it may happen that the maximum number of iterations is reached before the maximum likelihood equation is solved for one of the tests. In this case, the system allows the user to take the model into consideration if a marker was detected by a population genetics approach for instance.

Population genomics analysis

To assess the results provided by SAM, the AFLP data set was analysed with DFDIST and the microsatellite data set with FDIST2. Both programs are modified from FDIST (Beaumont & Nichols 1996). DFDIST is designed to handle dominant markers. The FDIST2 method is a F_{ST}-outlier test using coalescent simulations to model the behaviour of neutral loci under a symmetrical island model of population structure (Wright 1951). It is based on the principle that genetic differentiation between populations is expected to be higher for loci under divergent selection than for the rest of the genome, considered as neutral. In the DFDIST version, the Bayesian method developed by Zhivotovsky (1999) is implemented to estimate allele frequencies from the proportion of recessive phenotypes in the sample. For each locus, the allele frequencies are used to compute F_{ST} values conditional on heterozygosity. Loci showing atypical differentiation behaviour (i.e. F_{ST}) and lying outside the simulated neutral distribution are then detected as outliers. The FDIST method was selected to directly compare results for both codominant and dominant markers and because comparisons across studies have shown that the approach is relatively robust to violations of its assumptions on mutation, migration and population structure (Beaumont & Nichols 1996).

Pine weevil

The large pine weevil *Hylobius abietis* L. (Curculionidae) is one of the most important economic pests of European conifer forests; the larvae feed under the bark of stumps and roots of recently felled trees, and take from three months to two years to develop into adults, depending on location (Day *et al*. 2004), presumably because of climatic conditions and/or host plant quality (von Sydow & Birgersson 1997; Thorpe & Day 2002). The adults are active only under cool climatic conditions, usually in spring and autumn, and burrow into the soil during hot summers and cold winters; adults can fly large distances and can live up to four years (Day *et al*. 2004). Because of this complex life cycle, several climatic factors including temperature, precipitation, soil, frost and wind speed may have either a direct impact on larval/adult survival, or a more indirect impact on fitness through host plant quality, and represent therefore potential selective forces acting on the pine weevil genome at a large geographical scale. A total of 367 weevils (larvae and adults) were collected in 20 managed forest sites across Europe in Estonia, Poland, Finland, Ireland and France, covering most of the geographical range of the species in Europe (Fig. 1).

The genome was scanned using AFLP (Vos *et al*. 1995; see Conord *et al*. 2006). Briefly, total genomic DNA was extracted using the DNeasy Tissue Kit (Qiagen) following the manufacturer's instructions. It was then digested with restriction enzymes *Eco*RI and *Taq*I and ligated to specific adaptors. Preselective and selective amplifications were performed as indicated in Conord *et al*. (2006), using three different primer pairs. The fragments were separated on an ABI Prism 3100 DNA automated sequencer (Applied Biosystems) and visualized with GENESCAN® ANALYSIS 3.7 (Applied Biosystem) and GENOGRAPHER version 1.6.0 (Benham *et al*. 1999; available at http://hordeum.oscs.montana.edu/genographer/). AFLP profiles were recorded in a matrix as presence (1) or absence (0) of bands for each individual. A fragment was considered as absent in a given individual if the electropherogram showed a lack of signal after normalization of the corresponding peaks. A scoring

threshold was then set up which generally corresponded to 10% of the size of the highest peak found among all compared individuals. Fragments that could not be scored unambiguously were not included in the analyses. Overall, 367 individuals were screened for 83 AFLP markers.

Analyses with DFDIST were conducted with individuals grouped according to the forest region they belong to (i.e. Landes, Ardèche, Limousin, Estonia-Poland, Ireland, and Finland), because these have been shown to represent six homogeneous genetic entities (Conord *et al*. 2006). The sample size ranged from 20 (Landes) to 115 (Limousin). The expected F_{ST} for the island model was 0.0299 (0.0204 without the population of the Landes). The analysis was carried out using 50 000 simulations.

Sheep breeds

Fifty-seven sheep breeds, originating from European and Middle Eastern countries (Fig. 2) were sampled for 17–32 unrelated animals in the original region of the breed from herdbook flocks, where available. From each farm, three unrelated individuals were sampled from an average of 10 flocks per breed. Breeds were mainly autochthonous with one cosmopolitan (Merino) breed, sampled from Germany, Poland, Hungary, Romania and Spain included (Peter *et al*. 2007). The combined effects of demographic (genetic drift, inbreeding, introgression) and selection (artificial and natural) are strongly implicated in the loss of diversity of livestock breeds (Bruford 2004). This system therefore posed specific, yet potentially important challenges for detecting signatures of selection within the genome (for example, the island model is unlikely to apply in many cases), both to improve our understanding of the mechanisms underlying livestock genetic change (Luikart *et al*. 2003), but also to assist the discovery of genomic regions linked to quantitative trait loci implicated in selection and adaptation (e.g. for disease resistance).

Thirty-one bovine, ovine and caprine microsatellite markers were used (Peter *et al*. 2007) producing a total of 744 alleles whose frequency was compared to environmental parameters. Markers on all autosomal sheep chromosomes (OAR) except OAR8, OAR21, OAR22 and OAR23 were included. DNA extraction was performed by following the protocol of Montgomery & Sise (1990). Primer sequences, size ranges, multiplexing information and PCR protocols of the markers are available from the UN Food and Agriculture Organization website (http://dad.fao.org/en/refer/library/guidelin/marker.pdf; FAO 1998) which ranks markers by typing efficiency (e.g. PCR-amplification, scoring reliability and lack of ambiguity). Genotyping was performed on ABI Prism 377 and ABI 3100 semiautomated DNA analysers (Applied Biosystems), using standard methodologies (Peter *et al*. 2007). In order to link genotype results with future diversity studies, five 'reference' DNA samples from four breeds were also included in the genotyping procedure.

In FDIST2, individuals were grouped according to the breed they belong to (57 populations), with 100 demes, an expected F_{ST} for infinite allele, infinite island model of 0.055. The real sample size ranged from 17 to 32. The analysis was carried out using 20 000 simulations. Finally, the results obtained either with the stepwise or the infinite alleles model looked the same.

Environmental data

Altitude was estimated using the 30 arc-second digital elevation model (DEM) of the Shuttle Radar Topography Mission (SRTM30, NASA, http://www2.jpl.nasa.gov/srtm/). Climatic data comprised latitude/longitude grids with a resolution of 10 min (approximately 12 km at the latitude of Switzerland) containing nine monthly variables over global land areas described in Table 1. These climate data were collated and made available by

Table 1 List of environmental variables. Yearly means only were used in the pine weevil case study. In the sheep case study, monthly values were used in addition to yearly means, for a total of 13 periods per variable. Thus 118 environmental parameters were used in this analysis (altitude +9 climatic variables × 13 periods). With the exception of altitude (SRTM30, NASA, http://www2.jpl.nasa.gov/srtm/), environmental variables were computed by the Climatic Research Unit for the period 1961–2001, http://www.cru.uea.ac.uk (New *et al*. 2002)

Variable	Description
Altitude	Altitude computed with NASA SRTM30 Digital Elevation Model
DTR	Yearly mean and monthly values of mean diurnal temperature range in °C
FRS	Yearly mean and monthly values of number of days with ground frost
PR	Yearly mean and monthly values of precipitation in mm/month
PRCV	Yearly mean and monthly values of the coefficient of variation of monthly precipitation in percent
REH	Yearly mean and monthly values of relative humidity in percentage
SUN	Yearly mean and monthly values of percent of maximum possible sunshine
TMP	Yearly mean and monthly values of mean temperature in °C
WET	Yearly mean and monthly values of wet days (number of days with > 0.1 mm rain per month)
WND	Yearly mean and monthly values of wind speed in m/s, 10 metres above the ground

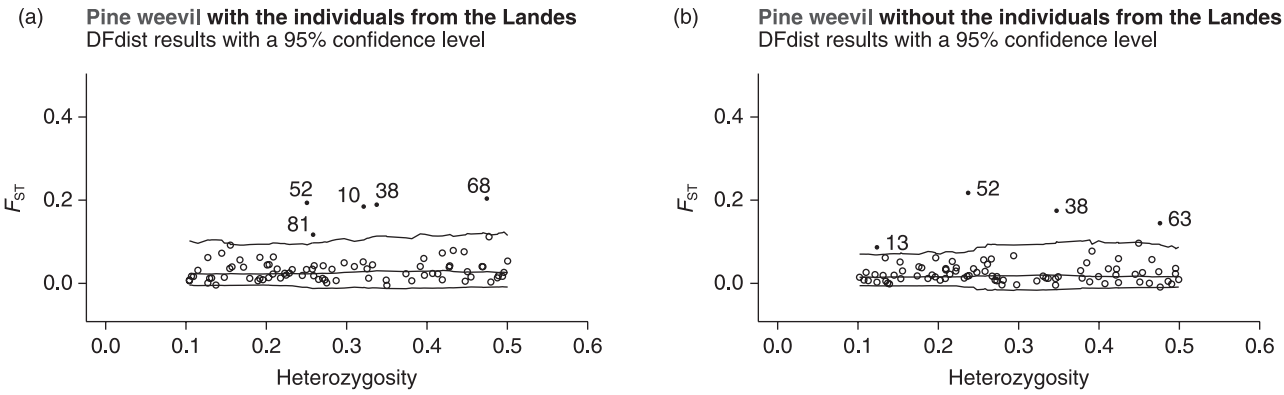

Fig. 3 Graphical representation of the results obtained with the DFDIST software (Beaumont & Nichols 1996) for pine weevils with the individuals from the Landes (a) and without the populations from the Landes (b).

the Climatic Research Unit, Norwich (CRU) and covers the period 1961–90 (New *et al*. 2002). Yearly means were used in the pine weevil case study (10 variables), whereas monthly and yearly values were utilized for the sheep data (118 variables, see Table 1). Monthly variables were considered in sheep because several management and production systems based on lambing periods (or seasons) are used in different breeds (e.g. autumn lamb production, winter lambing, or spring lamb production).

Results

Pine weevil

SAM identified 11 markers as significant with both tests, with a significance threshold set to 1.2E-5 (corresponding to a 99% confidence level including Bonferroni correction). Four markers (13, 38, 52 and 63) were highly significant with both tests (Table 2). Among the latter, markers 38, 52 and 63 were still significant for a confidence level of 1.2E-13. Four loci (52, 38, 10 and 68) were detected as outliers with DFDIST at the 99% confidence level, and when the analysis was reiterated at the 95% confidence level, one additional locus (marker 81) was detected (Fig. 3a). Only two of these loci, 38 and 52, were also detected by SAM. However, when the 20 individuals sampled in the Landes were excluded from the DFDIST analysis, loci 10 and 68 were no longer detected as outliers by DFDIST, while loci 63 and 13 appeared as outliers as also detected by SAM (Fig. 3b). This was done because locus 68 was monomorphic in the Landes, while locus 10 was nearly fixed (present in 18 out of 20 individuals). The removal of individuals from Landes only very slightly modified the results of SAM: marker 13 was detected by SAM (by *G* test only) at the 1.2E-15 confidence level (after Bonferroni correction), one order of magnitude higher (1.2E-14) than when Landes was included. The loci 38 and 52, which were strongly detected by both methods, appear as the best candidates for selection at the European scale. Locus 38 was positively correlated with the number of days of ground frost and negatively with precipitation, sunshine and diurnal temperature range, while locus 52 was more commonly found in regions with increasing diurnal temperature range (Fig. 4).

Sheep

Alleles absent from almost all breeds, as well as those present in almost all breeds were removed from the calculations because they do not contribute in discriminating environmental information. Therefore, the lowest and highest 5% of the distribution were not taken into account for processing of the models. 744 alleles located at 31 loci were analysed with 118 topo-climatic variables (see Table 1) for a total of 87 792 models whose significance was assessed.

A total of 141 alleles (19% of the total number of investigated alleles) were identified by SAM as significantly associated with at least one environmental variable with one of both statistical tests, with a confidence level of 99% (significance threshold [ST] set to 1.139E-7). Forty alleles at 21 loci were detected as significantly associated with at least one environmental parameter with a confidence level of 99.999% (ST = 1.139E-10), and with both the Wald and *G* tests. At this level of significance, there were 407 models for the Wald test representing 0.46% of the total number of processed models, and 577 models for the *G* test representing 0.65% of the total number of processed models. Figure 5(a) shows the corresponding loci with histograms of the number of related candidate alleles. Of these 21 loci, six were detected by both SAM and FDIST2 (HUJ616, OARFCB193, OARFCB304, MCM140, OARJMP29 and INRA63), with a confidence level of 99% in FDIST2 Fig. 6).

Table 2 SAM results for pine weevil with a significance level of 1.2E-14 (including Bonferroni correction). Only 49 markers on a total of 83 are displayed. On the second line, markers in bold were identified both by SAM and the population genetics theoretical approach. Each cell represents a model, i.e. a molecular marker vs. an environmental parameter. Grey cells with '1' indicate that for this model, the null hypothesis (H_O) is rejected with the corresponding test. Black cells with '2' indicate that H_O is rejected for both tests. Black arrows indicate models with *P*-value < 1E-20 for both tests (default limit set in the Matlab® procedure).
This table also illustrates the way the many results produced by the parallel processing of many univariate logistic regressions—a characteristic of the SAM method—can be sorted out and how significant models can be visually emphasized.

		Frequency of markers →	...	54	55	59	59	62	...	106	106	108	110	112	115	115	115	117	129	130	133	134	136	141	141	151	151	160	170	177	180	180	182	195	195	196	201	202	203	203	203	217	226	228	229	230	230	246	...	256	262	269	316
		Marker ID →	...	**13**	56	39	67	12	...	79	80	81	6	36	1	**52**	82	49	30	62	58	32	**38**	10	28	40	55	25	17	22	4	33	20	18	76	66	75	46	27	44	60	69	68	16	**63**	5	15	73	...	59	57	61	35
Likelihood Ratio Statistic (*G*) Null Hypothesis Test	Environmental variable	DTR		1	0	0	0	0		0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		FRS		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		PR		0	0	0	0	0		0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		WET		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		REH		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		SUN		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		TMP		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		WND		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		PRCV		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		Altitude		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
Wald Statistic (B1 parameter) Null Hypothesis Test	Environmental variable	DTR		0	0	0	0	0		0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		FRS		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		PR		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		WET		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		REH		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		SUN		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0		0	0	0	0
		TMP		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		WND		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		PRCV		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		Altitude		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
Null Hypothesis Test Cumulated	Environmental variable	DTR		0	0	0	0	0		0	0	0	0	0	0	2	←		0	0	0	0	2	←		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		FRS		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	2	←		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0		0	0	0	0
		PR		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	2	←		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		WET		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0		0		0	0	0	0
		REH		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		SUN		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	2	←		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0		0	0	0	0
		TMP		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		WND		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		PRCV		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
		Altitude		0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0

Significance level
with Bonferroni correction
= 1.20E-14

Probability to reject H0	1.00E-1
Number of environmental variables	10
Number of molecular markers	83
Total number of models	830

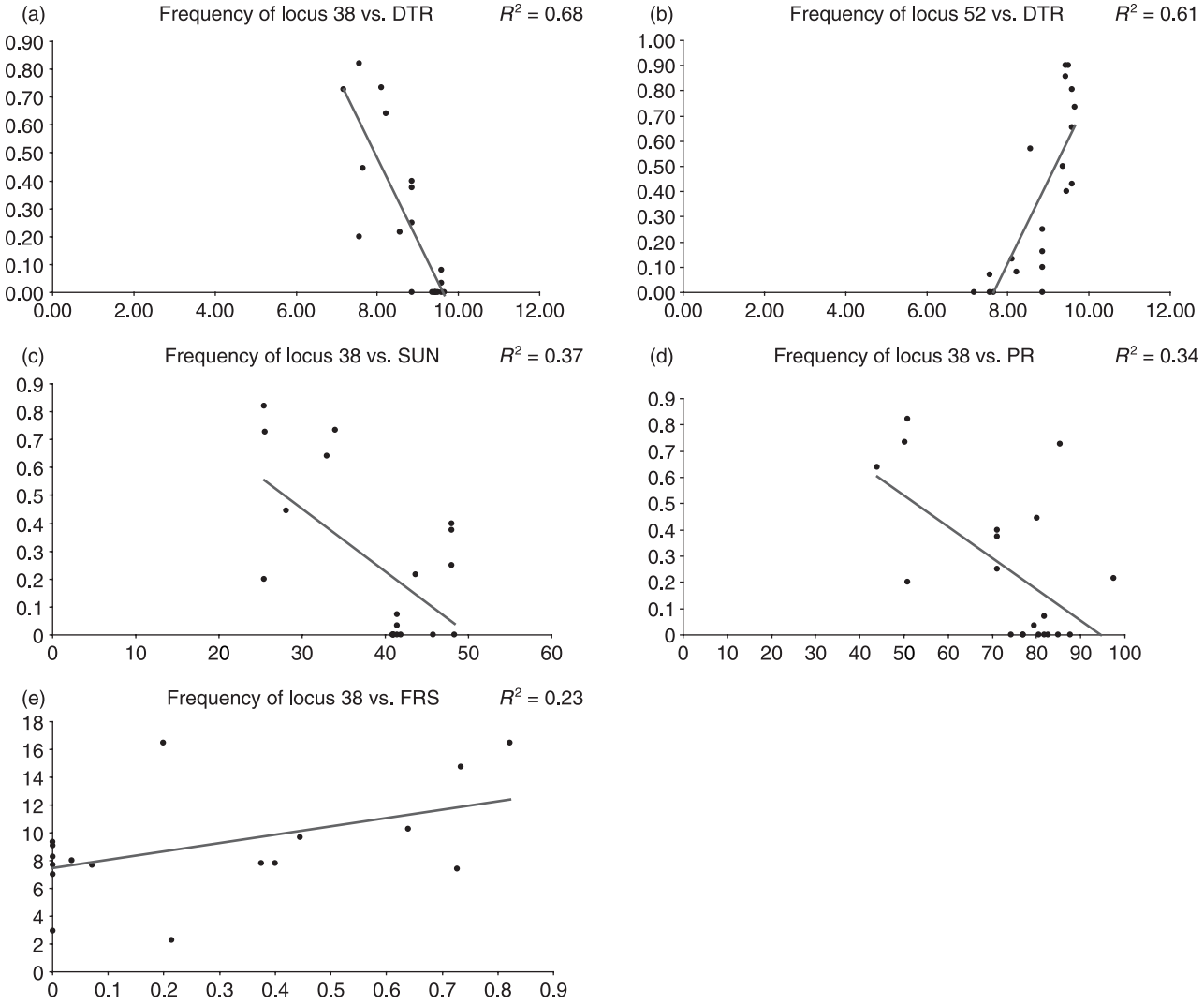

Fig. 4 Pine weevil: correlograms showing AFLP markers 38 and 52 significantly associated with diurnal temperature range (DTR), sunshine duration (SUN), number of days of ground frost (FRS) and precipitation (PR).

With a confidence level of 99.99999% (ST = 1.139E-12), SAM identified 18 alleles at 12 loci to be significantly associated with at least one environmental variable (Fig. 5b). These alleles are involved in 173 significant models according to the Wald test (0.19% of the total number of models processed), and in 251 significant models according to the *G* test (0.28% of the total number of models processed). Among these 12 loci, five were detected by SAM and FDIST2 (HUJ616, OARFCB193, OARFCB304, MCM140, OARJMP29), also with a confidence level of 99% in FDIST2.

Focusing only on the most significant associations detected by SAM, with both tests, five alleles are associated with environmental variables with a ST set to 1.139E-17. These alleles were from four loci, SRCRSP9 (two alleles), DYMS1 (one allele), ILSTS28 (one allele) and OARFCB304 (one allele). Only the last of these was also detected by FDIST2 with a confidence level set to 99%.

Table 3 shows the detail of these five alleles and the environmental variables which gave the most significant associations found out in the 57 sheep breeds investigated. Important observations are that locus SRCRSP9 (detected by SAM only) has two alleles (134 and 118) associated with the higher number of environmental variables (three each for the most conservative test, with number of wet days for allele 134, and wind speed for allele 118). DYMS1 (detected by SAM only) associated with the number of wet days, has previously been shown to be involved in parasite resistance (Buitkamp *et al*. 1996). ILSTS28 (allele 127), detected by SAM only, was also associated with the number of wet days. Locus OARFCB304 (allele 171), which was

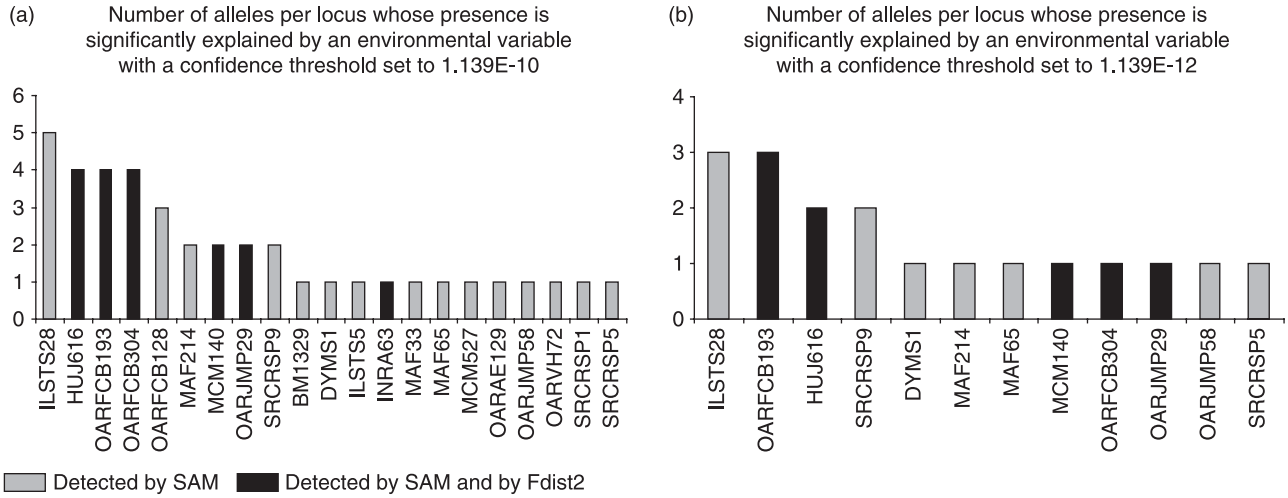

Fig. 5 (a) Histograms of candidate alleles in sheep: 40 alleles are significantly associated with at least one environmental parameter for a confidence level of 99.999% (significance threshold set to 1.139E-10), and with both Wald and *G* tests. The FDIST2 confidence level is set to 99%. (b) Histograms of candidate alleles in sheep: 18 alleles are significantly associated with at least one environmental parameter for a confidence level of 99.99999% (significance threshold set to 1.139E-12), and with both Wald and *G* tests. The FDIST2 confidence level is set to 99%.

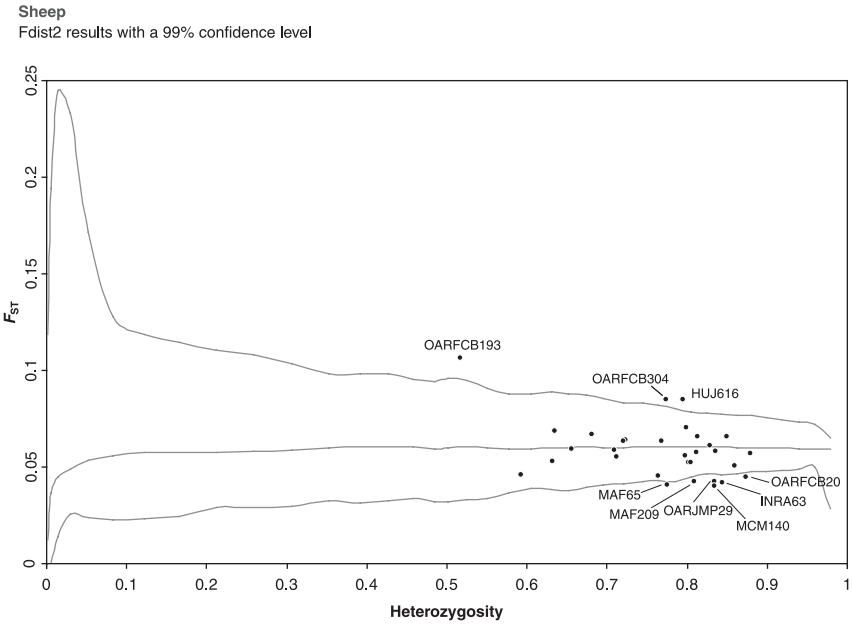

Fig. 6 Graphical representation of the results obtained with FDIST2 (Beaumont & Nichols 1996) for ECONOGENE sheep breeds, with a confidence level set to 99%. Three loci—all of which are detected by SAM with a significance threshold of 1.139E-14—possess values in excess of the neutral distribution. The six loci, which possess values lower than the neutral distribution, were also all detected by SAM: *OARFCB20* with a significance threshold (ST) of 1.139E-7, *MAF209* with a ST of 1.139E-8, *INRA63* with a ST of 1.139E-10, *MCM140* and *MAF65* with a ST of 1.139E-12 and finally *OARJMP29* with a ST of 1.139E-15.

detected by both methods, was associated with the quantity of precipitation.

Locus OARJMP29, detected by both SAM (with a ST of 1.139E-15) and FDIST2 (significance level of 99%) was a negative outlier to the neutral distribution of loci in FDIST2 (Fig. 6), and is therefore a candidate locus for balancing (Beaumont & Balding 2004) or stabilizing selection and is a locus that has been identified as being linked to disease resistance in a previous sheep study (Beh *et al*. 2002).

Discussion

For both the pine weevil and the sheep data, there was a good correlation between the results of SAM and those based on the population genetics approach chosen. Common statistical signals emerged from these analyses, which associated loci and environmental parameters, with these loci being clearly distinguishable from the theoretical distribution of neutral loci.

Table 3 Sheep breeds: the five alleles most likely to be under natural selection with the environmental variables involved in the corresponding most significant models. Values between brackets represent the number of variables involved in theses models. For example, on the first line, 'sunshine (4)' means that four sunshine monthly variables are significantly associated with allele SRCRSP9-134

Allele	*G* test: climatic variables	Wald Test: climatic variables	Monthly variable—Wald Test (more conservative)
SRCRSP9-134	No. of wet days (8), Rel. humidity (4), Sunshine (4)	No. of wet days (3)	January, October, November
DYMS1-181	No. of wet days (4)	No. of wet days (3)	March, September, Yearly mean
SRCRSP9-118	No. of wet days (1), Wind (3)	No. of wet days (2), Wind (1)	January, October, September
ILSTS28-127	Wind (1), Number of wet days (1)	No. of wet days (1)	October
OARFCB304-171	Precipitation (1)	Precipitation (1)	April

Pine weevil

In the pine weevil analysis, two loci were strongly detected by both approaches (loci 38 and 52). Although these two loci are not in linkage disequilibrium (Conord *et al*. 2006) and are therefore independently inherited, they varied in opposite directions with diurnal temperature range (DTR): AFLP fragment 38 decreased whereas 52 increased in abundance with increasing DTR (Fig. 4a, b). Increasing DTR is generally an indicator of harsh, continental climatic conditions, which here appeared to exert a selective pressure on at least two independent regions of the weevil genome. Moreover, the presence of AFLP fragment 38 was also associated with locations with a higher number of days of ground frost (Fig. 4e), and decreased with increasing precipitation (Fig. 4d); that is, its presence increased with drought, and cold and drought tolerance involve similar physiological processes in insects (Sinclair *et al*. 2003). Because both the larval and the adult stages overwinter under the soil, the number of days with ground frost is likely to strongly affect survival. Precipitation may also represent a more indirect selective pressure on larval survival through its effect on soil moisture and on the quality of the resource of dead wood, which may be more accessible for larval feeding under wet climatic conditions. Finally, the presence of fragment 38 decreased at locations with lower sunshine duration (Fig. 4c), a factor that has been shown to constrain adult activity (Nordlander *et al*. 2003).

The weevil analysis also shows evidence for discrepancies between the two approaches: two loci (68 and 10) strongly detected by the DFDIST method were not detected by SAM, whereas one locus (63) strongly detected by SAM was not detected by DFDIST. However, excluding 20 individuals (from Landes) out of 367 in the DFDIST analysis substantially changed the outcome of the analysis, allowing the detection of three fragments, 38, 52 and 63, also detected by SAM. Unlike all other sampled sites mainly composed of Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*), the Landes site is unique in being a monoculture of *Pinus pinaster*, at the southern edge of the pine weevil's geographical range, where both genetic drift and/or host-specific adaptation may occur. Moreover, only 20 individuals were sampled in Landes, representing the smallest group of our DFDIST analysis, and this relatively small sample may not adequately represent the genetic diversity present in Landes.

Because DFDIST first estimates allelic frequencies, and from that estimation, the genetic distances among samples (F_{ST}), it is strongly dependent on sample size, whereas SAM, being individual-centred is independent from local sample size effects. Therefore, for this data set, DFDIST appears to be more sensitive both to site-specific effects (genetic drift and/or local selection) and to sample size, than the SAM approach, perhaps an unsurprising finding given the fact that DFDIST assumes neutrality to conform to expectations under an island model to explain patterns of genetic diversity and differentiation.

Sheep

When the sheep dataset was analysed, a highly conservative approach to assessing the statistical power of the analysis was taken and here we have restricted our comments to a few examples, which follow. One striking example is an allele at locus DYMS1, which was associated with precipitation levels using SAM. Previously, this locus has been shown to be linked to parasite resistance in Scottish Blackface sheep, one of the breeds in this study and which is found in the highest precipitation environments (Buitkamp *et al*. 1996). In this study, the authors found that class I and class II major histocompatibility complex alleles were associated with faecal egg counts following natural, predominantly *Ostertagia circumcincta* infection. However, it is not clear why this locus is associated with disease resistance, since no further data are available which implicate this linkage group with disease resistance phenotypes. This locus was not detected as an outlier using FDIST (population genetic) analysis. Disease resistance is also implicated in the outlier

allele at locus OARJMP29, detected by both SAM and FDIST. OARJMP29 is the only locus tested here that is located on sheep chromosome 24. OARJMP29 has previously been shown to be linked to disease resistance (Beh *et al*. 2002), and was a negative outlier (i.e. potentially under stabilizing selection) in the FDIST analysis—this is an intriguing observation since disease resistance is a QTL which is sometimes associated with overdominant (heterozygous advantage) selection.

Environmental parameters included in the analysis

SAM is an exploratory method and is not an ideal tool to be used in the context of hypothesis-deductive reasoning. It executes a scan of all potential associations between all available environmental and molecular information, and thus permits investigation of data sets without being restricted to the choice of a given number of environmental variables corresponding to one or two working hypotheses. This has the advantage of preventing the user from missing important unexpected relationships. While being exhaustive and thus often dealing with a large amount of data, the method is also capable of rapidly highlighting significant relationships according to the significance level criteria set by the user in a computationally efficient way.

Non-independence of environmental parameters is a potential problem to which the user has to be attentive; however, sometimes slight differences (in temperature, in quantity of precipitation, etc.) may be sufficient to trigger associations to exist. This can be observed, for instance, with monthly environmental data used in the sheep case study where a given number of monthly environmental variables may be associated with a given allele while the other months for the same variable may not.

Finding accurate environmental data sets constitutes the main constraint on SAM. The scale of the study will determine the data available. On a local or regional scale, it is usually easier either to find existing environmental data, or to record data with specialized devices (temperature, precipitation, altitude, sunshine, etc.). On larger geographical scales, it is necessary to rely on national or international environmental data sets. At this latter level, the pine weevil case study is a good example to illustrate the use of free available data such as the SRTM30 digital elevation model described above, which has a resolution of about 90 meters, or like the climatology provided by the Climatic Research Unit in Norwich (New *et al*. 2002). With the exception of case studies for which environmental parameters can be measured on the field, the drawback of the SAM related to these elements is that it is necessary for researchers either to acquire the necessary skills in GIS to manage environmental data and to produce the matrices for computation, or to resort to GIS scientists to produce the data sets.

SAM and sampling

SAM implies different rules for sampling than standard analyses in population genetics. The goal is to obtain a statistically representative number of individuals per type of landscape, and not per population. Species, breeds or populations can be taken into consideration later in the analysis for comparisons, but it is possible to examine a given environmental factor on several organisms independently of the concept of population membership. Using SAM, it is more useful to sample individuals in a diversity of landscapes, in order to highlight a number of potential associations between loci and environmental parameters. Constraints lie in the recording of precise geographical coordinates of samples. It is possible, but rather difficult, to re-assign locations to previous data sets collected in the field without having recorded geographical coordinates. It means that sampling has to be planned with the intention of using SAM. Indeed, the goal is to be able to describe the area where a studied species occurs as accurately as possible with environmental data. Thus on the one hand, a precise location is required, and on the other hand, as high a resolution as possible of environmental data will allow the user to obtain the best models. The resolution of environmental data is the determining factor in the building of models as it is not useful having recourse to very high precision coordinates while environmental parameters characterize a wide area.

This version of SAM a priori better suits studies on large geographical scales, with data coming from several independent regions. Indeed, on finer scales results can be influenced by spatial autocorrelation. Further improvements are planned to develop a SAM that explicitly takes into account spatial structures into the ecological models, based on the raw-data or the matrix approaches described by Legendre (1993).

SAM and demography

One of the assumptions of the current population genetics approach is that all studied populations are at demographic equilibrium, e.g. that locus specific effects (i.e. outliers) are due to selection only. If a neutral mutation appears in an expanding population, while the remaining populations are at their demographic equilibrium, it may be detected as an outlier by the DFDIST approach. In contrast, if this particular mutation is not linked to environmental parameters (i.e. is neutral), it will not be detected by the SAM approach. Therefore, the latter is more robust than the population genetics approach to detect adaptive loci when populations are not at equilibrium.

Extensive conifer plantations in western Europe go back 200 years (Agnoletti & Anderson 2000) and this recent range expansion of the host plant together with modern

harvesting techniques (e.g. clear cutting that favours local pullulation of this stump-feeding insect) suggests that weevil populations are probably not at demographic equilibrium, and that colonization events followed by population expansion are frequent. However, using simulations, Beaumont & Nichols (1996) have tested the effect of different demographic scenarios (e.g. island vs. colonization models) on the power of DFDIST to detect outliers and found that these different models produce indistinguishable F_{ST} distributions when F_{ST} is not too high (less than 0.5). Average F_{ST} across weevil populations is 0.03 (Conord *et al*. 2006) so that the difference observed between DFDIST and SAM is unlikely to be due to populations in a nonequilibrium situation. The only parameter that was shown to produce skewed F_{ST} distributions in DFDIST, and consequently a loss of power in outlier detection, is heterogeneous rates of migration across populations. However, to obtain skewed distributions, Beaumont & Nichols (1996) simulated two groups of populations with extremely different F_{ST} (0.048 and 0.67, respectively). Across weevil populations, pairwise F_{ST} ranges from 0.02 to 0.09. Similarly, the demographic structure of many sheep populations and breeds is not at drift/mutation equilibrium (with genetic drift predominating). F_{ST} among sheep populations in the present study were higher than in pine weevils, with a mean of 0.124 but never achieved values close to 0.5.

Identifying the nature of a potential selective pressure

In comparison with standard methods of detecting signatures of natural selection, SAM has the main characteristic of identifying a potential environmental selection pressure. The outliers DFDIST and FDIST2 detect are believed to be involved in selection processes. While this inference is clear (Beaumont & Nichols 1996; Beaumont & Balding 2004), the effective selective pressure cannot be identified. This is the gap that the SAM can fill as shown by the results of this study. The knowledge of the nature of the selective pressure is likely to favour the emergence of working hypotheses about the role of the regions of the genome which are linked to the analysed markers, and thus to facilitate the hunting for genes involved in specific functional processes. It will then be necessary to take into account the fact that the distance from the linked site to the actual selected gene will vary depending on the recombination rate and time since selection (Wiehe 1998).

Moreover, SAM is independent of any theoretical genetic model and does not have constraints such as Hardy–Weinberg equilibrium assumption for example, which is far from being verified when considering markers involved in selection. Furthermore, the method is simple, fast, and provides results in agreement with population genetics models, with information about the nature of the selective pressure in addition. But this has to be balanced by the time required to find and prepare environmental databases. The option for mutual cross-validation with DFDIST/FDIST2 confirms loci detected by SAM and *vice versa*, with additional information on the nature of the potential selective pressure is another strength of this approach.

Working with codominant, multiallelic markers, a notable advantage of SAM is the statistical power to detect—among more than two alleles—the association of a particular allele (e.g. at the locus SRCRSP9 in sheep, see Table 3) with a given environmental parameter. It is clear that while such an indication remains a statistical association, circumspection is required about the potential implication of such an allele in adaptation, unless analysis is carried out on a locus known to be involved in adaptive processes and linkage relationships are established with the 'neutral' allele in question.

Notwithstanding the above, comparison with results provided by the population genetics approach is important to highlight the strongest signatures of selection when environmental parameters are not precisely targeted, and these were implicated in a large number of loci detected by SAM. The population genetics approach acts as a milestone of significance and potentially as a form of theoretical authentication of the results provided by SAM, until a sufficient number of applications carried out in parallel with DFDIST/FDIST2 and SAM allow further validation of this new method. Finally, DFDIST/FDIST2 will, in some cases, still be necessary to differentiate the type of selection markers are submitted to (divergent/direction vs. balancing/overdominant).

Towards landscape genomics

Enabling such simultaneous processing of a very large amount of data, the method is well suited to the population genomics approach, carrying out genome-wide tests to identify markers associated with environmental variables. It is designed to support the processing of hundreds of molecular markers against hundreds of environmental parameters. It is a new tool permitting to identify not only loci, but also specific ecological parameters which will help to interpret the role specific regions of the genome may play, likely to improve our understanding of the genetic mechanisms of evolution.

Acknowledgements

We would like to thank Guy Lempérière, Christina Peter, Trinidad Perez, Evren Koban and Erika Baus for sampling, Christian Miquel, Ludovic Gielly and Delphine Rioux for technical assistance, François Pompanon and Mark Beaumont for help with the data analysis, Godfrey Hewitt, Paolo Ajmone Marsan and four anonymous reviewers for advice on the manuscript. S. Joost was funded by the European Commission (Quality of Life, Contract QLK5-CT-2001-02461, 'ECONOGENE' http://www.econogene.eu).

A. Bonin and C. Conord were funded by the French Ministry of Research.

References

Agnoletti M, Anderson S (2000) *Methods and Approaches in Forest History*. CABI Publishing and the International Union of Forestry Research Organizations, Wallingford.

Agresti A (1990) *Categorical Data Analysis*. John Wiley and Sons, New York.

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 of London Series B-Biology Sciences*, **263**, 1619–1626.

Beh KJ, Hulme DJ, Callaghan MJ *et al.* (2002) A genome scan for quantitative trait loci affecting resistance to *Trichostrongylus colubriformis* in sheep. *Animal Genetics*, **33**, 97–106.

Benham J, Jeung JU, Jasieniuk M, Kanazin V, Blake T (1999) GENOGRAPHER: a graphical tool for automated fluorescent AFLP and microsatellite analysis. *Journal of Agricultural Genomics*, **4**, 399.

Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, **46**, 441–469.

Bruford MW (2004) Conservation of UK livestock: from molecules to management. In: *Farm Animal Genetic Resources* (eds Simm G, Villanueva B, Townsend S), pp. 151–169. Nottingham University Press, UK.

Buitkamp J, Filmether P, Stear MJ, Epplen JT (1996) Class I and class II major histocompatibility complex alleles are associated with faecal egg counts following natural, predominantly *Ostertagia circumcincta* infection. *Parasitology Research*, **82**, 693–696.

Clutton-Brock J (1999) *A Natural History of Domestic Animals*. Cambridge University Press, Cambridge, UK.

Conord C, Lempérière G, Taberlet P, Després L (2006) Genetic structure of the forest pest *Hylobius abietis* on conifer plantations at different spatial scales in Europe. *Heredity*, **97**, 46–55.

Conte E, de Maio A (2003) Distributed target detection in compound-Gaussian noise with Rao and Wald tests. *IEEE Transactions on Aerospace and Electronic Systems*, **39**, 568–582.

Day KR, Nordlander G, Kenis M, Halldorson G (2004) General biology and life cycles of bark weevils. In: *Bark and Wood Boring Insects in Living Trees in Europe, a Synthesis* (eds Lieutier F, Day KR, Battisti A, Grégoire JC, Evans HF). Kluwer Academic, Dordrecht.

Erickson DL, Fenster CB, Stenoien HK, Price D (2004) Quantitative trait locus analyses and the study of evolutionary process. *Molecular Ecology*, **13**, 2505–2522.

FAO (1998) *Secondary Guidelines for Development of National Farm Animal Genetic Resources Management Plans: Measurement of Domestic Animal Diversity (Modad): Recommended Microsatellites Markers*. Food and Agriculture Organization, Rome.

Ford MJ (2002) Applications of selective neutrality tests to molecular ecology. *Molecular Ecology*, **11**, 1245–1262.

Garcia LV (2004) Escaping the Bonferroni iron claw in ecological studies. *Oikos*, **105**, 657–663.

Goodchild MF (1996) Geographic information systems and spatial analysis in the social sciences. In: *Anthropology, Space, and Geographyraphic Information Systems* (eds Aldenderfer M, Maschner HDG), pp. 214–250. Oxford University Press, New York.

Hauck WW, Donner A (1977) Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, **72**, 851–853.

Hosmer DW, Lemeshow S (2000) *Applied Logistic Regression*. John Wiley & Sons, New York.

Legendre P (1993) Spatial autocorrelation—trouble or new paradigm. *Ecology*, **74**, 1659–1673.

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.

MacCallum C, Hill E (2006) Being positive about selection. *PloS Biology*, **4**, 293–295.

MacCullagh P, Nelder JA (1989) *Generalized Linear Models*. Chapman & Hall/CRC, London.

Mackay TFC (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics*, **35**, 303–339.

Menard S (2002) *Applied Logistic Regression Analysis*. Sage Publications, Thousand Oaks, California.

Montgomery GW, Sise JA (1990) Extraction of DNA from sheep white blood cells. *New Zealand Journal of Agricultural Research*, **33**, 437–441.

Narum SR (2006) Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conservation Genetics*, **7**, 783–787.

New M, Lister D, Hulme M, Makin I (2002) A high-resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nordlander G, Orlander G, Langvall O (2003) Feeding by the pine weevil *Hylobius abietis* in relation to sun exposure and distance to forest edges. *Agricultural and Forest Entomology*, **5**, 191–198.

Peter C, Bruford M, Perez T, Dalamitra S, Hewitt G, Erhardt G and the ECONOGENE Consortium (2007) Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds. *Animal Genetics*, **38**, 37–44.

Phillips PC (2005) Testing hypotheses regarding the genetics of adaptation. *Genetica*, **123**, 15–24.

Ryder ML (1983) *Sheep and Man*. G. Duckworth, London.

Shaffer JP (1995) Multiple hypothesis-testing. *Annual Review of Psychology*, **46**, 561–584.

Sinclair BJ, Vernon P, Klok CJ, Chown SL (2003) Insects at low temperatures: An ecological perspective. *Trends in Ecology and Evolution*, **18**, 257–262.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

von Sydow F, Birgersson G (1997) Conifer stump condition and pine weevil (*Hylobius abietis*) reproduction. *Canadian Journal of Forest Research*, **22**, 1254–1262.

Thorpe KV, Day KR (2002) The impact of host plant species on the larval development of the large pine weevil *Hylobius abietis* L. *Agricultural and Forest Entomology*, **4**, 187–194.

Tu W, Zhou Z-H (1999) A Wald test comparing medical costs based on log-normal distributions with zero valued costs. *Statistics in Medicine*, **18**, 274D–2761.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vos P, Hogers R, Bleeker M *et al*. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Wiehe T (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus: Hitchhiking of microsatellites. *Theoretical Population Biology*, **53**, 272–283.

Wilson AJ, Pemberton JM, Pilkington JG *et al*. (2006) Environmental coupling of selection and heritability limits evolution. *Plos Biology*, **4**, 1270–1275.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. *Molecular Biology and Evolution*, **22**, 506–519.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, **8**, 907–913.

Stéphane Joost is a quantitative geographer, postdoctoral fellow interested in the application of geographical information science to molecular ecology. Aurélie Bonin is a postdoctoral fellow whose research focuses on understanding the genetic basis of adaptation in various biological models. Mike Bruford is interested in the conservation genetics of wild and domesticated animal populations. Laurence Després teaches evolutionary ecology, population genetics and phylogeny at the University of Grenoble, France. Her main research interests are in speciation patterns and underlying evolutionary processes in plant-insects interactions. Cyrille Conord is a apostdoctoral fellow whose interest focuses on landscape genetics and more specifically on forest ecosystems from an evolutionary perspective. Georg Erhardt is full professor at the Justus-Liebig-University of Giessen, Germany, and his research interests include disease resistance, conservation genetics and genomics in livestock species. Pierre Taberlet is the Director of the Laboratory of Alpine Ecology (LECA) in Grenoble, France, and focuses on the conservation genetics and molecular ecology of many different plant and animal species.