A new F_{ST}-based method to uncover local adaptation using environmental variables.

Pierre de Villemereuil[*], & Oscar E. Gaggiotti[*†]

[*]: Université Jospeh Fourier, Centre National de la Recherche Scientifique,

LECA, UMR 5553, 2233 rue de la piscine, 38400 Saint Martin d'Hères, France

[†]Scottish Oceans Institute, University of St Andrews,

Fife, KY16 8LB, United Kingdom

Keywords: genome-scan, local adaptation, environment, F model, Bayesian methods, false discovery rate

Corresponding author: Pierre de Villemereuil, E-mail: bonamy@horus.ens.fr

Abstract

• Genome-scan methods are used for screening genome-wide patterns of DNA polymorphism to detect signatures of positive selection. There are two main types of methods: *(i)* "outlier" detection methods based on F_{ST} that detect loci with high differentiation compared to the rest of the genome, and *(ii)* environmental association methods that test the association between allele frequencies and environmental variables.

• We present a new F_{ST}-based genome-scan method, BayeScEnv, which incorporates environmental information in the form of "environmental differentiation". It is based on the F model, but, as opposed to existing approaches, it considers two locus-specific effects; one due to divergent selection, and another due to various other processes different from local adaptation (e.g. range expansions, differences in mutation rates across loci or background selection). The method was developped in C++ and is avaible at `http://github.com/devillemereuil/bayescenv`.

• A simulation study shows that our method has a much lower false positive rate than an existing F_{ST}-based method, BayeScan, under a wide range of demographic scenarios. Although it has lower power, it leads to a better compromise between power and false positive rate.

• We apply our method to a human dataset and show that it can be used successfully to study local adaptation. We discuss its scope and compare it to other existing methods.

Introduction

One of the most important aims of population genomics (Luikart *et al.*, 2003) is to uncover signatures of selection in genomes of non model species. Of special interest is the process of local adaptation, whereby populations

experiencing different environmental conditions undergo adaptive, selective pressures specific to their local habitat. As a result, populations evolve traits that provide an advantage in their local environment. Many experimental approaches focused on potentially adaptive traits have been developed to test for local adaptation (reviewed in Blanquart *et al.*, 2013), but only recently it has become possible to make inferences about the genomic regions involved in local adaptation processes. Indeed, the advent of next generation sequencing (NGS, Shendure and Ji, 2008) has fostered the development of so-called genome-scan methods aimed at identifying regions of the genome subject to selection. These methods are now widely used in studies of local adaptation (Faria *et al.*, 2014).

There are two main types of genome-scan methods. The first type detects 'outlier' loci using locus-specific F_{ST} estimates, which are compared to either an empirical distribution (Akey *et al.*, 2002), or to a distribution expected under a neutral model of evolution (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). The rationale behind these methods is that local adaptation leads to strong genetic differentiation between populations, but only at the selected loci (or marker loci linked to them). Thus, loci with very high F_{ST} compared to the rest of the genome are suspected to be under strong local adaptation and are referred to as outliers. The outlier approach was further extended to statistics akin to F_{ST} (Bonhomme *et al.*, 2010; Günther and Coop, 2013), and also to other unrelated statistics (Duforet-Frebourg *et al.*, 2014). One limitation of these methods is that they are not designed to test hypotheses about the environmental factors underlying the selective pressure.

A second type of methods focuses on environmental variables and aims at associating patterns of allele frequency to environmental gradients. The rationale is that selective pressures should create associations between allele frequencies at the selected loci and the causal environmental variables (Coop *et al.*, 2010). In the presence of population structure, performing a simple linear regression would be an error-prone approach (De Mita *et al.*, 2013; de Villemereuil *et al.*, 2014). Instead, existing methods account for population structure by modelling the allele frequency covariation across populations (Coop *et al.*, 2010; Frichot *et al.*, 2013; Guillot *et al.*, 2014). One disadvantage of most of these approaches is that the parameters that capture the effect of demographic history on genetic differentiation do not have a clear biological interpretation, which in turn makes the rejection of the null model hard to interpret in terms of detection of local adaptation. Especially, note that although the elements of the covariance matrix estimated by Coop *et al.* (2010) could in principle be interpreted as parametric estimates of the pairwise and population-specific F_{ST}, this is only true when levels of genetic drift are low (Nicholson *et al.*, 2002).

It is important to note that, regardless of the type of genome-scan method under consideration, processes other than local adaptation might be responsible for the observed spatial patterns in allele frequency or F_{ST}. These include demographic processes such as allele surfing (Edmonds *et al.*, 2004) or hierarchical population structure (Excoffier *et al.*, 2009), large differences in mutation rate across loci (Edelaar *et al.*, 2011), hybrid incompatibility following secondary contact (Kruuk *et al.*, 1999) and background selection (Charlesworth, 1998).

It is therefore possible that some of the loci identified as outliers are in fact false positives. Accounting for processes other than selection would require introducing parameters that could appropriately capture the effect of these other processes.

Here, we present a method that incorporates features of the two types of genome-scans described above. The objective is to allow inferences about the environmental factors underlying selective pressures, and simultaneously better discriminate between true and false genetic signatures of local adaptation. Note that our new method focuses only on local adaptation driven by a focal environmental variable and, therefore, differs from other F_{ST}-based methods that carry out "blind" genome scans. Thus, this new approach is aimed at testing hypothesis about specific drivers of local adaptation such as altitude (Bigham *et al.*, 2010; Foll *et al.*, 2014), salinity (Larsson *et al.*, 2007; Daub *et al.*, 2013), pathogens (Fumagalli *et al.*, 2011; Daub *et al.*, 2013), etc.

Our method is based on the Bayesian approach first proposed by Beaumont and Balding (2004) and later extended by Foll and Gaggiotti (2008). The original formulation considers population- and locus-specific F_{ST}'s, which are described by a logistic model with three parameters: a locus-specific term, α_i, that captures the effect of mutation and some forms of selection, a population-specific term, β_j, that captures demographic effects (e.g. N_e and migration) and a locus-by-population interaction term, γ_{ij}, that reflects the effect of local adaptation. The estimation of the first two terms benefits from sharing information across loci or populations, but this is not the case for the interaction term, which is therefore poorly estimated (Beaumont and Balding, 2004, but see Riebler *et al.*, 2008). In practice signatures of local adaptation are therefore inferred from the locus-specific effects (α_i) under the assumption that large positive values reflect adaptive selection. The implicit assumption is that background selection and mutation should not have much of an effect on this term. In order to relax this assumption and to better estimate the interaction term we introduce environmental data so that $\gamma_{ij} = g_i E_j$, where E_j is the "environmental differentiation" observed in population j and g_i is a locus-specific coefficient. In what follows, we first describe in detail the probabilistic model underlying our Bayesian approach. We then evaluate its performance using simulated data and then present an application using a human dataset. Finally, we discuss the scope of our method and compare it with other existing genome-scan approaches.

Statistical model

Modelling allele frequencies using the F model

Our new genome-scan approach is based on the F model (Beaumont and Balding, 2004; Foll and Gaggiotti, 2008) and extends the software BayeScan (Foll and Gaggiotti, 2008) by incorporating environmental data so as to explicitly consider local adaptation scenarios. Full details of the F model are given by Gaggiotti and Foll (2010), so here we only provide a brief description. The core assumptions of the F model is that all populations share a common pool of migrants, but that their effective sizes and immigration rates are population-specific.

Thus, population structure at each locus is described by local F_{ST}'s that measure genetic differentiation between each local population and the migrant pool.

The F model uses the multinomial-Dirichlet likelihood for the allele counts $\boldsymbol{a_{ij}} = (a_{ij1}, \ldots, a_{ijK_i})$ at locus i within population j (where K_i is the number of distinct alleles at locus i) with parameters given by the migrant pool allele frequencies, $\boldsymbol{f_i} = (f_{i1}, \ldots, f_{iK_i})$, and a population- and locus-specific parameter of similarity, $\theta_{ij} = \dfrac{1 - F_{\mathrm{ST}}^{ij}}{F_{\mathrm{ST}}^{ij}}$:

$$\boldsymbol{a_{ij}} \sim \mathit{multDir}(\theta_{ij} f_{i1}, \ldots, \theta_{ij} f_{iK_i}), \tag{1}$$

where *multDir* stands for the multinomial-Dirichlet distribution.

Although, for the sake of simplicity, we only present here the formulation for co-dominant data, the software implementing our approach also allows for dominant data (e.g. AFLP markers) using the same probabilistic model as Foll and Gaggiotti (2008). Note finally that, for bi-allelic co-dominant markers (e.g. SNP markers), the likelihood reduces to a beta-binomial model.

Alternative models to explain population structure

Our purpose is to better discriminate between true signals of local adaptation and spurious signals left by other processes. Therefore, we assume that genetic differentiation at individual loci is influenced by three type of effects: *(i)* genome-wide effects due to demography, *(ii)* a locus-specific effect due to local adaptation caused by the focal environmental variable, and *(iii)* locus-specific effects unrelated to the focal environmental variable. Although in principle one could consider all seven alternative models that can be constructed with different combinations of these three effects, most of them would not have any biological meaning. For example, all models should include genome-wide effects associated with genetic drift. Additionally, we do not consider the two types of locus-specific effects simultaneously in a full model. The reason for this is that the inclusion of α_i along with g_i is not justified biologically. This is because the joint effect of local adaptation and another locus-specific effect such as allele surfing or background selection on the same locus is extremely unlikely either because of the strong effect of genetic drift in the first instance or the implausibility of a favourable variant arising and increasing in frequency in a highly conserved region subject to strong purifying selection. Thus, we focus on three different models to explain the genetic structuring at individual loci.

Null model of population structure Under the null hypothesis that all loci are neutral, the local differentiation parameter F_{ST}^{ij} will be driven only by local population demography and, hence, should be common to all loci:

$$\log\left(\frac{F_{\mathrm{ST}}^{ij}}{1 - F_{\mathrm{ST}}^{ij}}\right) = \log\left(\frac{1}{\theta_{ij}}\right) = \beta_j. \tag{2}$$

A high β_j value means that the population j is strongly differentiated from the pool of migrants. This could be due to a lack of immigration from the other populations, a reduced effective size, or a particular spatial

structure.

Alternative model of local adaptation In this model, we focus on a particular signature left by a process of local adaptation. If selection is driven by a putative environmental factor, we expect that genetic differentiation for the locus or loci under selection will be stronger than expected under neutrality for populations with strong environmental differentiation. Any measure of distance between the environmental value of population j and the average environment could serve as a measure of differentiation. For the sake of simplicity, we here only consider the absolute value. Furthermore, in order to facilitate the calibration of prior distributions, we consider standardised environmental values with unit variance.

To model the effect of local adaptation on locus i, we consider the impact of environmental differentiation E_j of population j on the locus, we thus modify Eq. 2 as follows:

$$\log\left(\frac{F_{\mathrm{ST}}^{ij}}{1-F_{\mathrm{ST}}^{ij}}\right) = \beta_j + g_i E_j, \tag{3}$$

where g_i quantifies the sensitivity of locus i to the environmental differentiation.

Alternative model of locus-specific effect Local adaptation with respect to the focal environmental variable is not the only evolutionary phenomenon that could lead to departures from the neutral model. Other phenomena that could produce such locus-specific effects include local adaptation due to other unknown factors, large differences in mutation rate across loci, the so-called allele surfing phenomenon (Edmonds *et al.*, 2004) and background selection (Charlesworth, 2013).

This is accounted for by using the following parametrisation for local differentiation:

$$\log\left(\frac{F_{\mathrm{ST}}^{ij}}{1-F_{\mathrm{ST}}^{ij}}\right) = \alpha_i + \beta_j. \tag{4}$$

The main advantage of implementing both of the above alternative models is that we can distinguish between departures from the neutral model of unknown origin (using Eq. 4) and departures due to local adaptation caused by a particular environmental factor (using Eq. 3).

Material and Methods

Implementation of the statistical model

Our method, summarised in Fig. 1, uses two types of data: *(i)* the allele counts $\boldsymbol{a}$ for each locus in each population sample, and *(ii)* observed values $\boldsymbol{E}$ of an environmental variable (one value per population), which are transformed into environmental differentiation using an appropriate function. Indeed, our model aims at associating genetic distance (i.e. the F_{ST}^{ij}) with an environmental distance. Note that measuring an environmental

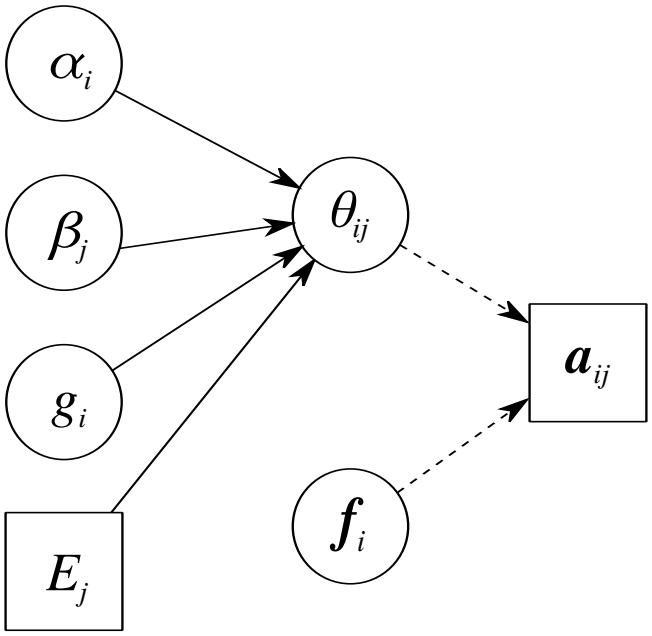

Figure 1: Directed Acyclic Graph (DAG) of the model. Squared nodes denote know quantities (E for environmental data, and A for genetic marker data). Circled nodes denote unknown parameters. Plain arrows stand for deterministic relationships, and dashed arrows stand for stochastic relationships.

distance requires to define a reference. The most natural reference would be the average of the environmental values, but this would not be always the case (see the example of adaptation to altitude in humans presented below). Also, it is strongly advised to standardise the environmental values by dividing by the standard deviation, in order to avoid effect size issues regarding the inference of the parameter g.

As stated in the previous section, there are three different models:

M1 Neutral model: β_j,

M2 Local adaptation model with environmental differentiation E_j: $\beta_j + g_i E_j$,

M3 Locus-specific model: $\alpha_i + \beta_j$.

Note that in our framework, the focal model being tested against the two others is **M2**. Thus, power and error rates (FPR and FDR) are computed for model **M2**. Model **M3** can be considered as a "nuisance model" whose role is to reduce the overall false positive rate by explaining the inflation of the variance in F_{ST} due to locus-specific effects other than selection driven by the focal environmental factor. Hence, the statistical significance of the parameter α_i is not of interest for BayeScEnv: only the significant values of g_i are considered. All three models were implemented using an RJMCMC algorithm (Green, 1995). In order to propose relevant values for new parameters during the jumps, the RJMCMC is preceded by pilot runs. These are aimed at both calibrating the MCMC proposals to reach efficient acceptance rates, and approximating the posterior distribution of parameters, as proposed by Brooks (1998) and already implemented in BayeScan (Foll and Gaggiotti, 2008). Our code is based on the source code of BayeScan 2.1 and is written in C++. The source

and binaries are available at `https://github.com/devillemereuil/bayescenv`.

Our prior belief in the three models is described by two parameters: the probability π of moving away from the neutral model and the preference p for **M3** against **M2** as alternative models. We can calculate the prior probability for each model as:

$$\begin{aligned} P(\mathbf{M1}) &= 1-\pi, \\ P(\mathbf{M2}) &= \pi(1-p), \\ P(\mathbf{M3}) &= \pi p. \end{aligned} \tag{5}$$

The mathematical details of the transition between models can be found in the Supplementary Material. Pilot studies showed that using values of p above 0.5 yielded extremely conservative results (note that setting $p = 1$ would mean that model **M3** is always favoured over **M2**, in which case the power of the method is zero, yielding no positives whatsoever).

We used a uniform Dirichlet prior for the allele frequencies $\boldsymbol{f}_i \sim Dir(1, \ldots, 1)$. The priors for the hyperparameters α and β, were Normal with mean -1 and variance 1 (note that the results of our method will be especially sensitive to the prior mean of α, but our pilot studies showed that -1 was a good default). Since under a local adaptation scenario the parameter g is only expected to be positive, it was assigned a uniform prior between 0 and 10.

Our method outputs posterior error probabilities and q-values, which are test statistics related to the False Discovery Rate (FDR) (Storey, 2002; Käll *et al.*, 2008). Contrary to the commonly used False Positive Rate (FPR), which is the probability of declaring a locus as positive given that it is actually neutral, the FDR is the proportion of the positive results that are in fact false positives, and is more appropriate for multiple testing (Käll *et al.*, 2008). See the Supplementary Information (SI) for more details.

Simulation analysis

We performed a simulation study to evaluate the performance of our method and compare it with that of BayeScan (Foll and Gaggiotti, 2008). We modelled 16 populations each with 500 individuals genotyped at 5,000 loci, among which one (monogenic scenario) or 50 (polygenic scenario) were under selection. We modelled three kinds of population structure: *(i)* a classical island model (IM), *(ii)* a one-dimension stepping-stone (SS) model and *(iii)* a hierarchically structured (HS) model.

The genome was composed of 5,000 bi-allelic SNPs spread along 10 chromosomes. The loci under selection due to an environmental variable E (see Fig. S2 and Eq. S7 and S8), one for the monogenic case and 50 for the polygenic case, were randomly distributed across the genome. Since all markers were independently initialised, our simulations yielded negligible linkage disequilibrium. Consequently, we considered as true positives only the loci subject to selection. For the IM and SS scenarios, we directly initialised all 16 populations. For the HS scenario, we initialised the ancestral population, which, following successive and temporally spaced-out fission

events, gave rise to 2, 4, ... , 16 populations. This hierarchical structure is reinforced by preferential migration between related populations. More details regarding migration and population history are available in the SI. This model is very close to that used by de Villemereuil *et al.* (2014). It should be particularly difficult for our method, because all populations are equally differentiated (i.e. the β_j parameters are expected to be roughly the same across populations), but a phylo-geographic covariance exists between related populations, which is not explicitly accounted for by our probabilistic model. More information regarding the environmental gradient and the fitness function are available in the SI, but, briefly, a polygenic multiplicative model was used with a selection strength of 0.02 (0.1 for the monogenic case).

The simulations were performed using the SimuPOP Python library (Peng and Kimmel, 2005) and the scripts are available online in the data section. Our simulated datasets were analysed using our C++ code and version 2.1 of BayeScan (Foll and Gaggiotti, 2008).

We generated 100 datasets for each scenario and computed the realised FDR, FPR and power yielded by BayeScan and our new environmental method (BayeScEnv). For the latter, we also compared several parametrisations using a prior probability π of jumping away from the neutral model of 0.1 (equivalent to the default prior odds used by BayeScan, which is 10) or 0.5, as well as a preference for the locus-specific model p of 0.5 (environmental and locus-specific models are equiprobable) or 0 (the locus-specific model is forbidden and only the environmental model is tested against the neutral one).

We supplemented these scenarios with a heterogeneous mutation rate case, based on the IM scenario above, where most of the genome had a high mutation rate of 0.05, whereas 50 loci had a low mutation rate of 10^{-7}. The result was an overall low F_{ST} of 0.05 for the whole genome, and of 0.10 for the low mutating loci.

HGDP SNP data analysis

In order to test our new method against a real dataset, we focused on 26 Asian populations from the Human Genome Diversity Panel (HGDP) SNP Genotyping data. This data set consists of 660,918 SNP markers genotyped using Illumina 650Y arrays. After cleaning the dataset from mitochondrial and sex-linked markers, we removed all markers with minor allele frequency below 5%. This left us with a total of 446,117 SNPs. For all populations, we obtained the following environmental variables from the BIOCLIM database: mean annual temperature, precipitation, and altitudinal data. We ran separate BayeScEnv analysis for each variable and compared the results with BayeScan (which doesn't use environmental variables). After standardisation of the environmental variables, we computed environmental differentiation from the mean for temperature and precipitation, and from the sea level for elevation. Gene ontology enrichment tests for the detected genes were performed using the "SNP mode" of the Gowinda software (Kofler and Schlötterer, 2012). The prior odds for BayeScan was 10 for this analysis. BayeScEnv prior parameters for this analysis were $\pi = 0.1$ and $p = 0.5$.

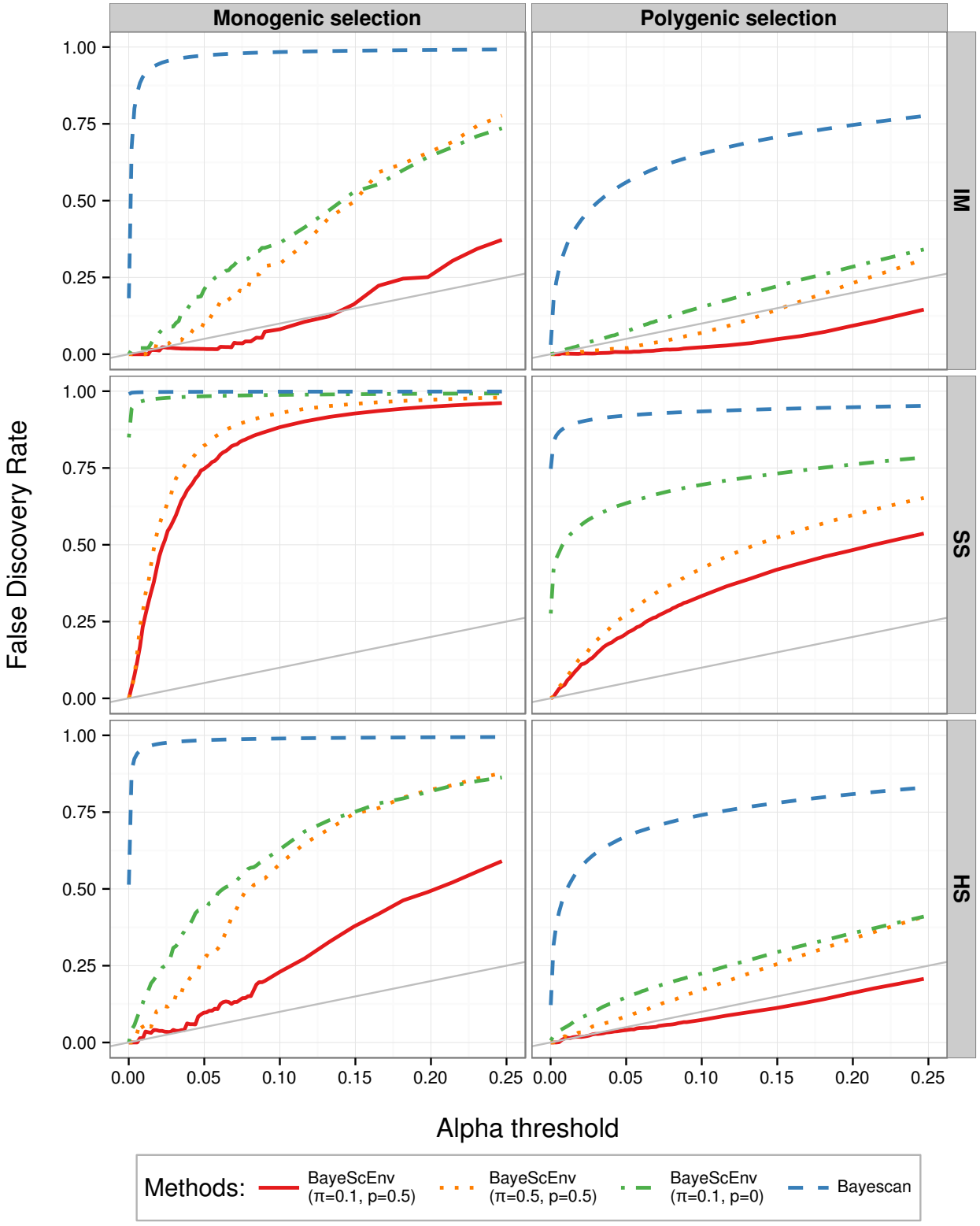

Figure 2: False Discovery Rate (FDR) against significance threshold α for three scenarios (IM: Island model, SS: Stepping-Stone model and HS: Hierarchically Structured model) and monogenic/polygenic selection. The grey line is the expected identity relationship between the FDR and α. The models tested are BayeScan (blue dashed), and BayeScEnv (orange dotted, green dot-dashed and solid red) with different probabilities π of jumping away from the neutral model (M1) and different preferences p for the locus-specific model (M3). Note that $p = 0$ means the environmental model (M2) is tested against the neutral one only.

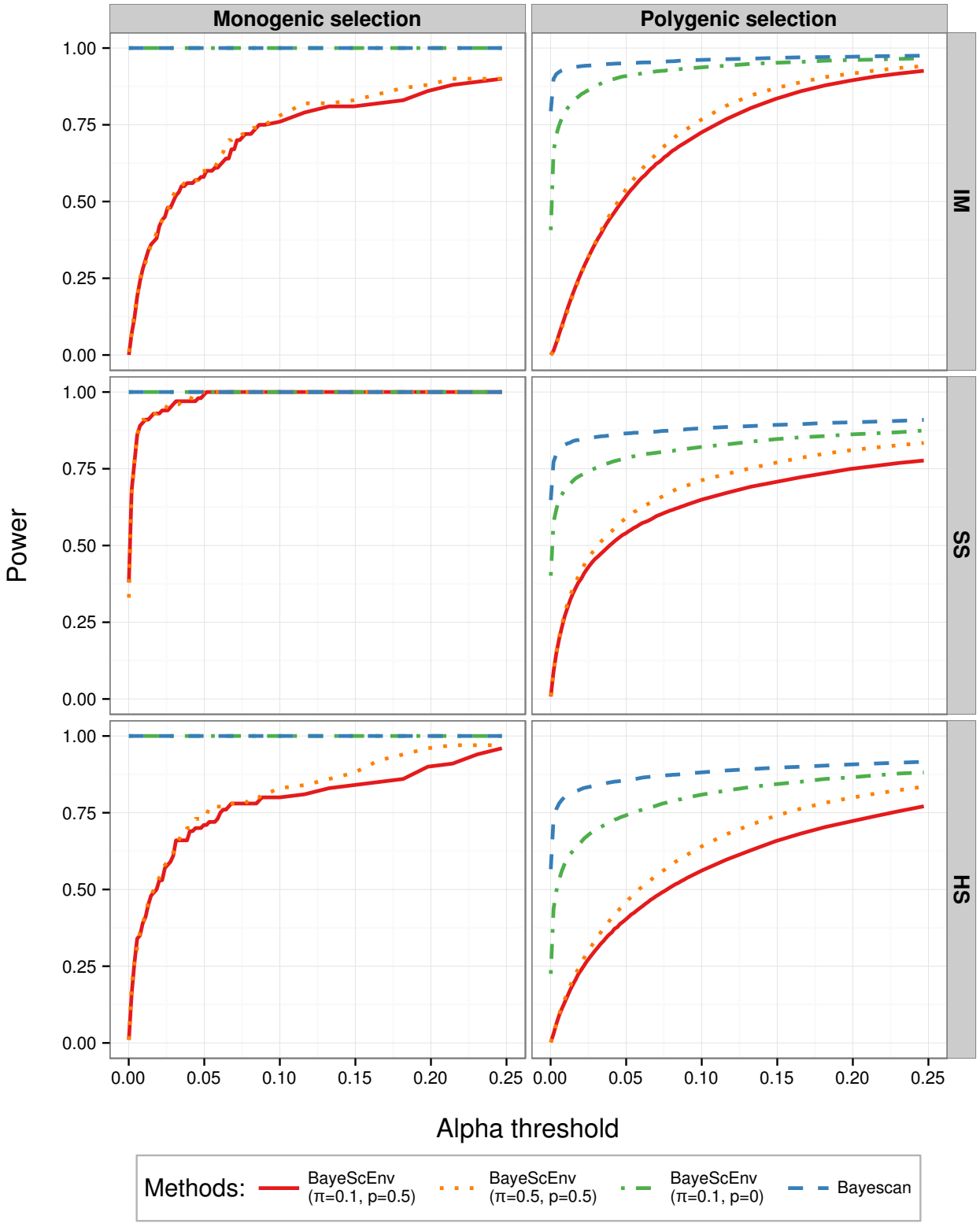

Figure 3: Power against significance threshold α for three scenarios (IM: Island model, SS: Stepping-Stone model and HS: Hierarchically Structured model) and monogenic/polygenic selection. The models tested are BayeScan (blue dashed), and BayeScEnv (orange dotted, green dot-dashed and solid red) with different probabilities π of jumping away from the neutral model (M1) and different preferences p for the locus-specific model (M3). Note that $p = 0$ means the environmental model (M2) is tested against the neutral one only.

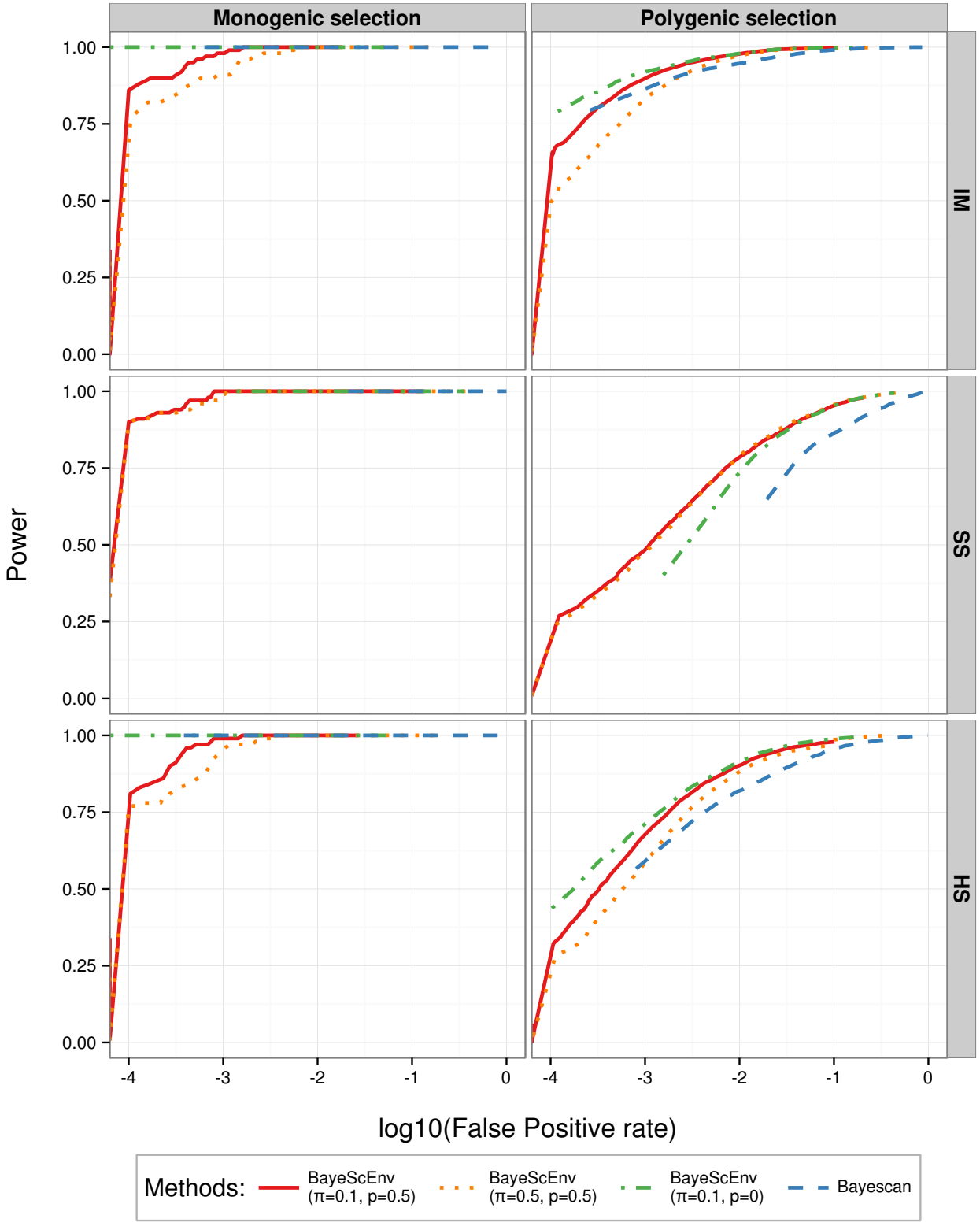

Figure 4: Power against False Positive Rate (FPR), a.k.a. ROC curve, for three scenarios (IM: Island model, SS: Stepping-Stone model and HS: Hierarchically Structured model) and monogenic/polygenic selection. The models tested are BayeScan (blue dashed), and BayeScEnv (orange dotted, green dot-dashed and solid red) with different probabilities π of jumping away from the neutral model (M1) and different preferences p for the locus-specific model (M3). Note that $p = 0$ means the environmental model (M2) is tested against the neutral one only.

Results

Simulation results

By definition, a threshold value of α used to decide whether q-values are significant or not is expected to yield an FDR of α on the long run, when the model is robust and priors are calibrated. Recall that in BayeScEnv all q-value tests below were performed on the parameter g to test for local adaptation. In the case of BayeScan, on the other hand, the q-values correspond to parameter α.

As shown in Fig. 2, BayeScan was less well calibrated, yielding higher FDRs than BayeScEnv under all scenarios and for both monogenic and polygenic selection. Additionally, for BayeScEnv, the implementation using $\pi = 0.1$ was fairly well calibrated (i.e. the curve is close the grey line in Fig. 2) under the IM scenario (for both monogenic and polygenic versions) and under the polygenic version of the HS scenario. This implementation was much more conservative than the one using $\pi = 0.5$. For $\pi = 0.1$ and $p = 0$, the FDRs were closer to those yielded by BayeScan, but still lower.

The higher FDR for BayeScan and BayeScEnv with $\pi = 0.5$ or $p = 0$ was mainly driven by a higher FPR rather than a lack of power (Fig. 3, see also Fig. S3 in the SI). Notably though, BayeScan had a quite high power, higher than that of BayeScEnv. Note, however, that BayeScEnv with $p = 0$ had, as BayeScan, a maximal power in the monogenic scenarios, and was almost as powerful as BayeScan in the polygenic scenarios. Yet its FDR was lower (sometimes much lower) than that of BayeScan. This indicates that the incorporation of environmental data helps to reduce the error rate both with or without the inclusion of spurious locus-specific effects (α_i). More details regarding the FPR results are available in the Supplementary Information (Fig. S3).

Another traditional way to apprehend the compromise between power and false positives is the so-called Receiver Operating Characteristics (ROC) curve, plotting power against FPR (Fig. 4). In these plots, the curve that is "more to the left" is preferred because this means it offers higher power for a lower FPR. Fig. 4 shows that BayeScEnv with $\pi = 0.1$ and $p = 0$ performed best under the IM and HS scenarios, whereas BayeScEnv with $\pi = 0.1$ and $p = 0.5$ performed better under the "harder" SS scenario. Overall, although BayeScan has higher power to detect local adaptation, it is still too liberal when deciding that a locus is under selection for the scenarios we investigated.

The heterogeneous mutation scenario lead to a dramatically high false positive rate for the low mutating loci in the case of Bayescan (62%). BayeScEnv, on the other hand, yielded a much lower false positive rate for these loci (4.9%). Of course, because the higher differentiation due to low mutation rate can be seemingly distributed according to the environmental variable, higher false positive rates will always be expected in such a scenario. Nevertheless, BayeScEnv is an improvement over Bayescan in that regard.

Method	Variable	Nr of significant SNPs	Nr of significant GO terms	Nr of genes associated with a significant GO term
BayeScEnv	altitude	154	32	11
	temperature	170	103	20
	precipitation	2728	439	359
BayeScan	—	66,316	469	5628

Table 1: Results from BayeScan and BayeScEnv on the human dataset. FDR significance threshold was set to 5%. The total number of tested markers was 446,117.

Analysis of human data from Asia

The results of the human dataset analysis (Table 1) show a dramatic discrepancy between the two methods. Whereas BayeScan yields a very large number (66,316) of markers considered as significant at the 5% threshold, many fewer markers (154 to 2728) are considered significant by BayeScEnv. Gene Ontology (GO) enrichment tests identified many significant terms (Table 1). Note, however, that in the altitude and temperature analyses they correspond to a small number of genes (11 and 20 respectively, see Table 1). The number of genes is larger for the precipitation analysis (359) and even larger for the analysis using BayeScan (5628).

Regarding the altitude, significant biological processes included the fatty acid metabolism (e.g. SCARB1), skin pigmentation (e.g. MLANA, SLC24A5), kidney activity (e.g. SLC12A1) and oxido-reductase activity (e.g. NOS1AP). Regarding the temperature, significant biological process included cardiac muscle activity (e.g. SLC8A1) and development (e.g. NRG1, FOXP1), fatty acid metabolism (e.g. FADS1, FADS2) and response to hypoxia (e.g. SLC8A1, SERPINA1). For the precipitation analysis with BayeScEnv, as well as the BayeScan analysis, the number of significant terms was too large for hand-picked examples to be feasible.

The significance results (q-values) are displayed as a Manhattan plot in Fig. 5, along with the above mentioned genes for the altitude and temperature analyses (Fig. 5, A and B). Other regions of the genome also include outlier loci but they correspond to non-coding regions, or are close to genes associated to GO terms that were not significant, or to proteins without a known function (e.g. C9orf91, which was the most significant gene in the temperature analysis). Pattern of linkage disequilibrium was visible, which sometimes strongly supported some candidate genes (Fig. 5, A, SLC12A1 and SLC24A5). Finally, comparing BayeScEnv (Fig. 5, A, B and C) and BayeScan analyses (Fig. 5,D), we see that BayeScan yielded too many significant markers for a Manhattan plot to be a useful display of the results. An interesting pattern is that BayeScan yielded far more outlier markers with maximal certainty (e.g. posterior probability of one) than BayeScEnv. For the present dataset, 22,516 markers had a posterior probability of one, whereas the maximal posterior probability yielded by BayeScEnv was 0.9998. Finally, almost all loci detected using BayeScEnv were also found when using BayeScan (between 98% for altitude to 100% for the two other variables).

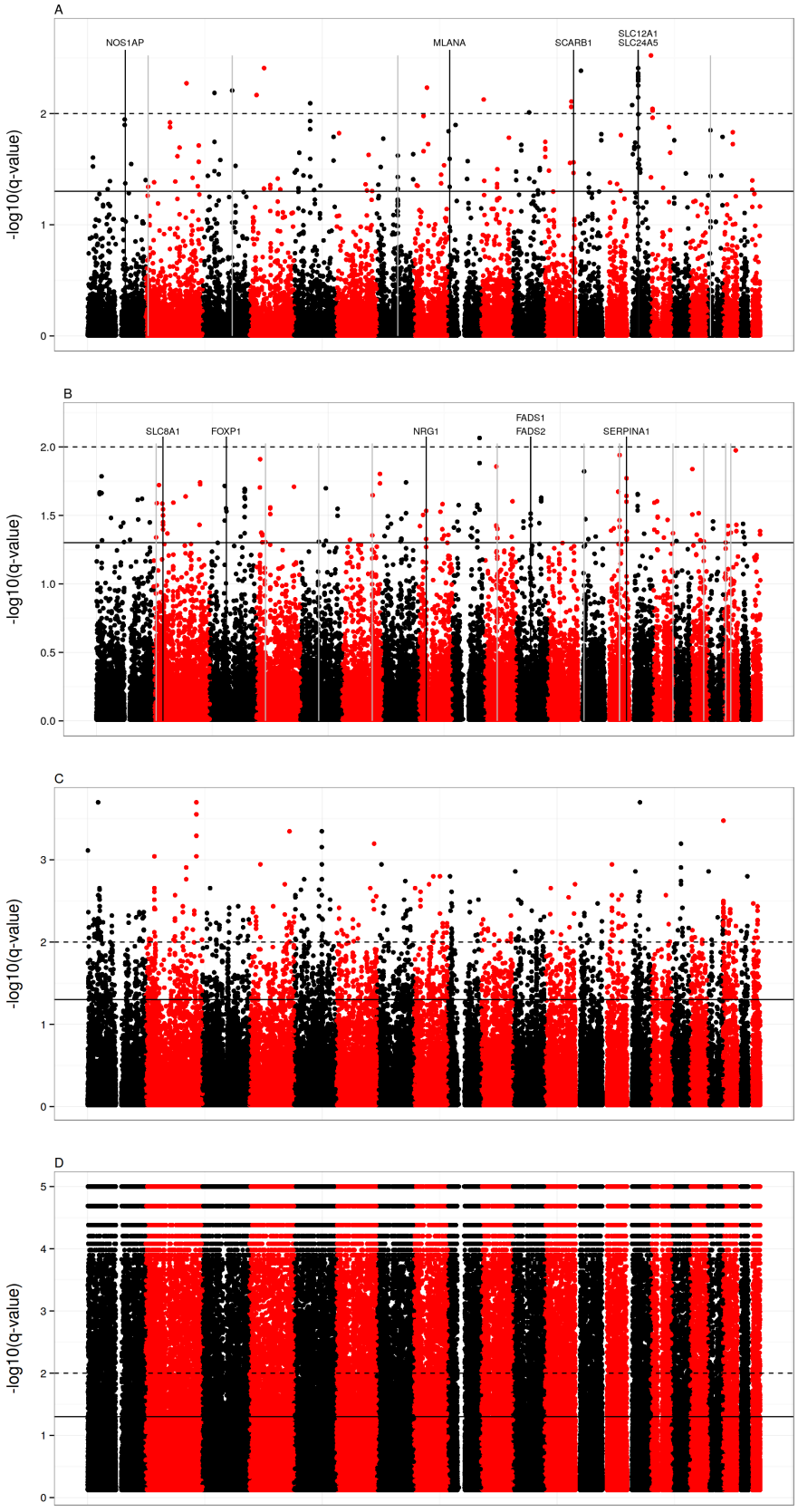

Figure 5: Manhattan plot of the q-values for the human dataset when using BayeScEnv with altitude (A), temperature (B), precipitations (C) or when using BayeScan (D). For altitude and temperature (A and B), genes mentioned in the text are displayed using black lines and genes associated with a significant GO term using grey lines. Top "stripes" for BayeScan (D) are artefacts due to finite number of iterations in RJMCMC (e.g. 0, 1, 2, 3... iterations outside of the non-neutral model), corresponding to determined posterior probabilities when divided by the total number of iterations.

Discussion

Features and performance of the method

The method we introduce in this paper, BayeScEnv, has several desirable features. First, just as BayeScan, it is a model-based method. This means that the null model can be understood in terms of a process of neutral evolution. One can thus predict what the method is able to fit or not. Second, we explicitly model a process of local adaptation caused by an environmental variable. Third, in order to render the model more robust, we account for locus-specific effects unrelated to the environmental variable under consideration. These departures can be due to another process of local adaptation (i.e. caused by unknown environmental variables), to large differences in mutation rates across loci, to background selection (Charlesworth, 2013) or complex spatial effects, such as allele surfing (Edmonds *et al.*, 2004) and hierarchical population structure (Excoffier *et al.*, 2009). Our simulation results show that when compared to BayeScan, BayeScEnv has a better control of its false discovery rate under various scenarios (Fig. 2), yielding fewer, but more reliable candidate markers. Obviously, this has a cost in terms of absolute power (Fig. 3), but BayeScEnv still performs better than BayeScan in terms of the investigated compromises between true and false positives (i.e. FDR and ROC, Fig.2 & 4).

Besides, the parametrisation of BayeScEnv allows for a fine and intuitive control of the false positive rate and power. For example, setting p to 0 increases both power and false positive rate, whereas setting $p = 0.5$ will allow for a more conservative test. This is because with $p = 0$, that is when the locus-specific effect model (M3) is excluded, the local adaptation model (M2) will absorb much of the signal in the data, yielding a higher probability of detecting true positives, but also a higher sensitivity to false positives. Our simulation results show that, if the species under study has moderate to large dispersal abilities (c.f. hierarchical structure or island model), the former parametrisation will be more appropriate, whereas for species with low dispersal abilities (c.f. stepping-stone model) the latter should be preferred. Thus, being able to choose the right parametrisation only requires limited knowledge about the dispersal abilities of the species.

We note that BayeScan was recently extended to consider species with hierarchical population structure (BayeScan3, Foll *et al.*, 2014). With BayeScan3 it is now possible to study widely distributed species covering several continents or geographic regions. It is also possible to better focus on local adaptation by considering groups that include pairs of populations inhabiting different environments such as low and high altitude habitats. Thus, BayeScan3, allows for the consideration of categorical environmental variables. Our new approach on the other hand, allows the study of local adaptation related to continuous environmental variables in species with a more restricted range.

How to quantify 'environmental differentiation'?

To model local adaptation, we compute an "environmental differentiation" in terms of the distance (absolute value) to a reference value. Although this reference can conveniently be chosen as the average of the environmental values across the sampled populations, other kinds of reference may be biologically more relevant. For example, in our analysis of the effect of elevation in humans, it seems appropriate to use sea level as the reference. Indeed, given the kind of environmental variables elevation is a proxy for (e.g. partial pressure of oxygen, temperature, solar radiation, etc.), for most systems we would consider the sea level as a neutral environment rather than the differentiated one.

Another way to account for environmental differentiation is to use Principal Component Analysis (PCA), providing one of the axes to BayeScEnv as a description of the distance between environments. Despite this practice being an elegant way to summarise environmental distance between populations, it also has the drawback of making it more difficult to identify the "causal" variable.

Note that the environmental variables must be standardised so as to avoid scale inconsistencies between g and α and β. If we choose the average environmental value as reference, then standardisation involves mean-centring and rescaling to have unit variance. However, if we choose another reference, then standardisation only involves rescaling to have unit variance.

Finally, the software implementation of our method only accepts one environmental variable at the time as including more than one variable would considerably slow the algorithm down, and render the biological interpretation of g quite tedious. Also, when using several correlated variables, it is important to realise that statistically distinguishing between the relative selective roles of each one would require many populations.

Comparison with other environmental association methods

There are several genome-scan approaches that incorporate environmental information, such as Bayenv (Coop *et al.*, 2010), LFMM (Frichot *et al.*, 2013) and gINLAnd (Guillot *et al.*, 2014). These methods perform a regression between allele frequencies and environmental values. Yet non-equilibrium situations combined with complex spatial structuring can lead to spatial correlations in allele frequencies, which in turn can lead to high false positive rates. To minimise this problem, the above methods take into account allele frequency correlations across populations while performing the regression.

BayeScEnv, on the other hand, assumes that all populations are independent, exchanging genes only through the migrant pool. However, it includes a locus-specific effect unrelated to the environmental variable that helps to take into account locus-specific spatial effects due to deviations from the underlying demographic model. The fact that this approach works is illustrated by our simulation study, which showed that BayeScEnv was fairly robust to isolation-by-distance and a hierarchically structured scenario. Moreover, the analyses of simulated datasets from de Villemereuil *et al.* (2014), available in the SI, show that even under very complex scenarios,

BayeScEnv can compete with other environmental association methods. In particular, most of these scenarios assume an environmental selective gradient confounded with population structure, which is particularly hard for genome scan methods (Frichot *et al.*, 2015): the results show that, in that case, BayeScEnv suffer from low power, but not from an excess of false positives. When compared with the other methods (including Bayenv and LFMM), BayeScEnv typically yields a medium FDR for most scenarios, and is less scenario-sensitive than Bayenv and LFMM. Nevertheless, we note that BayeScEnv is best suited for species with medium to high dispersal abilities such as marine species and anemophilous plants.

Another point that distinguishes BayeScEnv from these methods is that it does not assume any particular functional form for the relationship between environmental values and allele frequencies. While existing association methods all assume a clinal pattern, BayeScEnv only assumes that genetic differentiation increase exponentially with environmental differentiation. This allows for a more diverse family of relationships between allele frequencies and the environment.

Finally, BayeScEnv is one of the very few methods to study gene-environment associations that can be used with dominant data (but see also Guillot *et al.*, 2014).

Data analysis

When confronted with real datasets, BayeScEnv typically returned fewer significant markers than BayeScan. This is explained both by the focus on searching for outliers linked to a specific environmental factor and by the lower false positive rate of our approach. When applied to the human dataset, BayeScEnv identified several genomic regions that are enriched for gene ontology terms relevant to potential local adaptation to altitude or temperature. We emphasise that this study was not meant to exhaustively and rigorously investigate local adaptation in Asian human populations. However, our results tend to demonstrate that the candidates yielded by BayeScEnv have a biological interpretation. For example, skin pigmentation and cardiac activity could clearly be involved in responses to increased solar radiation and depleted oxygen availability at high elevation.

Much of the ontologies linked to temperature were potentially confounded with adaptation to altitude, such as the response to hypoxia and cardiac muscle activity. Also, fatty acid metabolism was associated to both altitude and temperature. Of course, the biological functions described here do not account for all the signals yielded by BayeScEnv (see Fig. 5, A and B). Other genomic significant regions include genes with less obvious biological function regarding local adaptation, non-coding regions and proteins without a known function. Finally, the analysis using the precipitation variable yielded too many significant markers for a detailed analysis of the biological functions involved. This may not necessarily be due to a confounding effect of the spatial structure (the human Asian populations being structured mainly from West to East, while the Eastern climate is characterised by strong precipitations during the monsoon), since precipitation may behave as a surrogate for several environmental variables.

Conclusion

The main improvement introduced by our new method, BayeScEnv, over existing F_{ST}-based genome-scan approaches is the possibility of focusing on the detection of outlier loci linked to genomic regions involved in local adaptation and better distinguishing between the signal of positive selection and that of other locus-specific processes such as mutation (see the heterogeneous mutation rate scenario in the Results) and background selection. Although it does not explicitly model complex spatial effects, the consideration of two different locus-specific effects make it more robust to potential deviations from the migrant pool model. This is reflected in its much lower false discovery rate when compared to BayeScan.

Our new formulation also allows for an improved control of the true/false positives compromise through the parameter p, which describes our preference for the model that includes a locus-specific effect unrelated to the environmental factor over the model that includes environmental effects. Although we recommend using $p = 0.5$, lower values (including 0) could be used if population structure is weak or maximising power is more important than reducing the false positive rate.

With this new method, there are now three alternative formulations of genome-scan methods based on the F model. BayeScan detects a wide range of locus-specific effects (including background selection). Although its false discovery rate is higher than that of the two extensions, it is able to detect regions of the genome subject to purifying selection. The hierarchical version of this original formulation, BayeScan3, allows the study of local adaptation due to categorical environmental factors. Finally, our new method, BayeScEnv, is more appropriate to detect genomic regions under the influence of selective pressures exerted by continuous environmental variables. Thus, all three methods are complementary and jointly cover scenarios applicable to a wide range of species

Acknowledgement

We thank M. Foll for providing the source code of BayeScan and for clarifying several issues related to the code, J. Renaud for his help on getting the average altitude out of the HGDP latitude/longitude data, S. Schoville for the BIOCLIM data, E. Bazin for his help on the HGDP data analysis. PdV was supported by a doctoral studentship from the French *Ministère de la Recherche et de l'Enseignement Supérieur*. OEG was supported by the Marine Alliance for Science and Technology for Scotland (MASTS).

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, **12**(12), 1805–1814.

404 Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome
405 scans. *Molecular Ecology*, **13**(4), 969–980.

406 Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, López Herráez D,
407 Brutsaert T, Parra EJ, Moore LG, Shriver MD (2010) Identifying Signatures of Natural Selection in Tibetan
408 and Andean Populations Using Dense Genome Scan Data. *PLoS Genet*, **6**(9), e1001116.

409 Blanquart F, Kaltz O, Nuismer SL, Gandon S (2013) A practical guide to measuring local adaptation. *Ecology*
410 *Letters*, **16**(9), 1195–1205.

411 Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, SanCristobal M (2010) Detecting selection
412 in population trees: the Lewontin and Krakauer test extended. *Genetics*, **186**(1), 241–262.

413 Brooks S (1998) Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society:*
414 *Series D (The Statistician)*, **47**(1), 69–100.

415 Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce vari-
416 ability. *Molecular Biology and Evolution*, **15**(5), 538–543.

417 Charlesworth B (2013) Background Selection 20 Years on: The Wilhelmine E. Key 2012 Invitational Lecture.
418 *Journal of Heredity*, **104**(2), 161–171.

419 Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci
420 Underlying Local Adaptation. *Genetics*, **185**(4), 1411–1423.

421 Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L (2013) Evi-
422 dence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution*, **30**(7),
423 1544–1558.

424 De Mita S, Thuillet AC, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y (2013) Detecting selection
425 along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing
426 populations. *Molecular Ecology*, **22**(5), 1383–1399.

427 Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome Scans for Detecting Footprints of Local Adaptation
428 Using a Bayesian Factor Model. *Molecular Biology and Evolution*, **31**(9), 2483–2495.

429 Edelaar P, Burraco P, Gomez-Mestre I (2011) Comparisons between QST and FST—how wrong have we been?
430 *Molecular Ecology*, **20**(23), 4830–4839.

431 Edmonds CA, Lillie AS, Cavalli-Sforza LL (2004) Mutations arising in the wave front of an expanding population.
432 *Proceedings of the National Academy of Sciences of the United States of America*, **101**(4), 975–979.

433 Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population.
434 *Heredity*, **103**(4), 285–298.

435 Faria R, Renaut S, Galindo J, Pinho C, Melo-Ferreira J, Melo M, Jones F, Salzburger W, Schluter D, Butlin R
436 (2014) Advances in Ecological Speciation: an integrative approach. *Molecular Ecology*, **23**(3), 513–521.

437 Foll M, Gaggiotti OE (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant
438 and Codominant Markers: A Bayesian Perspective. *Genetics*, **180**(2), 977 –993.

439 Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L (2014) Widespread Signals of Convergent Adaptation
440 to High Altitude in Asia and America. *The American Journal of Human Genetics*, **95**(4), 394–407.

441 Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for Associations between Loci and Environ-
442 mental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, **30**(7), 1687–1699.

443 Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O (2015) Detecting adaptive evolution based
444 on association with ecological gradients: Orientation matters! *Heredity*.

445 Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R (2011) Signatures of Environmental
446 Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLoS*
447 *Genet*, **7**(11), e1002355.

448 Gaggiotti OE, Foll M (2010) Quantifying population structure using the F-model. *Molecular Ecology Resources*,
449 **10**(5), 821–830.

450 Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.
451 *Biometrika*, **82**(4), 711–732.

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

455 Günther T, Coop G (2013) Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, **195**(1),
456 205–220.

457 Kofler R, Schlötterer C (2012) Gowinda: unbiased analysis of gene set enrichment for genome-wide association
458 studies. *Bioinformatics*, **28**(15), 2084–2085.

459 Kruuk LEB, Baird SJE, Gale KS, Barton NH (1999) A comparison of multilocus clines maintained by environ-
460 mental adaptation or by selection against hybrids. *Genetics*, **153**(4), 1959–1971.

461 Käll L, Storey JD, MacCoss MJ, Noble WS (2008) Posterior Error Probabilities and False Discovery Rates:
462 Two Sides of the Same Coin. *Journal of Proteome Research*, **7**(1), 40–44.

Larsson LC, Laikre L, Palm S, André C, Carvalho GR, Ryman N (2007) Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. *Molecular Ecology*, **16**(6), 1135–1147.

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**(12), 981–994.

Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefánsson K, Donnelly P (2002) Assessing Population Differentiation and Isolation from Single-Nucleotide Polymorphism Data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **64**(4), 695–715.

Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**(18), 3686–3687.

Riebler A, Held L, Stephan W (2008) Bayesian Variable Selection for Detecting Adaptive Genomic Differences Among Populations. *Genetics*, **178**(3), 1817–1829.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**(10), 1135–1145.

Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 479–498.

de Villemereuil P, Frichot E, Bazin E, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**(8), 2006–2019.

Data Accessibility

The Python code used to simulate data is available online in the Supplementary Information. The software and its source code are available online at GitHub: `http://github.com/devillemereuil/bayescenv`. The HGDP dataset is available at `http://www.hagsc.org/hgdp/files.html`. The BIOCLIM database is available at `http://worldclim.org/bioclim`.

Author contributions

PdV and OEG designed the statistical model. PdV modified the C++ code and performed the simulation and data analysis. PdV and OEG wrote the article.