RESOURCES
Molecular Ecology Resources (2016) doi: 10.1111/1755-0998.12629

High performance computation of landscape genomic models including local indicators of spatial association

S. STUCKI,* P. OROZCO-TERWENGEL,† B. R. FORESTER,‡ S. DURUZ,* L. COLLI,§ C. MASEMBE,¶ R. NEGRINI,§,** E. LANDGUTH,†† M. R. JONES,†† THE NEXTGEN CONSORTIUM,‡‡ M. W. BRUFORD,† P. TABERLET§§,¶¶ and S. JOOST*

*Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland, †School of Biosciences, Cardiff University, Sir Martin Evans Building, Cardiff CF10 3AX, UK, ‡Nicholas School of the Environment, University Program in Ecology, Duke University, Durham, NC 27708, USA, §BioDNA - Centro di Ricerca sulla Biodiversità e sul DNA Antico, Istituto di Zootecnica, Università Cattolica del S. Cuore, via E. Parmense 84, 29100 Piacenza, Italy, ¶Department of Zoology, Entomology and Fisheries Sciences, College of Natural Sciences, Makerere University, Box 7062, Kampala, Uganda, **Associazione Italiana Allevatori, 00161 Roma, Italy, ††Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA, ‡‡http://nextgen.epfl.ch, §§Laboratoire d'Ecologie Alpine (LECA), CNRS, Grenoble 38000, France, ¶¶Laboratoire d'Ecologie Alpine (LECA), Univ. Grenoble Alpes, Grenoble 38000, France*

Abstract

With the increasing availability of both molecular and topo-climatic data, the main challenges facing landscape genomics – that is the combination of landscape ecology with population genomics – include processing large numbers of models and distinguishing between selection and demographic processes (e.g. population structure). Several methods address the latter, either by estimating a null model of population history or by simultaneously inferring environmental and demographic effects. Here we present SAMβADA, an approach designed to study signatures of local adaptation, with special emphasis on high performance computing of large-scale genetic and environmental data sets. SAMβADA identifies candidate loci using genotype–environment associations while also incorporating multivariate analyses to assess the effect of many environmental predictor variables. This enables the inclusion of explanatory variables representing population structure into the models to lower the occurrences of spurious genotype–environment associations. In addition, SAMβADA calculates local indicators of spatial association for candidate loci to provide information on whether similar genotypes tend to cluster in space, which constitutes a useful indication of the possible kinship between individuals. To test the usefulness of this approach, we carried out a simulation study and analysed a data set from Ugandan cattle to detect signatures of local adaptation with SAMβADA, BAYENV, LFMM and an F_{ST} outlier method (FDIST approach in ARLEQUIN) and compare their results. SAMβADA – an open source software for Windows, Linux and Mac OS X available at http://lasig.epfl.ch/sambada – outperforms other approaches and better suits whole-genome sequence data processing.

Keywords: environmental correlations, genome scans, high performance computing, landscape genomics, local adaptation, spatial autocorrelation

Received 7 April 2015; revision received 5 August 2016; accepted 19 September 2016

Introduction

In the 1970s, several studies reviewed by Hedrick *et al.* (1976) implemented gene–environment associations to correlate the frequency of alleles with an environmental variable to look for signatures of selection (see also Mitton *et al.* 1977). Thirty years later, Joost *et al.* (2007, 2008) developed the concept to allow simultaneous processing of large numbers of logistic regressions to accommodate the increasingly larger numbers of molecular markers in use since the introduction of PCR (e.g. ALFPs, microsatellites). Since then, correlative approaches have been used in parallel with population genetics outlier-detection methods (e.g. Beaumont & Nichols 1996; Vitalis *et al.* 2003; Foll & Gaggiotti 2008) as cross-validation (e.g. Jones *et al.* 2013; Henry & Russello 2013) to detect signatures of local adaptation, that is a region of the geographic landscape where a particular genetic variant occurs at higher frequency and is correlated with an

Correspondence: Stéphane Joost, Fax: +41216935790; E-mail: Stephane.Joost@epfl.ch

environmental variable, potentially reflecting the higher fitness it confers to its carriers in that region (see a review in Vitti *et al.* 2013). Even though this kind of approach is still in vogue (Colli *et al.* 2014; Lv *et al.* 2014), there has been a recent revival in the interest of developing new statistical approaches for landscape genomics for use with genome-scale data sets, as such analyses enable the inference of environmental drivers of selection (Coop *et al.* 2010; Frichot *et al.* 2013; Günther & Coop 2013; Guillot *et al.* 2014; Frichot & François 2015; Gautier 2015; de Villemereuil & Gaggiotti 2015). For example, BAYENV (Günther & Coop 2013) implements a Bayesian method to compute correlations between allele frequencies and ecological variables taking into account differences in sample sizes and population structure. LFMM (Frichot *et al.* 2013; Frichot & François 2015) estimates the influence of population structure on allele frequencies by introducing unobserved variables as latent factors, while SGLMM (Guillot *et al.* 2014) extends the approach of Coop *et al.* (2010) by rooting it in a spatially explicit model and by implementing inference by means of the Integrated Nested Laplace Approximation and Stochastic Partial Differential Equation (SPDE) computational framework. Recently, Gautier (2015) introduces BayPass elaborating on the BAYENV model to capture some linkage disequilibrium information, among other important improvements, while de Villemereuil & Gaggiotti (2015) present BAYESCENV, an F_{ST}-based genome-scan method, which takes into account environmental differentiation between populations. It is based on the Beaumont & Balding's (2004) *F* model and similarly as implemented on BAYESCAN (Foll & Gaggiotti 2008), it considers that genetic variation at a given locus is affected by demographic processes that affect the entire genome (e.g. population expansions), selective events that change the allele frequencies at the locus as a response to an environmental variable (e.g. local adaptation to high temperature), and additional effects unrelated to the environmental variable tested. These methods aim at distinguishing between the effects of selection and those of demographic history; however, the increasing availability of large genomic data sets, has increased the computational intensity of this problem. In parallel, the geographic coordinates of samples are becoming frequently collected during field campaigns, enabling the computation of spatial statistics to shed an independent light on the interaction of selection and demographic signals.

Here we present the software SAMβADA, an extension of MATSAM (Joost *et al.* 2008), which offers an open source multivariate analysis framework to detect signatures of local adaptation in large-scale population genomics data sets. SAMβADA focuses on high performance computing to process whole-genome data and includes spatial statistics that measure indices of spatial autocorrelation to account for underlying patterns of spatial association in the data set due to population structure. The program is illustrated using two case studies: one in 5000 diploid individuals simulated for 100 SNPs in a heterogenous landscape, and the other one in 813 *Bos taurus* and *Bos indicus* individuals in Uganda genotyped for ~40 000 SNPs. Lastly, SAMβADA's performance is compared with other state-of-the-art software programs to detect signatures of selection.

Materials and methods

This section first presents SAMβADA's approach and implementation, with an overview of the accompanying modules. The second part introduces two case studies using simulation and a data set from Ugandan cattle, and how these data were collected and prepared for the subsequent analyses.

SAMβADA's approach

SAMβADA provides a locus-based approach to study local adaptation in a set of polymorphic markers using genome–environment associations. It aims at determining whether each investigated molecular marker is selected by one or a set of specific environmental variables (e.g. while multiple loci may be selected by the same environmental variable, it is also possible that different loci are affected by different environmental variables). As the analysis is performed independently for each locus, the number of possible combinations grows quickly with the size of both molecular (i.e. number of markers) and environmental data sets (i.e. number of variables) tested. To enable processing of large data sets, SAMβADA provides an automated procedure for selecting candidate loci associated with the environmental variables tested. For each locus, the set of predictor variables is kept parsimonious, because the main goal of the method is to detect which loci are potentially locally adapted rather than making predictions for the genotype of an individual based on its habitat. SAMβADA uses logistic regressions to model the probability of observing a particular genotype of a polymorphic marker given the environmental conditions at the sampling locations (Joost *et al.* 2007). As the state of a given genotype is considered as a binary presence/absence in each sample, SAMβADA can handle many types of molecular data (e.g. SNPs, indels, copy number variants and haplotypes), provided the user formats the input as required by SAMβADA and described in the software's documentation. Specifically, biallelic SNPs are recoded as three distinct genotypes (e.g. AA, AG and GG).

Univariate analysis. In the univariate case, each model involving a genotype and an environmental variable is compared with a constant model, in which the probability of the presence of the genotype is the same at each location in the landscape and is equal to its frequency in the data set. A maximum likelihood approach (Dobson & Barnett 2008) is used to fit the models. Significance is assessed with both log-likelihood ratio (G) and Wald tests (Joost *et al.* 2007). Bonferroni correction is applied for multiple comparisons (Bonferroni 1936; Shaffer 1995). To this end, the nominal significance threshold α is divided by the number m of hypotheses to be tested, that is the number of models that were fitted (e.g. if 10 000 SNPs are tested with five environmental variables, m = 150 000, as for each biallelic SNP there are three possible genotypes), to obtain the significance threshold α' ($\alpha' = \alpha/m$). The models having both P-values (computed from G and Wald scores) lower or equal to α' are considered as significant. To avoid numerous computations of P-values, the significance threshold α' is converted to a minimum score threshold using the quantile function of the χ^2 distribution. For each model, the property 'showing a score larger or equal to the score threshold' is equivalent to 'showing a P-value lower or equal to the threshold α''. Thus, the significance assessment can be performed directly on the scores.

In comparison with MATSAM (Joost *et al.* 2008), SAMβADA proposes several improvements: faster processing (see SAMβADA's implementation and Table S8, Supporting information), multivariate analysis and measures of spatial autocorrelation.

Multivariate analysis. In the multivariate approach, several environment variables can be used at the same time to model the presence of each genotype. In this case, the selection procedure is similar to a forward stepwise regression (Dobson & Barnett 2008) and is adapted to assess the significance of multivariate models. Both G and Wald tests refer to a null model to build the null hypothesis. The current model could be compared to the constant model (the same as in the univariate case) using multivariate χ^2 statistics. While rejecting the null hypothesis in this configuration would indicate that at least one parameter in the model is statistically significant, it would not provide information about which parameter (s) is relevant to the model. Therefore, SAMβADA assesses parameter significance in multivariate models with either a Wald test applied to each parameter separately (except the constant parameter) or with G tests excluding a parameter at a time: model selection is based on simpler models nested in the current one (see Supporting information).

Multivariate models allow the inclusion of pre-existing knowledge, provided the data constitutes a continuous variable. In particular, if population structure was analysed beforehand and can be represented as a coefficient of membership for each individual, this information can be included in the modelling. For models involving both an environmental variable and this coefficient, the selection procedure will assess whether the environmental variable is associated with the genotype while taking into account the possible effect of admixture. In case there are many ancestral populations, several coefficients may be included in the analysis.

Spatial autocorrelation. Beyond the detection of selection signatures, SAMβADA quantifies the level of spatial dependence in the distribution of each genotype. This measure of spatial autocorrelation refers to similarities or differences in genotypes occurrences between neighbouring individuals that cannot be explained by chance. Assessing whether geographic location has an effect on allele frequencies is especially important in landscape genomics, because statistical models assume independence between samples. Thus, if individuals with similar genotypes tend to concentrate in space, spurious correlations may co-occur with specific values of environmental variables. On the other hand, spatial independence of data strengthens the confidence in the detections. Spatial autocorrelation is a well-known concern (Legendre 1993) when investigating local adaptation, but few software allow its measurement [e.g. GEODA – Anselin *et al.* (2006) – or the libraries PySAL for PYTHON – Rey & Anselin (2010) – or SPDEP in R – Bivand & Piras (2015)].

SAMβADA measures the global spatial autocorrelation in the whole data set with Moran's I, as well as the spatial dependence of each point with local indicators of spatial association (LISA) (see Moran 1950; Anselin 1995 and see Sokal & Oden 1978 for application in biology). In practice, LISAs are computed by comparing the value of each point with the mean value of its neighbours as defined by a specific weighting scheme based on a kernel function (see Supporting information). The sum of LISAs on the whole data set is proportional to Moran's I (Anselin 1995). Both a spatially fixed kernel type relying on distance only and a varying kernel type considering the number of points can be used. SAMβADA includes three fixed kernels (moving window, Gaussian and bisquare) and a varying one (nearest neighbours). Significant spatial autocorrelation indices are determined based on an empirical distribution of the indices: for Moran's I, values (genotype occurrences) are permutated among the locations of individuals in the whole data set and a pseudo P-value is computed as the proportion of permutations for which I is equal to or more extreme (higher for a positive Moran's I or lower for a negative Moran's I) than the observed I. For LISA, the pseudo P-value is separately computed for each point (individual), by

keeping the individual of interest fixed and permuting the values of its neighbouring points with the rest of the data set.

samβada's implementation

samβada was developed as a standalone application written in C++, using the Scythe Statistical Library (Pemstein *et al.* 2011) which offers functions in matrix computation and probability distributions. samβada is distributed under an open source GNU General Public License to ease its use for research and teaching.

Desktop and high performance computing. When the development started, the estimations of computational load showed that it could prove difficult to both provide the new features described above and analyse whole-genome sequencing (WGS) data sets with a single computer. Thus, samβada is distributed with a module enabling High Performance Computing of large data sets.

Desktop version (samβada): samβada includes multivariate analyses and spatial autocorrelation computation. Many options are provided to facilitate formatting data and to customize analyses. For instance, the significance of models is assessed during the analysis and nonsignificant associations can be discarded on the fly. Moreover, models can be sorted out according to their scores before writing the results in order to facilitate their interpretation.

Parallel computing version (samβada and Supervision): To speed-up the analysis of large data sets, Supervision enables parallel processing with samβada by splitting data sets and merging results. The combination of samβada and Supervision makes it possible to analyse large data sets: (i) univariate logistic models identify candidate loci exhibiting selection signatures; (ii) these loci may be then investigated in the light of spatial autocorrelation measures and multivariate models. The former step may point out whether the observed correlation is due to similarities between neighbours, while the latter allows the inclusion of population structure, if any, in the model to assess the additional effect of the environmental variable after taking demography into account.

Modules. samβada includes several modules that enhance interfacing with other programs.

Geovisualization of spatial statistics: samβada provides an option to save spatial autocorrelation results as a shapefile (.shp), a common format for storing vector information in Geographic Information Systems (GIS). This feature relies on the shplib open source library (http://shapelib.maptools.org/), which is included and distributed with samβada.

Recoding molecular data: samβada is distributed with a utility for recoding molecular data into binary information, so that each genotype is considered on its own. Currently RecodePlink handles ped/map files, a standard format for SNP data used in genomics analysis (Purcell *et al.* 2007).

Supervision: For very large molecular data sets, samβada provides a module to share workload between computers. Supervision splits the input data in several files that can be processed separately, even on independent computers. At the end of an analysis, Supervision merges the results to provide the same output as if the whole data set had been processed at once. This module enables the processing of WGS data sets with samβada using a couple of desktop computers (see Table S9, Supporting information).

Alternative methods to detect selection

The performance of samβada was compared with other software for detecting signatures of selection. These analyses involved two other correlative approaches [bayenv – Coop *et al.* (2010) – and Latent Factor Mixed Models – Frichot *et al.* (2013); Frichot & François (2015)], and an F_{ST}-outlier-detection approach (Beaumont & Nichols 1996) included in arlequin 3.5 (Excoffier & Lischer 2010). Please note that these methods consider allele counts, whereas samβada recodes them into genotypes. An overview of bayenv, lfmm and arlequin is available in the supporting information.

Simulation study

As samβada and lfmm (Frichot *et al.* 2013; Frichot & François 2015) share a similar correlative approach, simulated data were used to compare their performance in scenarios where the selected loci are known. The analyses used a subset of the simulation data generated by Forester *et al.* (2016) who included lfmm in their work.

Simulated data. The simulations were run using the program cdpop v1.2 (Landguth & Cushman 2010), which models population genetic change across a landscape surface as a function of mutation, mating, gene flow, drift and selection. Each simulation had 5000 diploid individuals with 100 bi-allelic loci, one of which was subject to selection. All loci experienced a 0.0005 mutation rate per generation, free recombination and no physical linkage. Ten Monte Carlo (MC) replicates of each simulation were run for a total of 1250 generations, discarding the first 250 generations as burn-in (no selection imposed) to establish a spatial genetic pattern prior to initiating the landscape selection configurations.

The simulations used a discrete landscape selection configuration generated using the neutral landscape model QRULE (Gardner 1999) to simulate binary

landscape maps (1024 × 1024 pixels). Habitat fragmentation was controlled with the H parameter, which affects the aggregation of habitat pixels. A low value of H ($H = 0.1$) was used, resulting in less aggregated (more dispersed) habitat patches, and 10 landscape replicates were produced (one for each MC replicate) to average across stochastic variation among simulated landscapes. Discrete habitat types (type 'AA' or 'aa') represented habitat patches in which AA or aa genotypes were, respectively, favoured (see Fig. S3, Supporting information for an example of the landscape configuration).

The effect of varying selection strength was tested, mediated through density-independent (i.e. environment-driven) mortality (s) determined by genotypes of the selected locus. Selection strengths included $s = 0.01$ or '1%', $s = 0.05$ or '5%', and s = 0.10 or '10%'. AA individuals had no mortality in 'AA' habitat patches and experienced 1%, 5% or 10% mortality if they occurred in 'aa' patches. Individuals with 'aa' genotypes at the locus under selection experienced the opposite selection gradient. The Aa genotypes experienced uniform selection ($s/2$) across the entire surface.

Dispersal capacity for movement and mating was set to a maximum of 5% of the landscape surrounding an individual, with dispersal occurring once per generation. Mating pairs of individuals and dispersal locations of offspring were chosen based on a random draw from the inverse-square probability function of distance, truncated with the specified maximum distance. Mating parameters represented a population of unisexual individuals with females and males mating with replacement. The number of offspring produced from mating was determined from a Poisson distribution ($\lambda = 4$), which produced an excess of individuals each generation to maintain a constant population size of 5000 individuals at every generation. Carrying capacity of the simulation surface was 5000 individuals. Excess individuals were discarded once all 5000 locations became occupied, which is equivalent to forcing out emigrants once all available home ranges are occupied (Balloux 2001; Landguth & Cushman 2010). Combining the 10 landscape configurations and the three levels of selection strength, a total of 30 molecular data sets were analysed in this simulation study.

Simulation analysis. A set of 500 individuals were randomly selected from each simulation of 5000 individuals (the 500 individuals were chosen from the same position in the grid in each simulation and replicate) to carry out the selection analyses with SAMβADA and LFMM (see Fig. S3, Supporting information). Simulation data were filtered for a minimum allele frequency (MAF) of 1%; no simulation loci were found to have a MAF <1%. All analyses used three environmental predictor variables: the x-coordinate location of an individual ('x'), the y-coordinate location of an individual ('y') and the location of an individual in an AA or aa patch ('habitat'). Two types of analyses were run with SAMβADA: (i) Univariate analysis with the three environmental predictor variables; (ii) Multivariate analysis using the population structure to build the null models. For univariate analysis, the significance threshold was set to $\alpha' = 0.01/900$ (100 loci, three genotypes and three environmental variables) after Bonferroni correction. The second type of analyses was performed as follows for each replicate: Population structure was assessed with ADMIXTURE (Alexander *et al.* 2009) using the 99 neutral loci. ADMIXTURE (Alexander *et al.* 2009) estimates the maximum likelihood of individual ancestries from multilocus SNP genotype data sets and assumes that samples descend from a predefined number of ancestor populations that became mixed. ADMIXTURE estimates both the fraction of each sample coming from each population and the marker frequencies in these populations. The optimal number of populations K is assessed by a k-fold cross-validation procedure (see Table S4, Supporting information, for the value of K in each simulation). As the sum of the coefficients of admixture is 1.0 for each sample, only $(K - 1)$ values are required to specify the ancestry of each sample. Thus, $(K - 1)$ 'population variables' were created by computing a PCA on the coefficients of admixture and by taking the $(K - 1)$ first principal components. The set of predictor variables was composed by the three environmental variables ('x', 'y' and 'habitat') and the $(K - 1)$ 'population variables'. The $(K - 1)$ 'population variables' were used to compute a 'null model' including the population structure for each marker, and then, the models to be tested were built by adding one environmental variable to the set of 'population variables'. In the current implementation of SAMβADA, this is performed by computing all the models from 1 to K variables (i.e. the total number of clusters in the data) before extracting the models of interest. As the models to be tested included one variable more than their corresponding null model, the total number of models considered for the Bonferroni correction was the same as for the univariate analysis.

For LFMM, K was determined using the Patterson method (Patterson *et al.* 2006) as suggested by Frichot *et al.* (2013) for simulation studies (see Table S5, Supporting information, for the value of K in each simulation). LFMM models were run with the package LEA (v. 1.2.0; Frichot & François 2015) in R (v. 3.2.3; R Core Team 2016) using the following parameters: 10 000 iterations with a burn-in of 5000 iterations, and five replicate runs. The median z-score and P-value were chosen from each set of five runs; significant outliers were detected as those loci with a P-value <(0.001/300) after Bonferroni correction.

The significance thresholds α for SAMβADA and LFMM were estimated separately for each method.

For each of the three simulation scenarios, the following metrics were averaged across the 10 replicates: true-positive rate (TPR), false-positive rate (FPR) and a genotype–environment association index (GEA) that determines how effective a method is at identifying the predictor that is driving selection (Forester *et al.* 2016). The GEA index ranges from 3 (best performance) to 0 (worst performance) and is coded: 3 = correct identification of variable 'habitat'; 2 = 'habitat' is significant, but less than '*x*' or '*y*'; 1 = 'habitat' is not detected but '*x*' or '*y*' are; and 0 = no variable is detected as significantly associated with the locus under selection.

Ugandan cattle

In addition to the simulated data set, we illustrate the use of SAMβADA with an empirical data set of Ugandan cattle, which is composed of two main populations. Ankole (or Ankole-Watusi) cattle are a Sanga breed (taurine-zebu cross) that appeared in the Nile Basin around 2000 years BC. They migrated southward and are now found in southwest Uganda, Rwanda and Burundi (Ndumu *et al.* 2008; Ajmone Marsan *et al.* 2010). Shorthorn zebus were introduced in East Africa around the VIIIth century AD; they later spread as they were less affected than taurine and Sanga cattle by rinderpest, but their susceptibility to trypanosomiasis is presumed to have restrained their dispersion across Africa (Ajmone Marsan *et al.* 2010). Shorthorn zebus are now common in northeast Uganda and are being crossbred with Ankole cattle in the centre of the country.

Sampling design. In the context of the European Nextgen project (http://nextgen.epfl.ch), the sampling of Ugandan cattle was designed to cover the whole country, including each eco-geographic region, and to obtain a homogeneous geographic distribution of individuals across the country. To this end, a regular grid made of 51 cells of 70 × 70 km was produced. On average, four farms were visited in each cell and four unrelated individuals were selected from each farm, for a total of 917 biological samples retrieved from 202 farms. The sampling season took place between March 2011 and January 2012. Recorded information also included the location of the farm, the name of the breed, a picture and morphological information (e.g. withers height and horns length) for each individual. These elements were stored in a database accessible through a Web interface, enabling real-time monitoring of the sampling campaign.

Molecular data. Out of the 917 individuals, 813 samples were genotyped with a medium-density SNP chip (54 609 SNPs, BovineSNP50 BeadChip; Illumina Inc., San Diego, CA, USA). Only markers located on the autosomal chromosomes were considered in the analyses. The data set was filtered with PLINK (Purcell *et al.* 2007) with a call rate set to 95% for both individuals and SNPs, and a MAF set to 1%. The resulting data set after filtering contained 804 samples and 40 019 SNPs.

Population structure. Population structure was analysed with the software ADMIXTURE (Alexander *et al.* 2009) using a subset of 28 197 SNPs pruned for linkage disequilibrium as recommended in the manual. The SNPs were filtered with PLINK (option `–indep-pairwise`), $r^2 < 0.2$, sliding window of 10 SNPs, step size of 5 SNPs), and the number of populations *K* was chosen using the cross-validation index of ADMIXTURE. The best partition of the data set consisted of four populations, although the vast majority of the samples (96%) were allocated to one of two clusters on the basis of the ancestry coefficients (Fig. S1, Supporting information). Mapping these coefficients revealed that these two clusters (340 and 431 individuals of 804) occurred in the southwest and northeast of Uganda, respectively. Using pictures of sampled individuals, the first cluster was identified as Ankole cattle and the second one as zebu. These observations are in agreement with the known background of Ugandan cattle. The remaining two clusters (33 animals in total) possibly represent introgression from allochthonous gene pools. The results of the population structure analysis were used to define the parameters needed by each method to detect selection signatures.

Environmental data. Habitat characteristics of sampling locations were described with the WorldClim data set containing monthly values of precipitation, minimum, mean and maximum temperature as well as 19 derived variables, at 1 km resolution (Hijmans *et al.* 2005). This data set provides appropriate data as it consists of representative climate information collected during 30 years (WMO standard climate normal, Arguez & Vose 2010) and its high resolution suits the scale of our study. These environmental variables were originally stored in four tiles (portions of map) which were pasted using the Geospatial Data Abstraction Library (GDAL Development Team 2013) and a customized Python script. The topography is described by the 90 m resolution SRTM3 (Shuttle Radar Topography Mission) digital elevation model (DEM) (Farr *et al.* 2007). SAGA GIS (www.sagagis.org) was used to paste the 36 tiles covering the country and to derive slope and orientation from the SRTM DEM. Longitude and latitude were also taken into account as a rough proxy for population structure. Finally, the values of the 72 environmental variables were extracted for each sampling locality using the

'Point Sampling Tool' extension (http://hub.qgis.org/projects/pointsamplingtool) in QuantumGIS (www.qgis.org).

Variable selection for univariate analysis: Considering all environmental variables in the computation of the multiple logistic regressions would have provided a comprehensive analysis with a low risk of missing detections. Nonetheless, some variables are highly correlated; thus, the corresponding models for a genotype are likely to represent the same phenomenon. To lower the dependency between models and spare computation time, we used the variance inflation factor (VIF) to control for multicollinearity (Dobson & Barnett 2008). A maximum VIF of 5 was chosen, corresponding to a coefficient of correlation of 0.9 between pairs of variables. The number of variables was reduced iteratively by randomly removing one of the two most correlated variables until the maximum correlation was lower than the threshold (0.9). This procedure led to a set of 23 environmental variables that were used for univariate landscape genomic analyses (Table S1, Supporting information).

Variable selection for multivariate analysis: The multivariate analysis with SAMβADA consisted in bivariate models along with their corresponding univariate and constant models. A maximum of two explanatory variables were considered to ease the interpretation of their respective effects. Moreover, SAMβADA's conservative approach to assess model significance tends to reject models including numerous environmental variables. In this study, the multivariate models were used to take population structure into account. The information on population structure was derived from the analysis of individual ancestries. To this end, a new variable 'population structure' was defined by performing a principal component analysis (PCA) on the coefficients of ancestry and was used to represent the population structure in SAMβADA analyses (see 'Protocol of analysis' for details). It was thus added to the set of 23 environmental variables and the correlation-based variable selection method was reapplied to limit the coefficient of correlation between pairs of variables to 0.81, which corresponds to limiting the VIF to 2.9. On this basis, 15 predictor variables (including the 'population structure' variable) were considered for SAMβADA multivariate analysis (see Table S1, Supporting information).

Protocol of analysis. Four approaches were applied to detect selection signatures among the 40 019 SNPs from 804 samples. As SAMβADA processes each genotype independently, while BAYENV, LFMM and ARLEQUIN treat each locus as a whole, we defined a locus as 'detected' by SAMβADA if at least one of its three genotypes showed a significant association with an environmental variable. For BAYENV, LFMM and ARLEQUIN, the selection signatures are analysed per locus.

Data preparation: Since Ugandan cattle globally comprises two admixing populations (Fig. S1, Supporting information), the 33 samples from the two smaller populations were excluded from the analyses with SAMβADA and LFMM, leading to a set of 771 samples for these methods. To estimate whether the population structure could be efficiently summarized by the Ankole and zebu clusters, a PCA was run on the coefficients of ancestry for the subset of 771 samples taken from the results of ADMIXTURE for $K = 4$. The first principal axis of this PCA accounted for 95% of the variance among all molecular markers, so that a single coefficient is sufficient to provide an overall view of an individual's ancestry. Given this configuration, SAMβADA's multivariate analysis needed a single variable, that is the first axis of the PCA, to summarize the population structure. As the cattle population is essentially constituted of two clusters, the number of latent factors tested with LFMM covered a range of values of K that included the estimated K as described by Frichot & François (2015). This range consisted of values of K from $K = 1$ to $K = 4$. For BAYENV and ARLEQUIN, as these approaches require the samples to be clearly assigned to a population, the 804 samples were classified into populations based on their coefficient of ancestry and using a threshold of 0.85, below which samples were excluded from the analysis. This led to, respectively, three clusters of 162 Ankole cattle, 8 zebus and 10 cattle from the third population; samples from the fourth population were highly admixed and none satisfied the condition. This method was preferred over a classification based on sampling locations or phenotypic traits because Ugandan cattle are generally admixed (see Fig. S1, Supporting information). The univariate correlative approaches – SAMβADA, BAYENV and LFMM – used a selected set of 23 environmental variables, while SAMβADA multivariate analysis used a set of 15 environmental variables (see 'Environmental data' for details).

Computational set-up for correlative Bayesian approaches: BAYENV (v. 2.0, Coop *et al.* 2010; Günther & Coop 2013) first estimated the interpopulation covariance matrix with a run of 100 000 iterations over a set of 1000 loci selected at random among the loci identified as neutral by SAMβADA's univariate analysis. Then, the full data set was analysed for another 100 000 iterations to detect the signatures of selection. LFMM models were run with the package LEA (v. 1.4.0; Frichot & François 2015) in R (v. 3.3.0; R Core Team 2016) using the following parameters: 10 000 iterations with a burn-in of 5000 iterations, and five replicate runs for each value of the number of latent factors.

Models selection: The statistical significance threshold for SAMβADA, LFMM and ARLEQUIN was set to $\alpha = 0.01$ before

applying the Bonferroni correction. The analysis of SAMβADA's multivariate models followed the same protocol as its counter-part on the simulation data: the univariate models involving the 'population structure' variable were used as 'null models' for assessing the significance of bivariate models involving the 'population structure' variable and one environmental variable; all other models were discarded. For LFMM, the median *z*-score and *P*-value were chosen from each set of five runs. The number of latent factors was set to $K = 2$ based on the quantile – quantile (QQ) plots (see Fig. S2, Supporting information). For BAYENV, model selection was based on the Jeffreys' scale of evidence (Jeffreys 1961) and on the distribution of Bayes Factors (BF) for neutral loci (Coop *et al.* 2010). This distribution was estimated by selecting a random subset from the loci identified as neutral by SAMβADA. BAYENV's results were analysed separately for each environmental variable and models showing a BF higher than 10 (strong evidence) or higher than the 1st percentile of the neutral distribution (if higher than 10) were used to build the set of candidate loci.

Results

Results for the simulated data

Detection of selection signatures. Univariate models in SAMβADA show that on average both the TPR and the genome–environment association index (GEA index) increase with the strength of selection (see Table 1a and Table S3, Supporting information, for detailed results). TPR ranges from 60% for the weak (1%) selection, to 90% for intermediate (5%), and to 100% for strong selection (10%), while the GEA index takes the values of 0.7, 1.6 and 2.1 for the corresponding selection pressures. The FPR is high (43–45%) but consistent among the different scenarios. When population structure is taken into account using multivariate models, the TPR index and the GEA index decrease for the weak and intermediate levels of selection compared to the univariate models, but their values remain unchanged for the stronger level of selection, whereas the FPR decreases for all levels of selection (2–4%, see Table 1b and Table S4, Supporting information, for detailed results). Overall, LFMM behaved very similar to the SAMβADA univariate approach showing the same TPR and FPR and marginally better GEA values (Table 1c and Table S5, Supporting information, for detailed results).

Spatial autocorrelation. Spatial statistics were computed for one genotype per locus for each replicate of the three selection scenarios. The choice of the genotypes was based on SAMβADA's univariate models: for each locus, the genotype in the model with the highest *G* score was chosen to represent the locus in the subsequent analyses. Spatial autocorrelation was measured using Moran's *I*, and the spatial ponderation was based on the number of nearest neighbours. The weighting schemes included 5, 15, 30, 45 and 60 neighbours. The threshold of pseudo-*P*-values was set to 0.01 (99 permutations) for assessing the significance of global and local values of Moran's *I*. Figure 1 presents an overview of the correlograms obtained for each simulation scenario. For each scenario, the loci were ordered in three groups: loci under selection (L0), neutral loci detected by SAMβADA (i.e. false-positive detections) and neutral loci not detected by SAMβADA (i.e. true-negative detections). On average, the group of false positives shows a higher value of Moran's *I* than the group of true negatives. The loci under selection show values of Moran's *I* similar to the group of true negatives for the weak selection scenario, while their values of Moran's *I* tend to be higher than both groups of neutral loci for the intermediate and strong selection scenarios (see Table 1). The individual correlograms for each replicate of the three selection scenarios are found in Figs S4–S6, Supporting information.

Local indicators of spatial association were summarized for each locus by counting the number of sampling points showing a significant value. The amount of significant LISA points is generally higher for the locus under selection than the averaged values of each of the two groups of neutral loci (see central part of Fig. S6,

Table 1 Average true-positive rate (TPR), false-positive rate (FPR) and genotype–environment association index (GEA index) across the 10 replicates for each simulation scenario. All simulations use a dispersal level of 5% and a discrete landscape with an aggregation index *H* of 0.1. TPR scales from 0% (worst performance, locus under selection not detected) to 100% (best performance, locus under selection detected); FPR scales from 0% (best performance, no false detection) to 100% (worst performance, 99 neutral loci detected as significant); GEA index scales from 0 (worst performance, no detection) to 3 (best performance, correct detection). Results for (a) SAMβADA univariate models, (b) SAMβADA multivariate models taking into account the population structure, (c) LFMM

Selection (%)	TPR (%)	FPR (%)	GEA index
(a) SAMβADA univariate			
1	60	45	0.7
5	90	43	1.6
10	100	45	2.1
(b) SAMβADA multivariate			
1	10	4	0.1
5	50	2	0.5
10	100	2	2.1
(c) LFMM			
1	50	43	0.6
5	90	43	2.0
10	100	43	2.8

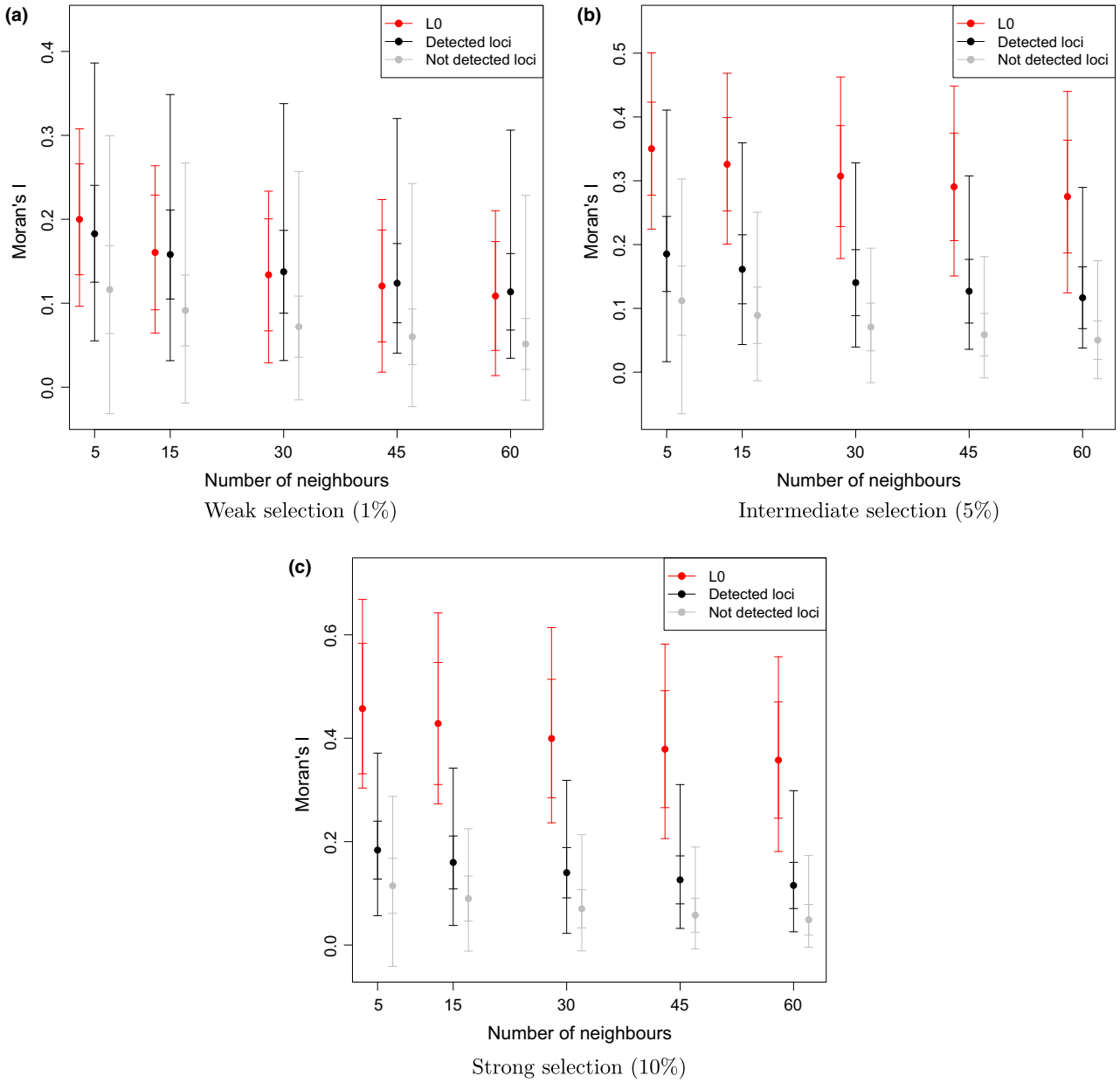


Fig. 1 Summary of correlograms computed for the simulation data. Spatial autocorrelation was measured using Moran's *I*, and the spatial ponderation was based on the number of nearest neighbours. The weighting schemes included 5, 15, 30, 45 and 60 neighbours. Each locus was represented by its genotype involved in the model with the highest *G* score. Each graph summarizes the correlograms of one of the selection scenario s: a) weak, b) intermediate, and c) strong selection. The loci were sorted in three groups: the loci under selection (L0 – red bars), the neutral loci detected by samβada (black bars) and the neutral loci not detected by samβada (grey bar). For each group, the averaged Moran's *I* is represented by the dot on the bar, the two marks above and below indicate the standard deviation and the outer bounds show the minimal and maximal values of Moran's *I* for this group. [Colour figure can be viewed at wileyonlinelibrary.com]

Supporting information). For the replicates where the locus L0 was detected by samβada's univariate models, all detected loci were ordered according to the decreasing number of significant LISA points. For the intermediate and strong selection scenarios, the locus L0 is often found among the first loci. For instance, L0 is found between positions 1 and 5 for the LISA computed with 15 neighbours in the intermediate selection scenario (see right part of Fig. S6, Supporting information).

Results for the Ugandan cattle

Detection of selection signatures. Using univariate models, samβada identified 2354 SNPs (5.9%) potentially subject

to selection, BAYENV 1169 (2.9%), LFMM 970 (2.4%) and ARLEQUIN did not identify any locus as significant. Among the 2354 loci detected by SAMβADA, 967 were <100 000 base pairs apart from another detected locus, suggesting that some loci may be detected simply due to physical linkage to selected regions. Figure 2 counts the number of common detections between landscape genomic approaches. SAMβADA's results partially match those of BAYENV with 214 common loci (i.e. 9% of SAMβADA' and 18% of BAYENV's detections). Concerning the third correlative approach, LFMM is more conservative than SAMβADA and the overlap is smaller because 79 loci (i.e. 3% of SAMβADA' and 8% of LFMM's detections) are detected by both SAMβADA and LFMM, while 24 loci (i.e. 2% of BAYENV's and 2% of LFMM's detections) are detected by both BAYENV and LFMM. However, 110 SNPs detected only by LFMM are <100 000 base pairs apart from loci detected by SAMβADA, potentially identifying the same selection signature. Lastly, ARLEQUIN's best results involved 17 SNPs with *P*-values lower than 10^{-4}. Although these results are not significant – the threshold corrected for multiple comparisons was $\alpha' = 2.5 \times 10^{-7}$ – it is interesting to compare them with the other methods. Among these 17 SNPs, one was common with SAMβADA, 16 were common with BAYENV and none with LFMM, suggesting that population-based methods, whether using outliers or environmental correlations, tend to overlap substantially in detecting selection signatures. Quantile – quantile (QQ) plots of SAMβADA and LFMM results are presented on Fig. S2 (Supporting information).

The loci detected by SAMβADA's univariate analysis with the highest *G* scores were compared among methods. Table 2 shows that BAYENV generally agreed with SAMβADA's detections, while LFMM's results differed. Some of the most significant loci detected by SAMβADA were ignored by LFMM. A total of eight loci were identified by the three correlative methods and four of them were among the most significant models detected by SAMβADA (see Table 2). Three of these SNPs occur close to each other on chromosome five.

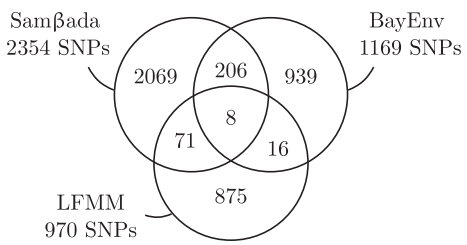


Fig. 2 Comparison of the selection signatures identified by the three landscape genomic approaches. The total number of SNPs detected by each method is indicated below the name. The diagram shows how these sets of SNPs overlap between methods.

SAMβADA's multivariate analysis identified 12 significant bivariate models, corresponding to 8 loci (see Table S2, Supporting information). In SAMβADA's framework, this means that these models involving one environmental variable and the variable 'population structure' provided a significantly more accurate estimation of the genotype's frequency than their univariate parent involving the variable 'population structure' only. Therefore, although population structure might partly explain the distribution of these genotypes, adding an environmental variable provided a significantly more accurate estimation of their distribution ($\alpha' = 5.9 \times 10^{-9}$). The loci detected by SAMβADA's multivariate analysis include three loci that were detected by all correlative approaches (Hapmap28985-BTA-73836, ARS-BFGL-NGS-106520 and BTA-73842-no-rs, see lines 7, 8 and 9 in Table 2).

Computation time was measured for the three correlative approaches using a desktop computer with 8-core CPUs at 4.0 GHz and 16 Gb of RAM, except for BAYENV, which used a slightly less powerful computer (8-core CPU at 3.1 GHz and 8 Gb of RAM). SAMβADA analysed the univariate models within 1.5 h using a single processing thread and both univariate and bivariate models in 2.6 h using four threads. LFMM analysed the data set in 26.9 h for each value of *K* using five threads (one per run) and BAYENV in 41.3 h with a single thread, for one run. Ratios between computation times tend to increase with larger data sets (see Table S7, Supporting information).

Spatial autocorrelation. Global and local indicators of spatial autocorrelation were computed for two genotypes with a weighting scheme based on the 20 nearest neighbours and a pseudo *P*-value threshold of 1%: (i) ARS-BFGL-NGS-46098 (genotype GG) (hereafter ARS-46 (GG)), which was detected by SAMβADA only with one of the highest *G* scores (Table 2, line 4), and (ii) Hapmap28985-BTA-73836 (genotype GG) (hereon HM-28 (GG)), which was detected by SAMβADA while the corresponding locus HM-28 was detected by BAYENV and LFMM (Table 2, line 7). SAMβADA identified isothermality, the stability of temperature across the year, as strongly associated with both genotypes. Figure 3 shows local indices of spatial autocorrelation for these two genotypes. On the one hand, ARS-46 (GG) was positively autocorrelated for the majority of points and the index was significant for half of them. Although the distribution of this genotype shows spatial dependence, nonsignificant associations were found at the edge of Lake Victoria and in a corridor in the North of the Lake with some occurrences in the West of Uganda. On the other hand, the local indices of spatial association of HM-28 (GG) showed lower values in general and were only significant in the

Table 2 List of SNPs detected by SAMβADA corresponding to the univariate models with the highest *G* scores. Loci are identified by their name, their chromosome and their position in million base pairs (Mbp). The following columns show whether SAMβADA (univariate), BAYENV and LFMM detected them with the corresponding environmental variables and *P*-values (SAMβADA, LFMM) or Bayes Factor (BAYENV). Loci in bold type are the common discoveries of SAMβADA univariate and bivariate, LFMM and BAYENV. Local indicators of spatial autocorrelation were analysed for SNPs on lines 4 and 7

			SAMβADA		BAYENV		LFMM	
Loci	Chr.	Pos (Mbp)	Env	*P*-value	Env	BF	Env	*P*-value
1. Hapmap41074-BTA-73520	5	48.35	prec7	48.35×10^{-47}	tmin10	136		
			latitude	1.41×10^{-43}	bio9	89.7		
			bio7	6.07×10^{-43}	prec6	74.2		
2. ARS-BFGL-NGS-113888	5	48.32	prec7	4.86×10^{-47}	tmin10	39.3		
			latitude	1.06×10^{-43}	bio9	27.6		
			bio7	1.26×10^{-42}	prec6	24.9		
3. Hapmap41762-BTA-117570	5	18.94	prec7	2.74×10^{-44}	bio9	15.3		
			latitude	3.95×10^{-41}	prec6	13.3		
			prec6	4.95×10^{-37}	prec5	12.6		
4. ARS-BFGL-NGS-46098	20	2.95	prec7	2.94×10^{-44}				
			latitude	2.58×10^{-39}				
			prec6	4.35×10^{-39}				
5. BTA-73516-no-rs	5	48.75	prec7	2.51×10^{-39}	bio9	12.8		
			latitude	4.57×10^{-36}	prec6	11.8		
			prec6	7.61×10^{-33}	prec5	11.5		
6. Hapmap41813-BTA-27442	5	49.04	prec7	6.06×10^{-39}	bio9	16.7		
			latitude	7.37×10^{-36}	prec6	15.3		
			prec6	2.26×10^{-32}	prec5	14.9		
7. Hapmap28985-BTA-73836	**5**	**70.34**	bio3	6.98×10^{-36}	bio9	12.5	bio3	4.01×10^{-19}
			prec6	1.18×10^{-35}	prec6	11.5	bio7	3.94×10^{-14}
			bio7	1.61×10^{-33}	prec5	11.1	latitude	6.63×10^{-10}
8. ARS-BFGL-NGS-106520	**5**	**70.2**	bio3	6.26×10^{-35}	tmin10	79.5	bio3	3.61×10^{-17}
			bio7	3.55×10^{-33}	bio9	23.3	bio7	1.18×10^{-12}
			latitude	1.13×10^{-31}	prec6	18.7	prec6	2.03×10^{-10}
9. BTA-73842-no-rs	**5**	**70.18**	bio3	8.95×10^{-34}	bio9	13.4	longitude	3.19×10^{-15}
			bio7	2.64×10^{-30}	prec6	11.3	prec6	1.35×10^{-9}
			latitude	4.13×10^{-30}	prec5	10.7	bio15	2.55×10^{-9}
10. Hapmap31863-BTA-27454	5	48.99	prec7	1.08×10^{-33}				
			latitude	3.00×10^{-30}				
			prec6	3.26×10^{-27}				
11. Hapmap50523-BTA-98407	5	46.74	prec7	6.36×10^{-32}	bio9	14.4		
			prec6	7.61×10^{-28}	prec6	12.8		
			latitude	9.69×10^{-28}	prec5	12.3		
12. BTB-01400776	20	2.7	prec7	4.71×10^{-31}				
			latitude	5.23×10^{-30}				
			prec6	1.65×10^{-25}				
13. ARS-BFGL-NGS-10586	2	128.64	latitude	9.47×10^{-29}	bio9	11.5		
			bio7	1.73×10^{-25}	prec6	10.1		
			prec7	1.81×10^{-25}				
14. Hapmap23956-BTA-36867	15	47.2	latitude	1.59×10^{-28}	bio9	23.1		
			prec7	2.17×10^{-26}	prec6	20		
			prec6	8.85×10^{-25}	prec5	19		
15. ARS-BFGL-NGS-94862	11	103.53	longitude	1.23×10^{-27}	bio9	45.6	longitude	9.52×10^{-10}
			prec7	1.26×10^{-22}	prec6	42.1		
			latitude	4.26×10^{-20}	prec5	40.8		
16. BTA-122374-no-rs	14	16.44	latitude	1.97×10^{-27}				
			prec7	1.05×10^{-23}				
			prec11	1.26×10^{-23}				
17. ARS-BFGL-NGS-43694	5	49.65	prec7	8.16×10^{-27}				
			latitude	3.41×10^{-25}				
			prec6	5.93×10^{-24}				

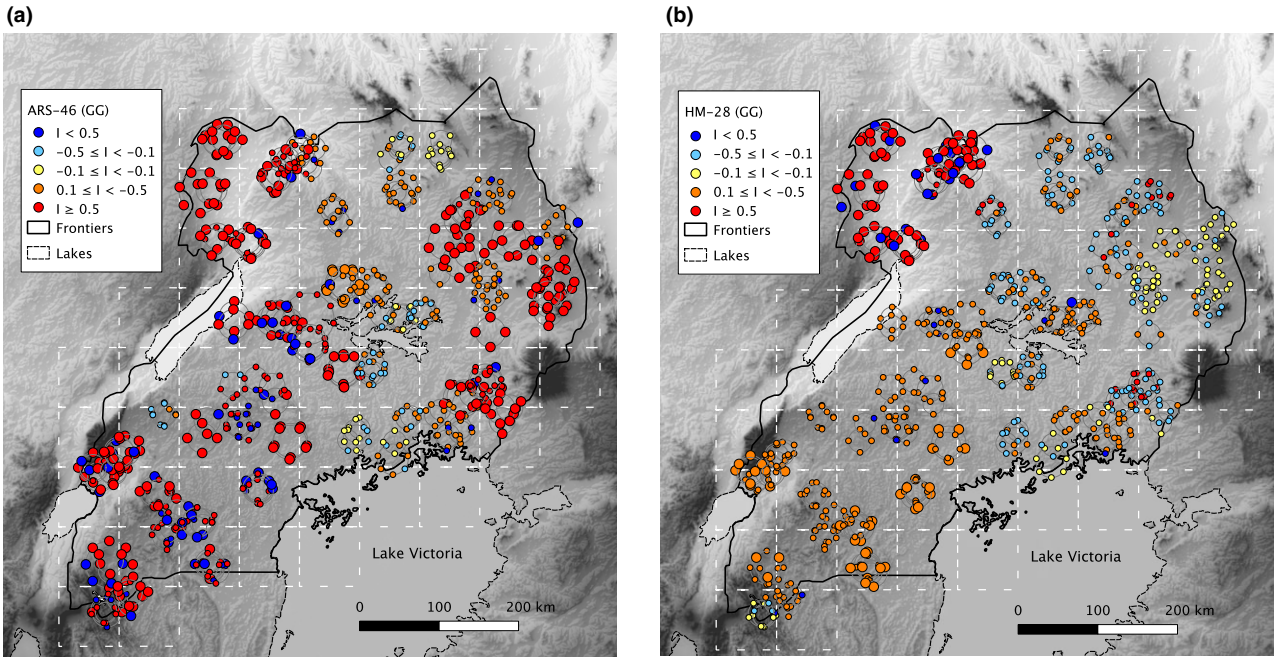
Table 2 (Continued)

			SAMβADA		BAYENV		LFMM	
Loci	Chr.	Pos (Mbp)	Env	*P*-value	Env	BF	Env	*P*-value
18. BTB-01356178	20	2.49	latitude	1.49×10^{-26}	tmin10	62.7		
			prec7	6.28×10^{-26}	bio9	33		
			prec6	6.69×10^{-23}	prec6	27.9		
19. BTA-108359-no-rs	14	16.31	longitude	2.35×10^{-26}				
			prec7	3.87×10^{-26}				
			prec11	6.28×10^{-25}				
20. ARS-BFGL-NGS-15960	5	28.02	prec7	3.20×10^{-26}	bio9	76.8		
			prec6	7.57×10^{-24}	prec6	74.1		
			longitude	1.78×10^{-23}	prec5	72.9		
21. ARS-BFGL-NGS-116294	2	128.58	latitude	6.05×10^{-26}	tmin10	43		
			prec7	3.34×10^{-23}	bio9	18		
			bio7	6.44×10^{-23}	prec6	15.2		
22. Hapmap52789-rs29018750	5	70.26	bio7	1.05×10^{-25}				
			bio3	1.32×10^{-24}				
			latitude	1.08×10^{-23}				
23. ARS-BFGL-NGS-86183	8	43.5	prec7	4.73×10^{-25}				
			prec6	1.27×10^{-21}				
			latitude	3.35×10^{-21}				
24. ARS-BFGL-NGS-16554	20	1.44	bio7	1.18×10^{-24}	tmin10	55.4		
			prec7	1.27×10^{-24}	bio9	15.2		
			latitude	4.91×10^{-23}	prec6	12.7		
25. ARS-BFGL-NGS-30091	22	47.94	longitude	1.25×10^{-24}				
			prec7	3.08×10^{-14}				
			tmax10	3.63×10^{-14}				

northwest of Uganda. This particular region also showed the lowest values of isothermality in Uganda, that is a high variability of temperatures. This correlation between HM-28 (GG) and isothermality also appeared with bivariate LISAs, where the presence of the genotype was compared with the mean value of isothermality among neighbouring points (not shown).

Discussion

The main features of SAMβADA are the processing speed, the multivariate modelling and the measurement of spatial autocorrelation. Processing speed is key when dealing with high-throughput data, while multivariate modelling and spatial autocorrelation measurements improve the interpretation of results, particularly when the data set includes population structure. Models may indeed include the global ancestry coefficients provided by a preliminary analysis (e.g. ADMIXTURE). This facilitates the detection of genotypes correlated with the environment while taking population structure into account. Additionally, introducing measurements of spatial autocorrelation into these analyses takes into account the valuable contribution of spatial statistics in landscape genomics. Unlike most current and nonspatial approaches (e.g. Coop *et al.* 2010; Frichot *et al.* 2013; Frichot & François 2015), SAMβADA allows the determination of whether the observed data reflects independent samples, a requirement of the underlying statistical model. Spatial autocorrelation measurements help assess whether the occurrence of a genotype is related to its frequency in the surrounding locations. More specifically, local indices of spatial autocorrelation allow the mapping of areas prone to spatial dependence. The results of the present analysis show that using spatial statistics in conjunction with correlative models may lower the risk of false positives in landscape genomics. This is important when the individuals under study share demographic history (e.g. individuals within breeds of a livestock species – Orozco-terWengel *et al.* 2015 – or absence of gene flow in a divergence-after-speciation model configuration – Cruickshank & Hahn 2014), in the presence of isolation by distance (Meirmans 2012) or cryptic relatedness (Corbett-Detig *et al.* 2015), and when genetic background are ignored (François *et al.* 2016). However, while some population structures do not show significant spatial autocorrelation, one has to keep in mind that particular demographic structures may totally mimic selection signatures (Holderegger *et al.* 2008) and that in this case, correlative approaches are not able to recognize the cause of the spatial pattern observed. SAMβADA can analyse such cases with the



ARS-46 (GG) HM-28 (GG)

Fig. 3 Local indicators of spatial association of markers ARS-46 (genotype GG) and HM-28 (genotype GG). The weighting scheme is based on the 20 nearest neighbours. Red points tend to be similar to their neighbours, while blue points differ from them. Yellow points are independent from their neighbourhood. Small points indicate nonsignificant values ($P > 0.001$). The map in the background represents the relief, the darker the shade, the higher the altitude. Samples coming from the same farm have been spread on a circle around their actual location. [Colour figure can be viewed at wileyonlinelibrary.com]

multivariate models including the global ancestry coefficients.

Simulation study

The simulation study shows that SAMβADA univariate models and LFMM are able to detect the locus under selection in discrete, low-agglomerated landscapes, provided that the strength of selection is high enough. In the weak selection scenario, the mortality at birth is compensated by the dispersal of individuals in approximately half the replicates, so that the locus under selection is not detected. On the contrary, it is only missed once for the intermediate selection strength and is always detected for the strong selection scenario. However, this power of detection comes at the cost of high FPRs. The relatively low dispersal capacity of individuals leads to isolation by distance, so that frequencies of neutral alleles vary across space (Forester *et al.* 2016). This induces some spurious correlations with the '*x*' and '*y*' coordinates, used as proxies for continuous gradient-like environmental variables. These false detections affect both the SAMβADA univariate models, which do not correct for population structure, and LFMM, which tries to model it as unobserved variables. Besides their comparable TPR and FPR, LFMM seems to recognize the variable 'habitat' as the driver of selection in more replicates than SAMβADA which tends to assign better scores to models involving '*x*' or '*y*'. The GEA index of both methods increases with the selection strength, showing that higher selection strengths increase the power of detection and the ability to distinguish the environmental variable driving local adaptation.

SAMβADA's multivariate analysis leads to a considerably lower FPR than the previous methods (2–4% vs. 39–45%). Therefore, including population structure as a set of covariates improves the ability of SAMβADA to distinguish between signals of selection and differences in allelic frequencies due to isolation by distance. In the strong selection scenario, the multivariate models have the same power of detecting the locus under selection as the univariate models. However, the TPR is lower for the intermediate level of selection and very low for the weak selection scenario. Thus, controlling for population structure in multivariate models with a conservative significance threshold (e.g. Bonferroni correction) may decrease the power of detecting loci under weak to moderate selection strengths. These results illustrate the trade-off which exists between the power of detection of correlation-based approaches and the specificity of the

said detections obtained by taking the population structure into account.

The analysis of spatial autocorrelation enables the comparison of the locus under selection (L0) to neutral loci detected by samβada (false positives) and neutral loci not detected by samβada (true negatives). False-positive loci tend to have higher values of Moran's *I* than the group of true negative for all selection scenarios (see Fig. 1 and Figs S4–S6, Supporting information, for details). This illustrates the fact that spatial dependency in neutral loci increases their probability of being detected as potentially subject to selection. The spatial autocorrelation of both groups of neutral loci (false-positive and true-negative) stays stable with increasing selection pressure, while the spatial autocorrelation of true positive clearly increases with the selection pressure. The latter effect may be emphasized by the fact that several genotypes are positively selected in distinct habitats and negatively selected in the other habitats. Therefore, loci with high values of spatial autocorrelation can also be subject to selection and should not be discarded from the analysis on this sole criterion. Local indicators of spatial autocorrelation draw the same picture as the global Moran's *I*: when counting the number of sampling points showing a significant LISA value, the locus under selection is often among the loci showing the most significant LISA points, and this trend also increases with selection pressure (Table S6, Supporting information).

Ugandan cattle

In the study of Ugandan cattle, samβada detected the highest number of SNPs as potentially subject to selection among the four approaches. However, samβada's detection rate may reflect false positives probably due to population structure. This interpretation is supported by the shape of the quantile–quantile plots, where samβada univariate analysis shows an excess of models with small *P*-values (see Fig. S2, Supporting information, part a). Indeed, the distribution of cattle populations follows roughly a north–south axis which corresponds to the gradient shown by some environmental variables. This overlay may result in some spurious associations. Regardless, environmental conditions can underlie the intensity of some health threats, such as the trypanosomiasis. The two cattle species bore some specific traits before they met in Uganda (e.g. drought tolerance and disease resistance). These specificities have contributed to shape their respective distribution in the country. In this case, the observed genome–environment associations can reflect the local adaptation of cattle in Uganda. Moreover, the discrepancy between the results may indicate that the more conservative approaches induce some false negatives. The zebus are indeed highly admixed with Ankole cattle and only eight of them were retained in the reference population used by bayenv and arlequin (compared with 162 Ankole cattle). This difference in sample size may have affected arlequin's analysis and prevented the detection of selection signatures. Another potential source of discrepancy between approaches is the use of a pre-existing SNP chip to analyse local adaption. Some ascertainment bias could result from the choice of the set of loci as neither Shorthorn zebus nor Ankole cattle were included in the SNP chip development. However, using the observed heterozygosity of both populations as a proxy of the effect of ascertainment bias, we can see that the average observed heterozygosity of Ankole is ~0.27 and that of the one of zebu is ~0.25, largely reflecting that if there is a bias it probably affects both groups similarly. Additional data from the BovineHD Genotyping BeadChip (Illumina Inc., San Diego, CA, USA) suggest that both Ankole and zebu here have similar observed heterozygosity (L. Colli, personal communication).

Comparing these results in the light of spatial dependence gives information about the differences between samβada's, bayenv's and lfmm's detections. The locus ARS-46 was detected by samβada only, and its genotype GG showed a widespread pattern of spatial autocorrelation (Fig. 3a). This pattern could originate from the underlying population structure, as Ankole cattle are more common in the southwest, while zebus are more common in the northeast of the country. This spatial dependence in the occurrence of this genotype is in contradiction with the assumptions of samβada's statistical model. Thus, the correlation detected by logistic regressions between ARS-46 (GG) and environmental variables could be spuriously driven by demographic factors, as described above. Patterns of spatial dependence for HM-28 presented a different situation (Fig. 3b). The low value of spatial autocorrelation for HM-28 (GG) implies that the distribution of this genotype was mostly independent of location, thus the logistic models are reliable for this genotype. HM-28 was also detected by the three landscape genomic approaches and by samβada multivariate analysis, and this supports a possible adaptive origin of the observed correlation with isothermality. Maps of local spatial autocorrelation for the genotypes ARS-46 (GG) and HM-28 (GG) illustrated a general trend: bayenv and lfmm discarded SNPs showing significant local spatial autocorrelation for a large proportion of the sampling locations, while samβada detected them. Thus, in this case, measuring the local autocorrelation of candidate genotypes may help distinguishing between the effects of local adaptation and those of population structure among samβada's detections.

Regarding common detections, three of the SNPs identified by samβada when population structure was included as a covariate were among the common

detections of the three correlative approaches. SAMβADA bivariate analysis is rather conservative with only eight detected loci; however the distribution of *P*-values is close to the expected distribution, suggesting that population structure was taken correctly into account (see Fig. S2, part b, Supporting information). Thus, pre-existing knowledge on demography may be built on to refine correlation-based detections of selection signatures. One possible approach consists of assessing population structure and then including one or a few variables summarizing this structure in the constant model used by SAMβADA. In this way, only genotypes showing a significant correlation with the environment while taking the population structure into account are detected. In case there are more than two main populations, hence requiring several variables to summarize the samples' ancestry, these summary variables could for instance be derived from a PCA of the samples' coefficients of ancestry. In the present study, the coefficients of ancestry for the Ankole and zebu populations are essentially complementary for most samples, thus using the first principal axis of the PCA is similar to using one of these coefficients of ancestry as the summary variable.

Concerning the biological function of frequently detected loci, these three loci are located on chromosome 5, near the gene POLR3B whose mouse counterpart is involved in limiting infection by intracellular bacteria and DNA viruses (UniProt, www.uniprot.org). Moreover, genotype HM-28 (GG) shows spatial autocorrelation in the northwestern part of Uganda and this area overlaps with one of those where the highest load of tsetse fly (*Glossina* spp.) occurs in the country (Abila *et al.* 2008; MAAIF *et al.* 2010). Hence, the risk of cattle trypanosomiasis is high in this region and the detected mutations may be involved in parasite resistance.

Comparison between simulated and empirical data

The analyses of the simulation and cattle data lead to some common observations. SAMβADA's univariate modelling detects some spurious associations in scenarios with population structure. As a countermeasure, multivariate analysis, which includes predictors variables accounting for this population structure, lowers the rate of false positives. However, the assumption that the main axis of molecular variation represents only the population structure may induce some false negatives, especially when the selection pressure is low (simulated data) or when the full data set was used to assess the said population structure (cattle data). The comparison of the two types of data also reveal some differences: the environmental variable 'habitat' which drives selection in the simulation data is discrete with a complex spatial distribution (low-agglomeration), while there are many continuous environmental variables describing the habitat in Uganda and most of these present a north – south gradient. Another difference is the spatial distribution of individuals: each sample came from a distinct location in the simulation data, while several individuals were sampled at each location in Uganda. These differences may be reflected in the observed patterns of spatial autocorrelation. The simulated data show that molecular markers displaying a high spatial dependence can actually be subject to selection. In fact, as many environmental variables are auto-correlated in nature, it can be expected that the distribution of a molecular marker selected by one of these variables will also present some spatial correlation. Therefore, it is currently not possible to distinguish between true and false positives solely on the basis of their spatial dependence. The most efficient approach involves comparing the results of several methods taking the population structure into account, and to observe the patterns of spatial autocorrelation to analyse how the detected GEAs are linked to the spatial distributions of markers and environmental variables.

Perspectives

The increasing availability of large molecular data sets raises challenges regarding their analysis. Correlative approaches in landscape genomics enable fast detection of candidate loci to local adaptation. However, these methods must take into account the effect of population structure (De Mita *et al.* 2013; Frichot *et al.* 2013; Joost *et al.* 2013; Frichot & François 2015). Limited dispersal of individuals leads to spatial autocorrelation of marker frequencies, which may cause spurious correlations with the environment. SAMβADA addresses these issues by rapidly detecting selection signatures with the possibility of including prior knowledge of the population structure in the analysis and by measuring the level of spatial autocorrelation for candidate loci. The next methodological step involves developing spatially explicit models that directly include autocorrelation. SGLMM (Guillot *et al.* 2014) provides such a model; however, the current R-based implementation does not enable whole-genome analysis.

The recent availability of whole-genome sequence (WGS) data also raises issues regarding the statistical assessment of multiple comparisons. Indeed, while many individuals and few genetic markers were available 10 years ago, the current high costs of WGS limit the number of sequenced samples. Therefore, standard procedures for multiple comparisons, such as the Bonferroni correction, are over-conservative and may lead to discard some adaptive loci. In this context, alternatives procedures focus on controlling the ratio of false positives in a set of significant results. Among them, Storey and

Tibshirani's false discovery rate (2003) was especially designed for large molecular data sets and suits any detection method relying on significance tests. This method is available as an R package (*q* value, Storey *et al* 2015) and its implementation in SAMβADA is ongoing.

Computation time is critical when processing large data sets. In this context, SAMβADA is able to swiftly analyse high-density SNP-chips and variants from WGS. When taking population structure into account, SAMβADA's multivariate analysis is approximately 10 times quicker than LFMM and 16 times than BAYENV for a data set comparable to this study, and these ratios increase with larger data sets (see Table S7, Supporting information). SAMβADA's simple underlying model has the advantage that the computation time grows linearly with the size of the genetic data under study. Therefore, SAMβADA's module for parallelized processing enables the analysis of WGS data sets on desktop computers (see Table S9, Supporting information). SAMβADA's processing speed, combined with its ability to analyse the spatial autocorrelation in molecular data and to incorporate prior knowledge on population structure, suits a wide range of applications, especially those involving whole-genome sequence data.

Acknowledgements

We thank Sergio Rey for his advice on assessing the significance of LISA, Stephan Morgenthaler for fruitful discussions on assessing the significance of multivariate logistic models, Olivier François and Eric Frichot for their explanations on LFMM and Gilles Guillot for providing us with SGLMM for testing purposes. We thank Kevin Leempoel for his help in analysing the spatial autocorrelation and Estelle Rochat for her careful reading and useful comments on the manuscript.

Funding

This research was funded by EU FP7 project NextGen (Grant KBBE-2009-1-1-03).

Resources

Software availability

SAMβADA is an open source software written in C++ available at http://lasig.epfl.ch/sambada (under the license GNU GPL 3). Compiled versions are provided for Windows, Linux and MacOS X.

Data availability

NextGen data are described at http://projects.ensembl.org/nextgen/. Ugandan cattle SNP data are available at ftp://ftp.ebi.ac.uk/pub/databases/nextgen/bos/variants/chip_array/ in PLINK format (files UGBT.bovineSNP50.UMD3_1.20140307.[ped/map].gz) with the following data policy ftp://ftp.ebi.ac.uk/pub/databases/nextgen/documentation/README_data_use_policy. Simulation data, landscape surfaces and individual sample files are available at Dryad doi:10.5061/dryad.v0c77.

References

Abila PP, Slotman MA, Parmakelis A *et al.* (2008) High levels of genetic differentiation between Ugandan *Glossina fuscipes fuscipes* populations separated by Lake Kyoga. *PLOS Neglected Tropical Diseases*, **2**, e242.

Ajmone Marsan P, Garcia JF, Lenstra JA, the Globaldiv Consortium (2010) On the origin of cattle: how aurochs became cattle and colonized the world. *Evolutionary Anthropology*, **19**, 148–157.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Anselin L (1995) Local Indicators of Spatial Association – LISA. *Geographical Analysis*, **27**, 93–115. GISDATA (Geographic Information Systems Data) Specialist Meeting on GIS (Geographic Information Systems) and Spatial Analysis, Amsterdam, Netherlands, Dec 01–05, 1993.

Anselin L, Syabri I, Kho Y (2006) Geoda: an introduction to spatial data analysis. *Geographical Analysis*, **38**, 5–22.

Arguez A, Vose RS (2010) The definition of the standard WMO climate normal: the key to deriving alternative climate normals. *Bulletin of the American Meteorological Society*, **92**, 699–704.

Balloux F (2001) EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.

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 B: Biological Sciences*, **263**, 1619–1626.

Bivand R, Piras G (2015) Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, **63**, 1–36.

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Colli L, Joost S, Negrini R *et al.* (2014) Assessing the spatial dependence of adaptive loci in 43 European and Western Asian goat breeds using AFLP markers. *PLoS One*, **9**, e86668.

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

Corbett-Detig RB, Hartl DL, Sackton TB (2015) Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology*, **13**, 1–25.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

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

Dobson AJ, Barnett AG (2008) *An Introduction to Generalized Linear Models*, 3rd edn. Chapman & Hall, Boca Raton, FL.

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

Farr TG, Rosen PA, Caro E *et al.* (2007) The shuttle radar topography mission. *Reviews of Geophysics*, **45**, RG2004.

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

Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR (2016) Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular Ecology*, **25**, 104–120.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

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

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

Gardner RH (1999) *RULE: Map Generation and a Spatial Analysis Program*, pp. 280–303. Springer, New York, NY.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

GDAL Development Team (2013) *GDAL – Geospatial Data Abstraction Library, Version 1.10*. Open Source Geospatial Foundation, Beaverton, Oregon.

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

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

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics*, **7**, 1–32.

Henry P, Russello MA (2013) Adaptive divergence along environmental gradients in a climate-change-sensitive mammal. *Ecology and Evolution*, **3**, 3906–3917.

Hijmans R, Cameron S, Parra J, Jones P, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

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

Jeffreys H (1961) *The Theory of Probability*, 3rd edn. Oxford University Press, Oxford, UK.

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

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

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

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

Landguth EL, Cushman SA (2010) CDPOP: a spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, **10**, 156–161.

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

Lv F-H, Agha S, Kantanen J *et al.* (2014) Adaptations to climate-mediated selective pressures in sheep. *Molecular Biology and Evolution*, **31**, 3324–3343.

Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**, 2839–2846.

Ministry of Agriculture, Animal Industry and Fisheries, Uganda, Uganda Bureau of Statistics, Food and Agriculture Organization of the United Nations, International Livestock Research Institute, and World Resources Institute (2010) *Mapping a Better Future: Spatial Analysis and Pro-Poor Livestock Strategies in Uganda*, pp. 30–37. World Resources Institute, Washington, DC and Kampala.

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

Moran PAP (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Ndumu DB, Baumung R, Hanotte O *et al.* (2008) Genetic and morphological characterisation of the Ankole Longhorn cattle in the African Great Lakes region. *Genetics Selection Evolution*, **40**, 467–490.

Orozco-terWengel P, Barbato M, Nicolazzi EL, Biscarini F, Milanesi M (2015) Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Frontiers in Genetics*, **6**, 191.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, 2074–2093.

Pemstein D, Quinn KM, Martin AD (2011) The Scythe statistical library: an open source C++ library for statistical computation. *Journal of Statistical Software*, **42**, 1–26.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rey SJ, Anselin L (2010) PySAL: A python library of spatial analytical methods. In: *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (eds Fischer M, Getis A), pp. 175–193. Springer, Berlin.

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

Sokal RR, Oden NL (1978) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.

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

Storey JD with contributions from Bass AJ, Dabney A, Robinson D (2015). qvalue: Q-value estimation for false discovery rate control. R package version 2.4.2. http://github.com/jdstorey/qvalue

de Villemereuil P, Gaggiotti O (2015) A new F_{ST} method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, **6**, 1248–1258.

Vitalis R, Dawson K, Boursot P, Belkhir K (2003) DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity*, **94**, 429–431.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

P.T., S.J., M.B., L.C. and R.N. designed research. S.S., P.O.T.W., L.C., S.J., B.F., C.M., R.N. and S.D. performed research. S.S., S.J. and P.O.T.W. contributed to new analytical tools. S.S., S.J., P.O.T.W., B.F., S.D., M.J. and E.L. wrote and reviewed the manuscript. All the authors undertook revisions, contributed intellectually to the development of this manuscript and approved the final manuscript.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Population structure computed with ADMIXTURE.

Fig. S2 Quantile-Quantile plots of the detections of SAMβADA and LFMM in the Ugandan cattle dataset.

Fig. S3 Example of a landscape selection configuration used for simulations.

Fig. S4 Correlograms for the 10 replicates of the simulation data under weak selection pressure (1%).

Fig. S5 Correlograms for the 10 replicates of the simulation data under intermediate selection pressure (5%).

Fig. S6 Correlograms for the 10 replicates of the simulation data under strong selection pressure (10%).

Table S1 Environmental variables used to detect selection signatures with correlative approaches.

Table S2 List of SNPs detected by SAMβADA with bivariate models including the variable 'pop' representing the population structure.

Table S3 Detections of SAMβADA in the simulation data using univariate models.

Table S4 Detections of SAMβADA in the simulation data using multivariate models taking into account the population structure.

Table S5 Detections of LFMM in the simulation data.

Table S6 Summary of the local indicators of spatial association (LISA) for the simulation data.

Table S7 Comparison of approximate computation times among methods.

Table S8 Comparison of computation times between MATSAM (v2) and SAMβADA.

Table S9 Computation times for parallel processing of larger datasets with SAMβADA.

Appendix S1 Multivariate analysis.

Appendix S2 Spatial autocorrelation.

Appendix S3 Alternative methods to detect selection.