Genome scan methods against more complex models: when and how much should we trust them?

Pierre de Villemereuil[*], Éric Frichot[†], Éric Bazin[*], Olivier François[†] & 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

[†]: Université Joseph Fourier Grenoble, Centre National de la Recherche Scientifique,

TIMC-IMAG UMR 5525, 38042 Grenoble, France

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

Fife, KY16 8LB, United Kingdom

Keywords: genome scan, adaptation, Bayesian methods, false discovery rate, power simulation study

Abstract

The recent availability of Next Generation Sequencing (NGS) has made possible the use of dense genetic markers to identify regions of the genome that may be under the influence of selection. Several statistical methods have been developed recently for this purpose. Here, we present the results of an individual-based simulation study investigating the power and error rate of popular or recent genome-scan methods: linear regression, Bayescan, BayEnv and LFMM. Contrary to previous studies, we focus on complex, hierarchical population structure and on polygenic selection. Additionally, we use a False Discovery Rate (FDR) based framework, which provides an unified testing framework across frequentist and Bayesian methods. Finally, we investigate the influence of population allele frequencies *versus* individual genotype data specification for LFMM and the linear regression. The relative ranking between the methods is impacted by the consideration of polygenic selection, compared to a monogenic scenario. For strongly hierarchical scenarios with confounding effects between demography and environmental variables, the power of the methods can be very low. Except for one scenario, Bayescan exhibited moderate power and error rate. BayEnv performance was good under non-hierarchical scenarios while LFMM provided the best compromise between power and error rate across scenarios. We found that it is possible to greatly reduce error rates by considering the results of all three methods when identifying outlier loci.

Introduction

The detection of signatures of selection has been a long-standing interest of population geneticists and evolutionary biologists. However, until recently, the paucity of molecular markers available limited the power of statistical methods to detect selection because other biological process such as structure and migration have confounding effects on polymorphism and linkage disequilibrium. This situation has changed radically with the advent of Next Generation Sequencing (NGS, see Shendure and Ji, 2008), which can generate dense arrays of markers, typically Single Nucleotide Polymorphisms (SNPs), spread across the genome. These new data can be used to distinguish between neutral processes that have a genome-wide effect (e.g. demographic history) and processes that have a local effect, particularly selection (Luikart *et al.*, 2003). Several so-called genome-scan methods have been developed for this purpose (reviewed in De Mita *et al.*, 2013).

One of the most popular types of methods is based on an idea first proposed by Lewontin and Krakauer (1973). The underlying rationale is that loci influenced by directional selection will show larger genetic differentiation than neutral loci while the opposite is true for loci subject to balancing selection. Thus, loci that exhibit unusually high or low F_{ST} are good candidates for being influenced by selection. Several variants of this test exist (e.g. Beaumont and Nichols, 1996; Vitalis *et al.*, 2001; Beaumont and Balding, 2004; Foll and Gaggiotti, 2008) and have been frequently applied to non-model species. Another recent group of genome-scan methods is based on the idea that many selected loci should be correlated with the environmental factors underlying the selective pressure (Joost *et al.*, 2007; Coop *et al.*, 2010; Frichot *et al.*, 2012). Genotype-environment association methods identify loci that show strong correlations with one or more environmental variables, and those loci are interpreted as potentially under selection.

All genome scan methods are based on the premise that it is possible to clearly distinguish between the genetic signals left by neutral and non-neutral processes. However, this assumption is frequently violated in real life scenarios (Hermisson, 2009). Several demographic processes such as allele surfing (Edmonds *et al.*, 2004) and bottlenecks can leave signatures that mimic those left by positive selection. Moreover, complex spatial structuring can increase the variance of genetic parameters across the genome leading to high false positive rates (Excoffier *et al.*, 2009). Sensitivity analyses published thus far (Pérez-Figueroa *et al.*, 2010; De Mita *et al.*, 2013) focus on these confounding effects of demographic covariance among populations arising through migration. The overall pattern that emerges from these studies is rather positive. Although all evaluated methods suffer from either low power (differentiation-based methods) or high false-positive rates (genotype-environment association methods), a strategy based on the use of both types of methods seems to lead to reliable identification of outlier loci. Nevertheless, one limit of existing studies is that they consider the effect of selection on a single locus. This is a quite unjustified assumption because selection for a specific quantitative phenotypic trait will influence several regions across the genome (Rockman, 2012). There is only two studies (Narum and Hess, 2011; Vilas *et al.*, 2012) that considers several selected loci. However, the first is limited both by the

number of loci (only 5) and the number of replicates of the simulated data, while the second focuses on the question of whether or not detected outlier markers are physically close to selected loci.

In this study we focus attention on more realistic scenarios than those considered in previous analyses. In particular, we investigate biases that may arise when selection acts upon traits determined by several genes. Indeed, as Rockman (2012) recently pointed out, there is a paucity of empirical and theoretical support for the abundance of large-effect Quantitative Trait Nucleotides (QTNs) in the wild. Instead it is likely that "alleles that matter for evolution" are numerous small-effect loci. It is unclear if current genome-scan methods will simply have low power or if they will also have a high false discovery rate when applied to these situations. Another important consideration about real populations and species is that they are unlikely to be at migration-drift equilibrium. Thus, we evaluate scenarios where they have experienced recent divergence from an ancestral population, a process that may also affect power and false discovery rates of existing methods.

Instead of evaluating the performance of a very large number of methods we focus on a few that have proven popular or that are very recent and tested only under some restricted scenarios. More precisely, we focus on two genotype-environment association Bayesian methods that explicitly take into account the covariance of allele frequencies across populations (Coop *et al.*, 2010; Frichot *et al.*, 2012) and we compare these methods to one of the most frequently used genome-scan methods based on population differentiation (Foll and Gaggiotti, 2008). We did not include more population differentiation methods as they have been shown to be less efficient than this particular one (Pérez-Figueroa *et al.*, 2010; Vilas *et al.*, 2012; De Mita *et al.*, 2013). We further consider a naive frequentist regression approach without any correction for population structure. The comparison is done using a rigorous statistical framework based on false discovery rates (FDR, see Benjamini and Hochberg, 1995) and *q-values* (Storey and Tibshirani, 2003; Storey *et al.*, 2004), which allow for a unified comparison of the performance of the methods.

Material & Methods

Simulation model

We carried out simulations using the SimuPop package for Python (Peng and Kimmel, 2005). We focused on highly structured population scenarios where selection acts on a multigenic trait. For the sake of clarity we describe each component of the simulation model separately and also present the main attributes of each scenario in Table 1. We simulated 100 replicates for each scenarios (but only used 50 for Bayescan, see below).

Demographic process Our main scenario is a dichotomous process of population fission in which an ancestral population of 500 individuals gives birth to two descendant populations after 50 generations of drift. The fission is instantaneous with local populations reaching carrying capacity of 500 individuals in a single generation. This dichotomous fission process is repeated until 16 populations are obtained (see dendrogram, Fig. 1.A).

Migration occurs all along the process and preferentially between historically close populations: two populations issued from the same fission event will exchange twice as many migrants as two populations issued from two distinct fission events. In other words, the proportion of migrants between two populations is determined by phylogeographic distance. We aimed at capturing the main features of a spatial expansion in a heterogeneous habitat. For example, a post-glaciation colonisation scenario, where new valleys and sub-valleys are progressively reached. The further apart two populations are along the population tree, the lower the migration rate between them. This model corresponds to a highly structured Isolation with Migration model (noted HsIMM). We assume a recent demographic origin for all populations (500 generations in total since the initial fission event). In addition, we consider two simpler scenarios: an isolation with migration (IMM) model where the sixteen populations are issued from a single fission event and a stepping stone scenario (SS) where all sixteen populations are issued from a single fission event. For these two models the length of the runs was 400 generations. These settings allow us to stop the simulation at a near-equilibrium situation. In all scenarios, each population consists of 500 individuals. The proportion of individuals in a local population that do not migrate, $(1-m)$ is the same under all three scenarios but the proportion of individuals that migrate between pairs of populations differ. Under the HsIMM it is $m/2^{(i+1)}$ where i is the number of fission events between each local population and the most recent common ancestral population (SI eq. 1). Under the IMM it is $m/15$ for all pair of populations (SI eq. 2). Under the SS model it is equal to $m/2$ for neighbouring populations and zero for all other pairs of populations (SI eq. 3). For all simulations we chose $m = 0.0045$, which yielded pairwise F_{ST} roughly equal to 0.1.

More information about the simulation process can be found in the supplementary information (SI, section 1). The python code used can also be found online in the data accessibility section.

Genetic process We simulated 5000 SNP regularly spread along 10 chromosomes. The recombination rate between adjacent pairs of SNPs is set to 0.002 in order to have, on average, one recombination event per population per generation. This amounts to spacing 500 SNPs uniformly along each chromosome. The mutation rate is set to 10^{-7} per generation at each SNP. We consider two genetic architectures: either a single locus case, or 50, randomly distributed, loci influencing a phenotypic trait directly linked to fitness. In each case, we assume co-dominance.

We use a multiplicative fitness function to describe the 'cumulative' effect of all loci on fitness :

$$W = (1+s_P)^{n_{11}}(1-s_P)^{n_{00}} \tag{1}$$

where s_P is the local coefficient of selection (depending on the local value of the environment, see next paragraph) and n_{11} and n_{00} are the number of $(1,1)$ and $(0,0)$ homozygous loci, respectively. Note that fitness is normalized such that the relative fitness of any heterozygous locus is 1. For small s, this multiplicative fitness function is

equivalent to an additive one.

Environmental variable underlying the selective pressure In the case of a highly structured model (HsIMM) we consider two spatial patterns of selection intensities, which are determined by an environmental variable E_S: *(i)* at each population fission, the values of E_S for each descendant population are drawn from a uniform distribution centred on the value of the ancestral population (HsIMM-U) and, *(ii)* at each population fission, the values of E_S for the new populations are set such that they produce an environmental gradient along a linear habitat (HsIMM-C). For the isolation with migration (IMM) and stepping-stone (SS) scenarios, the values of E_S are also set to form an environmental gradient, like in case *(ii)* (Table 1).

The local coefficient of selection s_P is calculated as a logistic transformation of the environmental variable :

$$s_P = s \frac{1 - e^{-\beta E_S}}{1 + e^{-\beta E_S}} \tag{2}$$

where s is the 'baseline' selection coefficient and β is the 'slope' of the logistic transformation. For the scenario with a single selected locus we set s to 0.1 and β to 1. In the case of the polygenic scenario we use $s = 0.004$ and $\beta = 5$. The difference in parameter values between the two scenarios is necessary because, for size effect s in the monogenic case and s/N in the polygenic case, local adaptation progresses much more slowly under the polygenic architecture. Therefore, it was necessary to increase both the effect size and slope of the gradient for the polygenic case so as to generate local adaptation patterns under both scenarios in a similar evolutionary time. The values were scaled so that the mean allelic frequency pattern in the polygenic case was similar to the one in the monogenic case.

We also investigate the potential for spurious selection signals due to the consideration of environmental factors unrelated to any selective pressure. For this we consider scenarios that include a selectively neutral environmental variable E_0 whose values are randomly drawn from a normal distribution. Selection starts at the second fission events in the HsIM scenarios, and at the (only) first one in the two other scenarios.

Table 1: Description of the scenarios considered in this study

Scenario	Spatial Model	Demographic History	Selection Pattern
HsIMM-U	Hierarchical	Multiple Binary Fissions	Correlated with demographic history
HsIMM-C	Hierarchical	Multiple Binary Fissions	Environmental Gradient
IMM	Standard IMM	Instantaneous Fission	Environmental Gradient
SS	Stepping-Stone	Instantaneous Fission	Environmental Gradient

Statistical analysis

Error rate For all methods, we use *q-values* as a significance test statistic (Storey and Tibshirani, 2003; Storey *et al.*, 2004). The *q-value* is tightly linked to the False Discovery Rate (FDR) (Storey and Tibshirani, 2003). For a statistical test, the FDR is equal to the number of false positives over the total number of positives

(true and false). Thus, it is the proportion of "false discoveries" among all the "discoveries" of the test. If the assumptions of the test hold, then a given threshold α_q for assessing the significance of *q-values* should lead to a FDR of α_q. For example, if one decides a cut-off threshold of 5%, then the test will yield 95% of true positives and 5% of false positives. Note that, in this sense, a cut-off of 5% for *q-values* is much more stringent than the same cut-off for *p-values*. It is important to distinguish between false positive rate, false discovery rate and power: their relationship is explained further in SI. Note that, for the same dataset, an increase in power would lead to a decrease of FDR, whereas an increase in false positive rate (FPR) would lead to an increase of the FDR.

Power For monogenic selection scenarios, the definition of power is straightforward: it is the proportion of truly selected loci that are significant (see also Eqn. 6 in SI). For polygenic selection, this definition leads to a value of power for each locus. We computed power for each locus for each simulation, and then averaged over all loci, in order to get a mean power comparable to the case of monogenic scenarios. Note that, in the case of polygenic scenarios, we have less sampling error than in monogenic scenarios, because we have 50 times more selected loci.

Data specification Some methods can be applied either to population allele frequency data or to individual genotype data. In principle, using genotypic data is more appropriate when it is difficult to clearly define population boundaries. It can also avoid potential biases introduced by differences in sample sizes across populations. We investigated the influence of data specification for the linear regression and the Latent Factor Mixed Model methods (see below).

Genome scan methods to detect selection

There are several genome-scan methods aiming at detecting selection by identifying outlier loci. Here we focus on two genotype-environment association methods that explicitly take into account the allele frequency covariance across populations and we compare these methods to a genome-scan method based on population differentiation. We further consider a naive frequentist approach that test for correlations between allele frequencies and environmental factors.

BayEnv A first method that takes into account the allele frequency covariance across populations generated by demographic history and spatial effects was developed by Coop *et al.* (2010) and is implemented in the software BayEnv. This method consists in a two-step procedure. First, a model using all loci (or a part of the data set that is known to be neutral) estimates the population structure using a variance-covariance matrix of allele frequencies between populations. Second, a model incorporating the empirical covariance matrix tests for the correlation between the allele frequencies at each locus (or only at loci of interest) and each environmental

variable. The software returns Bayes factors weighting the strength of evidence in favour of a correlation between allele frequencies and the environmental variable. We transform the Bayes Factors into posterior odds using a prior probability of the null model $\pi_0 = 0.99$, and use these odds to compute *q-values* (Storey and Tibshirani, 2003; Storey *et al.*, 2004) which are used to assess the significance of each locus. The procedure is explained in the supplementary information (SI), where we also provide the MCMC parameters used.

Latent Factor Mixed Model Latent Factor Mixed Models (LFMMs, see Frichot *et al.*, 2012) are very general and flexible models and provide an alternative approach to detect relationships between allele frequencies and environmental values, while taking into account population structure. The model can be seen as an approximate Principal Component Analysis combined with a regression. It is computationally faster than BayEnv and Bayescan (Frichot *et al.*, 2012). The K value (number of factors) needed by the software are estimated to be 15 for every scenarios, using Tracey-Widom tests. The *p-values* returned by the method are transformed into *q-values* following a standard procedure (Storey and Tibshirani, 2003). We used the version 1.2 of the software.

Bayescan Bayescan (Foll and Gaggiotti, 2008) is an F_{ST}-based model (Beaumont and Balding, 2004). This method is not searching for a potential correlation between allele frequencies and the environment. Instead, it is searching for loci exhibiting extreme F_{ST} values. Large F_{ST}s are then interpreted as signatures of local adaptation. It is testing for outliers independently of any environmental knowledge. The statistical significance is assessed by the use of *q-values* (Storey and Tibshirani, 2003; Storey *et al.*, 2004) using a prior odds of 100. The MCMC parameters used are detailed in SI. Because of computation time issues, we only used 50 replicates for this method.

Corr: Allele frequencies-environment regression This is the most naive method and only aims at detecting a correlation between population allele frequencies and an environmental variable. Typically, significance is evaluated using the *p-value* returned by Student's test on the slope of the regression. However, to correct for multiple tests and to easily compare with results of other methods we transform the *p-values* into *q-values* using the method presented in Storey and Tibshirani (2003).

Results

Genetic structure produced by the population models

As expected, our simulation models produce highly structured population genetic data. Fig. 1 shows the structure of the correlation in allele frequencies across populations, as estimated by the software BayEnv (Coop *et al.*, 2010). For the binary fission model (Fig. 1.A), the strength of the correlation decreases with the phylogeographic distance. The isolation with migration model (Fig. 1.B) produces no apparent spatial pattern

while the stepping stone model (Fig. 1.C) leads to a typical isolation-by-distance pattern.

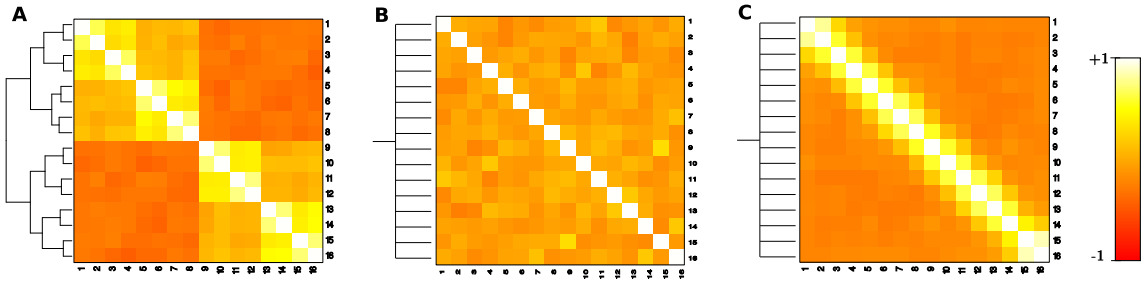

Figure 1: Heatmap of allele frequencies correlation between all simulated 16 populations. Panel A: HsIMM model; Panel B: IMM model; Panel C: SS model. The red to white gradient corresponds to the $[-1, 1]$ interval. The dendrogram illustrates proximity between populations (inferred for HsIMM, drawn for IMM and SS).

Monogenic selection

Error Rates The expectation is that the False Discovery Rate (FDR) increases linearly with the threshold used to decide the significance of q-values. However, the results differ radically from this expectation. Indeed, the FDR of all methods was higher than expected under all scenarios (Fig. 2), except for LFMM in the IMM scenario, which is even quite conservative. Note also that BayEnv has an acceptable FDR for the SS scenario and stringent thresholds (Fig. 2, SS). This inflation in FDR is partly due to the fact that, when only one locus is truly selected, even a small false positive rate, when combined with high power, leads to very high FDRs. Regarding hierarchical scenarios, when the spatial selection pattern is a function of phylogeographic distance (Fig. 2, HsIMM-U), FDRs are highest for Bayescan and lowest for LFMM, while the FDR values for BayEnv and the linear regression methods are intermediate. When selection is a function of an environmental gradient (Fig. 2, HsIMM-C), the FDR is highest for the linear regression method, intermediate for Bayescan and lowest for BayEnv and LFMM. Thus, the spatial pattern in selection intensities greatly influences the relative performance of the different methods. Note that the individual genotype data specification for the linear regression and LFMM (light lines) always lead to higher FDRs. This is especially the case for the linear regression with FDRs of almost 1. Note also that the linear regression method yields intermediate FDRs for small α thresholds for both scenarios. Finally, recall that FDRs are not on the same scale as False Positive Rates (FPR). Since here we are considering a monogenic scenario, a FDR of 75% corresponds to the truly selected locus plus 3 false positives, thus to a FPR of 0.06%.

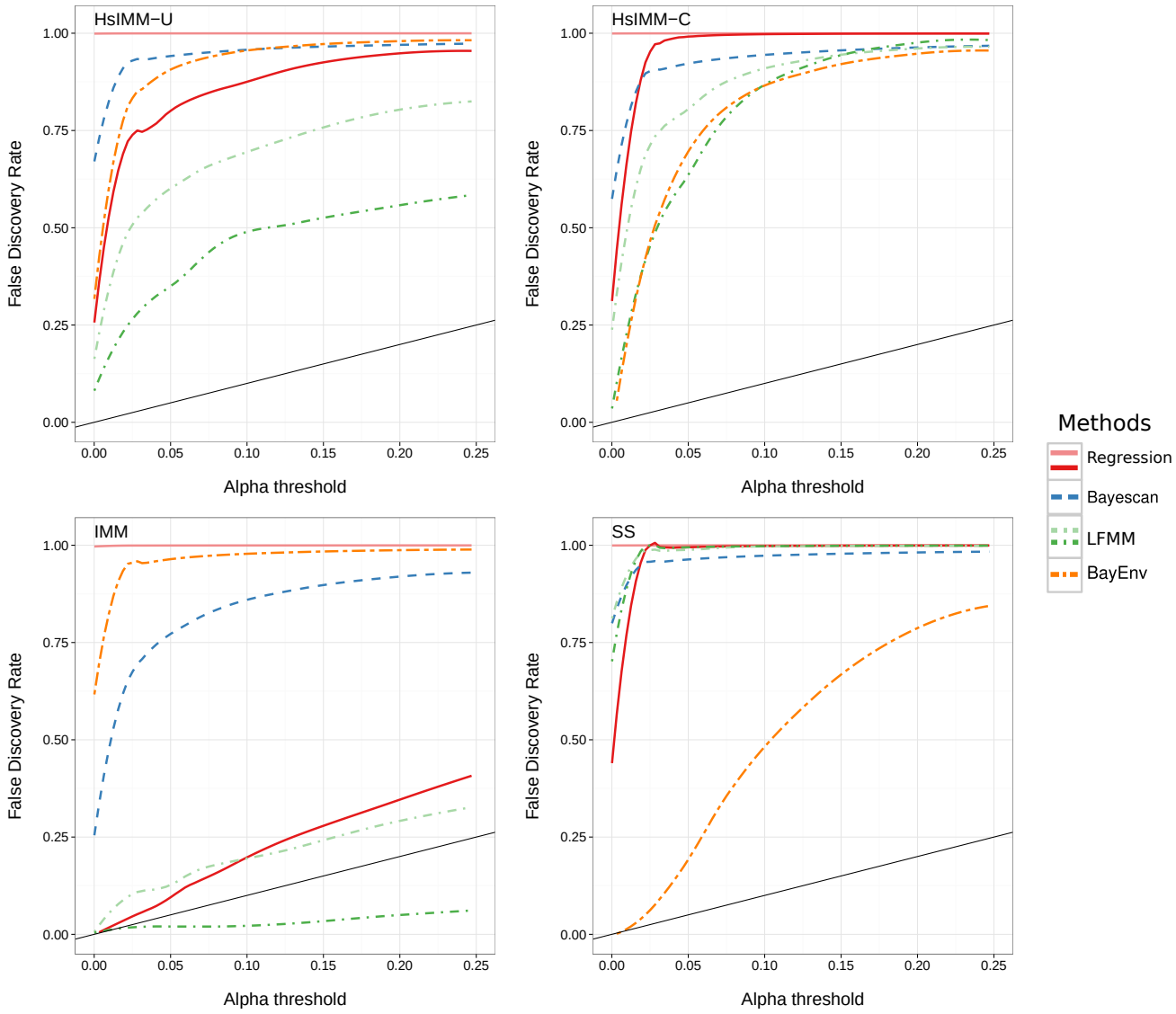

Figure 2: False discovery rate against significance threshold (α_q) for monogenic selection. Black line : Expected relationship between FDR and α_q. Lines are LOESS smooth for linear regression (plain red line), Latent Factor Mixed Model (LFMM, green dot-dashed line), Bayescan (blue dashed line) and BayEnv (orange two-dashed line). Light lines are for individual genotype data specification for the linear regression (light red) and LFMM (light green).

Statistical power Under the scenario HsIMM-U the power of all methods is moderate with a maximum between 75 and 80% for very permissive thresholds (Fig. 3., HsIMM-U, except the case of linear regression for individual genotype data specification, light red line). In this case, all recent methods had roughly similar power, although LFMM yields a lower one. The regression method has the lowest power for allele frequency data specification but the highest one when using individual genotypes. Under the other scenarios (Fig.3, HsIMM-C, IMM and SS), the power of all methods is very high, being perfect for some of them regardless of the threshold value used. Note that, in all cases, the regression model is always among the least powerful methods. Also,

whereas the "individual genotypes" specification always increase the power for the regression (light red lines in Fig. 3), this is not always the case for LFMM (light green lines).

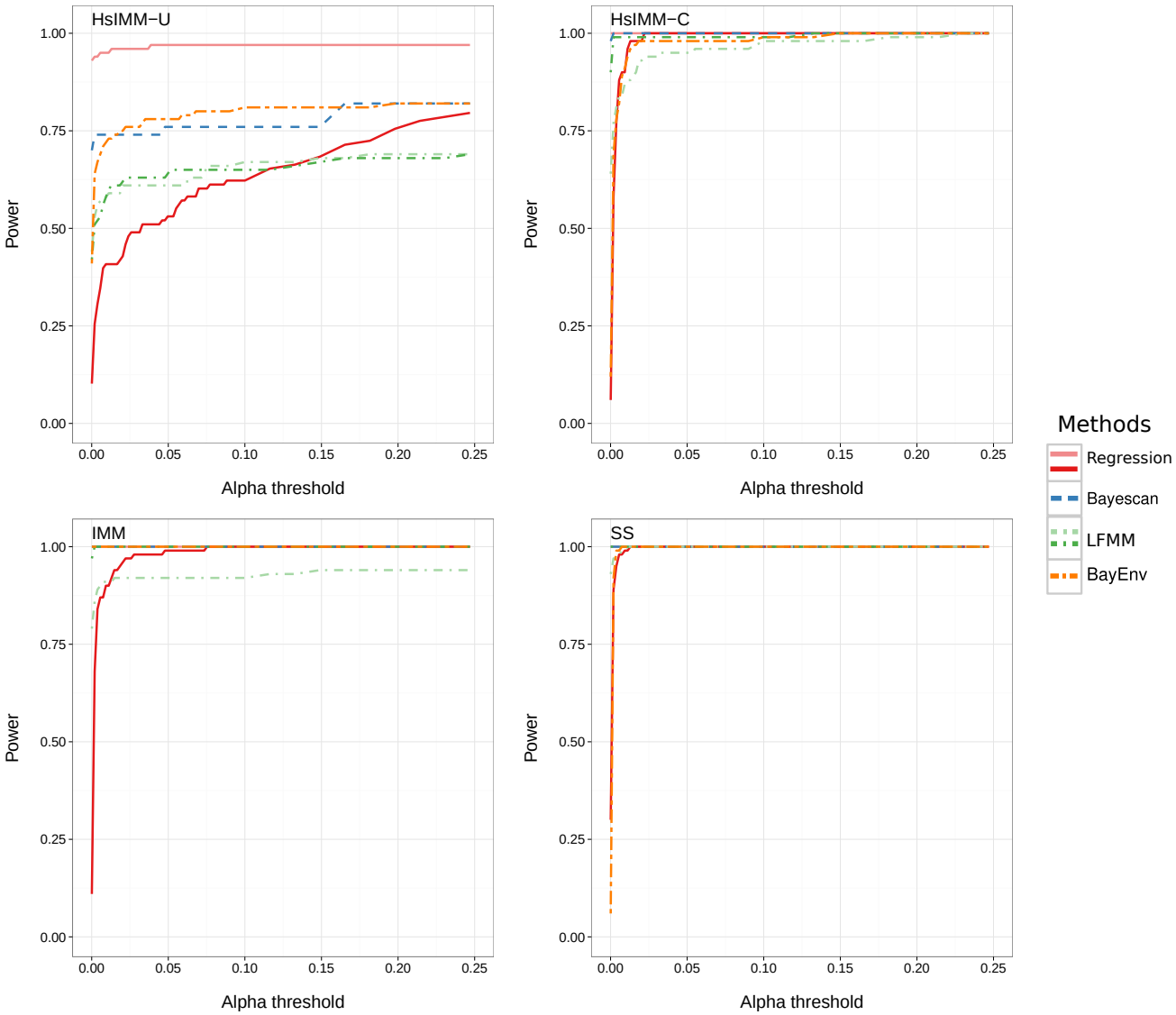

Figure 3: Statistical power against significance threshold for monogenic selection. Lines are linear regression (plain red line), Latent Factor Mixed Model (LFMM, green dot-dashed line), Bayescan (blue dashed line) and BayEnv (orange two-dashed line). Light lines are for individual genotype data specification for the linear regression (light red) and LFMM (light green).

Polygenic Selection

Error Rates As it was the case for the monogenic selection scenario, the false positive rate of all methods under all scenarios was higher than expected. Fig. 4 shows that the expected linear increase in FDR with increasing threshold values only holds for BayEnv under the stepping-stone model (Fig. 4, SS). Interestingly, LFMM shows a very conservative pattern for the IMM scenario, when using the population frequency data

specification (Fig. 4, IMM, dark green line). All other combinations of scenarios and methods are more error-prone than our theoretical expectation. Note in particular that all methods have very high FDR under the hierarchically structured IMM scenarios. While LFMM is the most conservative method in the case of environment correlated with demography (Fig. 4, HsIMM-U), BayEnv and Bayescan are the approaches that are the least error prone for a clinal environment (Fig. 4, HsIMM-C). The behaviour of LFMM and BayEnv changes radically across scenarios and they seem specially well adapted to a specific scenario (IMM and SS respectively) while the error rate of Bayescan is more intermediate across the different scenarios, although it is one of the worst under the standard IMM model. Regarding data specification, LFMM seems to be quite robust to its influence, although the individual specification still tends to yield more erroneous results than its allele frequency counterpart. The linear regression model, however, is much less robust: its individual genotype specification version is always the most error-prone, while its population allele frequencies specification can yield relatively conservative results (e.g. see Fig. 4, IMM).

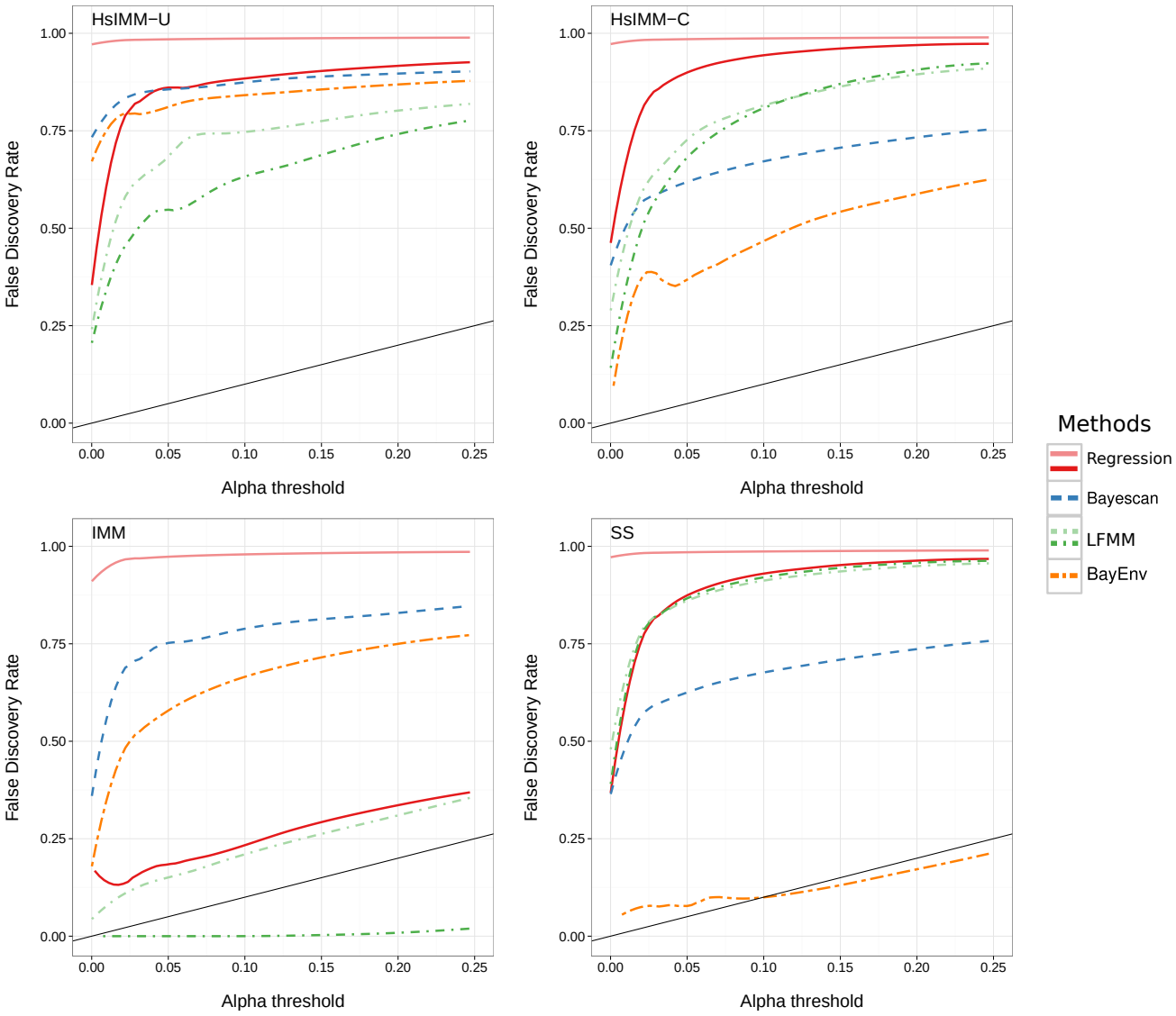

Figure 4: False discovery rate against significance threshold (α_q) for polygenic scenarios. Black line : Expected relationship between FDR and threshold value α_q. Lines are LOESS smooth for linear regression (plain red line), Latent Factor Mixed Model (LFMM, green dot-dashed line), Bayescan (blue dashed line) and BayEnv (orange two-dashed line). Light lines are for individual genotype data specification for the linear regression (light red) and LFMM (light green).

Statistical power Because of the small effect size of each locus under the polygenic model, the power of all methods should be lower than under the single-gene model. Indeed, we do observe an overall decrease in power for all scenarios (Fig. 5 compared to Fig. 3). The linear regression is the method that had the highest power under scenarios HsIMM-U, HsIMM-C. This power performance is followed by LFMM (Fig. 5 HsIMM-U and HsIMM-C). These two methods are comparable for the SS scenario (Fig. 5). Regarding these three models (HsIMM-U, HsIMM-C and SS), Bayescan shows intermediate power and BayEnv is the least powerful method. Interestingly, the behaviour of the methods is very different for the IMM scenario (Fig. 5): here, BayEnv is one

of the most powerful methods, only outperformed by the error-prone linear regression in its individual genotype specification. LFMM and Bayescan are the two worst methods.

While having a high power is an interesting feature, it needs to coincide with reasonable False Positive and False Discovery rates to be relevant. Power against False Positive (ROC curves) and False Discovery rates are provided in SI. The ROC curves (Fig. IV in SI) illustrate the compromise between the number of true and false positives and show that all methods are comparable in this regard. The "power against FDR" graphs (Fig. VII in SI) provide information about how many true positives are detected by the methods. For a given FDR, more power means more true (and false) positives.

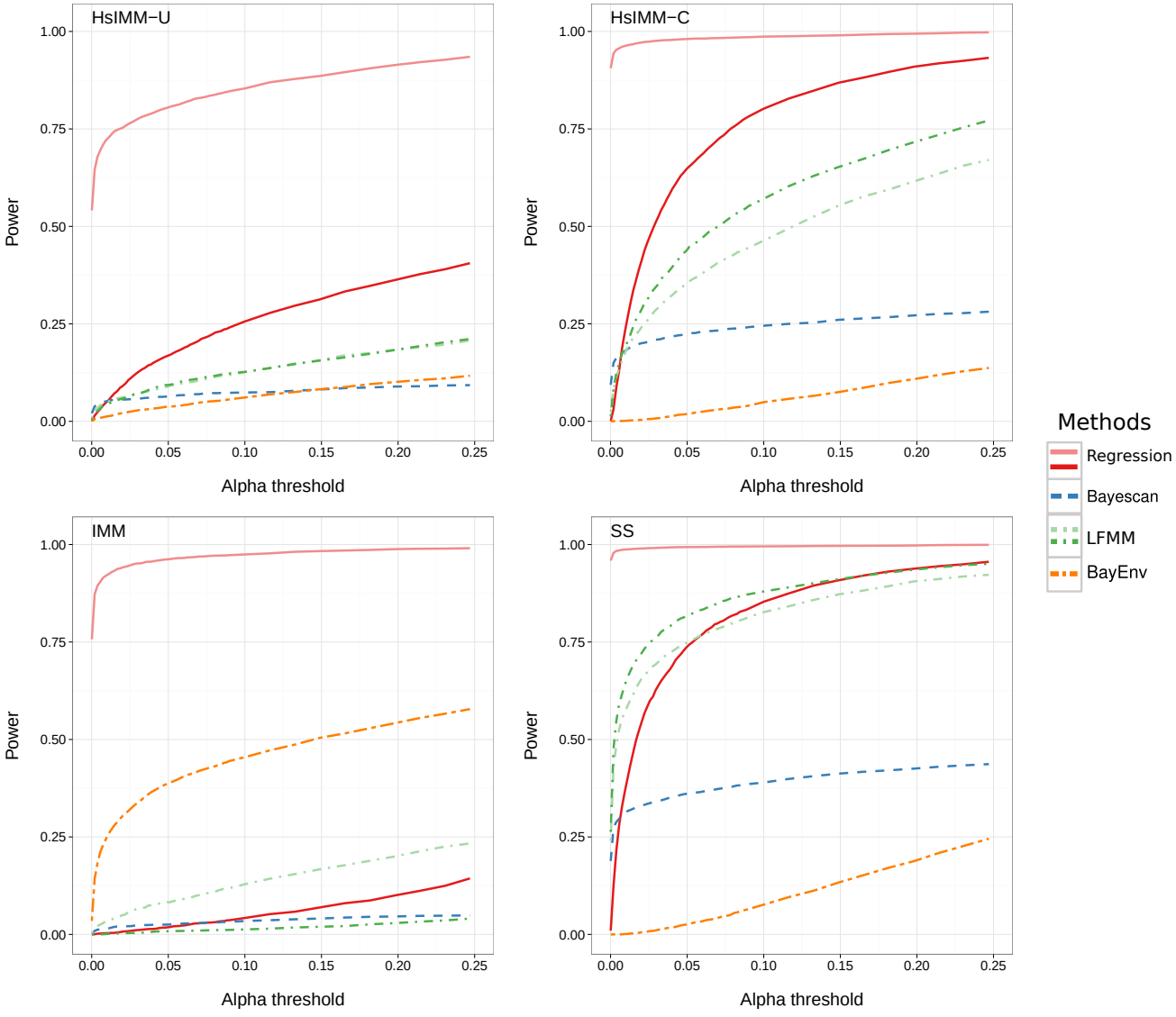

Figure 5: Statistical power against significance threshold for polygenic selection. Lines are for linear regression (plain red line), Latent Factor Mixed Model (LFMM, green dot-dashed line), Bayescan (blue dashed line) and BayEnv (orange two-dashed line). Light lines are for individual genotype data specification for the linear regression (light red) and LFMM (light green).

Consistency between methods Overall the methods tend to disagree from each other, in terms of which loci should be considered as selected (true or false positives). The percentage of overlap between loci considered as positives by two different methods is around 1% to 5%, except for the regression and LFMM (14% to 48% depending on the scenarios). Notable exceptions are the HsIMM scenarios, where Bayescan and LFMM reach an agreement on 13% of loci under selection for HsIMM-U and 18% for HsIMM-C. Still, the methods are more often in agreement regarding true positives than regarding false positives. This means that using 3 methods to assess the outlier behaviour of loci leads to a substantial decrease of the FDR. This decrease varies between 0.4 and 0.65, depending on the scenarios. For the IMM model, this strategy yields a FDR of 0% (all positives are true positives). Unfortunately, using several methods leads to a decrease of power of roughly the same magnitude as the decrease in FDR (between 0.25 and 0.55).

Spurious environmental variable Methods that use environmental variables to identify outliers assume that the chosen variables exert a selective pressure or are highly correlated to the one directly involved. One possible outcome in this situation is that the statistical tests identify a truly selected locus, but assign it to the wrong environmental variable. Although detecting a locus under selection is desirable, one does not want to link it to a spurious environmental variable. We call this error rate "spurious power" and define it as the proportion of truly selected loci considered as positive using a spurious, unrelated environmental variable. Fig. 6 shows that for HsIMM scenarios, the linear regression and BayEnv methods do not differ much in their "spurious power" (here, we only focus on the polygenic selection case). However, LFMM has a very low spurious power. For the IMM scenario, BayEnv is the most prone to erroneous choice of selective variable. By contrast, the linear regression is the most prone to error for the SS scenario.

Note that, in principle, the spurious power should be equal to the overall false positive rate (FPR), because we expect no association between the spurious environmental variable and the selected loci. This is roughly the case for all methods, except for LFMM in the scenarios IMM and SS. It tends to detect (false) association for selected loci more often than for non selected loci (see Fig. VIII in SI, note that the scale on these graphs are totally different from Fig. 6, since the methods differ in their False Positive Rate).

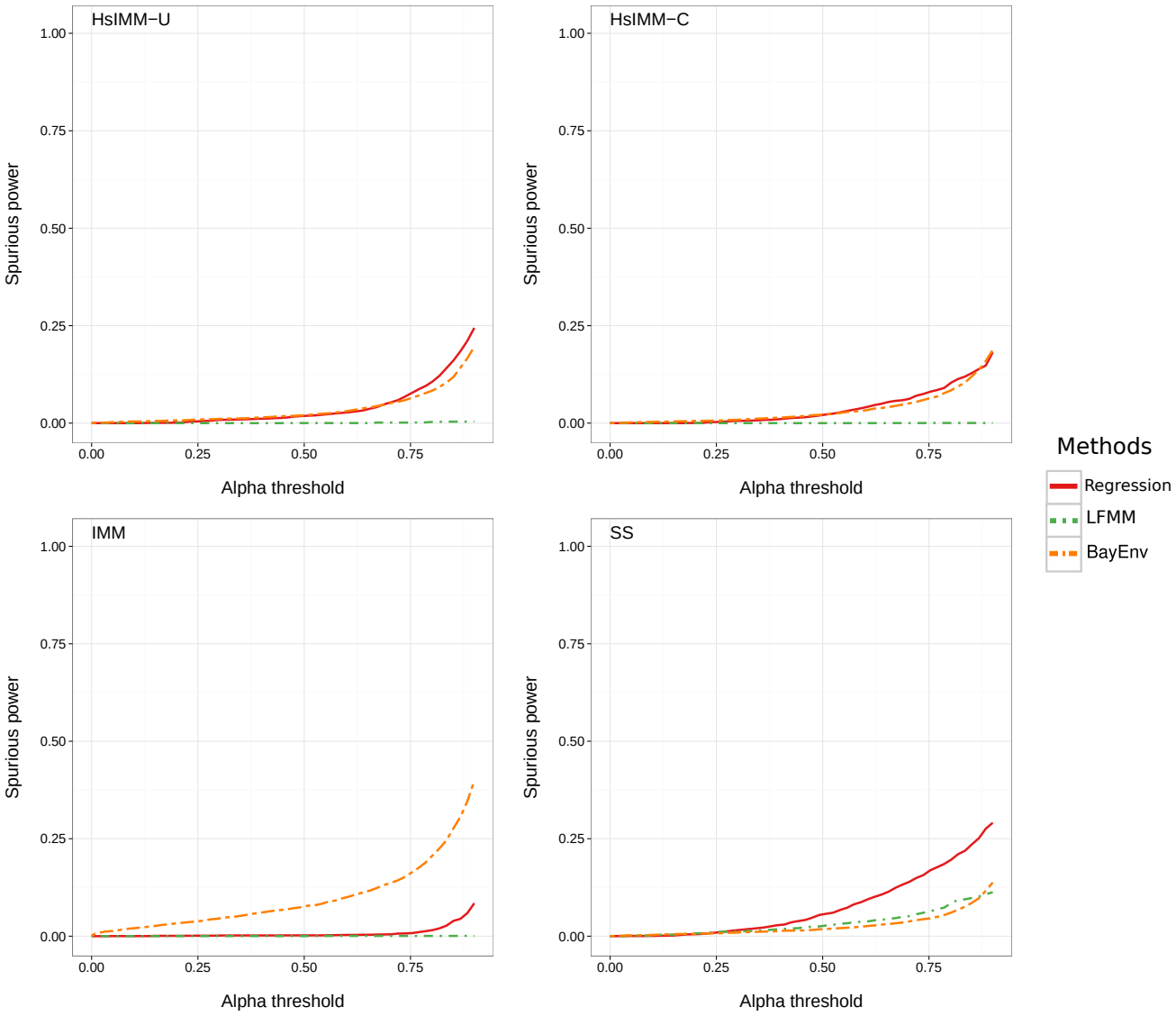

Figure 6: Spurious "power" (i.e. power to detect selected loci using an independent spurious variable) against significance threshold for the null environmental variable. Line are for linear regression (plain red line), Latent Factor Mixed Model (LFMM, green dot-dashed line), and BayEnv (orange two-dashed line). Light lines are for individual genotype data specification for the linear regression (light red) and LFMM (light green).

Discussion

Performances of the methods against difficult scenarios This study aimed at assessing the performance of recent and/or popular genome scan methods, in terms of power and error rate, when applied to difficult scenarios. The relative ranking of the methods, for the polygenic case, is summarised in Table 2. Note that the relationship in ranking between the FDR and the FPR is strong. Therefore, the methods have an inflated FDR mainly because of too many false positives, not because of too few true positives.

The most important challenge to the performance of all methods is the polygenic selection process. Obviously,

one would expect an overall decrease in power for all methods when using a polygenic selection model compared to a monogenic one, something that was actually observed. This decrease in power went hand in hand with an overall decrease of FDR (mostly due to the increased number of selected loci, see Eq. 7 in SI). However, the impact on performance differed among methods leading to a radical change in their ranking in terms of power/FDR. While all methods performed roughly equally in the monogenic scenarios –especially regarding power, for polygenic scenarios we observed large differences. First, the regression method became one of the most powerful but also most error prone methods. Second, the relative ranking between Bayescan, BayEnv and LFMM was changed, both in terms of power and error rate.

The second most important challenge was a strongly hierarchical spatial structure. This is evident when comparing the results for scenarios HsIMM-C and SS, both of which consider selection along an environmental gradient: the HsIMM-C scenario led to lower power for all methods. Note that the FDR for BayEnv was also inflated in the HsIMM-C scenario whereas it was almost perfect in the SS scenario. Apart from this overall changes in behaviour, the ranking of the method was conserved between the two spatial scenarios (although LFMM and the linear regression tend to be alike under the SS scenario).

The last challenge under study was the correlation between the environmental variable underlying the selective pressure and demographic history. The effects of this process can be visualised by comparing scenarios HsIMM-U and HsIMM-C, which only differ in this particular aspect. Overall, we see that a correlation between environment and demography led to low power for all methods, and higher FDR for Bayescan and BayEnv, which became even more prone to error than LFMM. The ranking, in terms of power, of the methods was conserved between the two kind of scenarios.

Another source of error to be considered in the case of association methods (e.g. the regression, LFMM and BayEnv) is that of associating the selected loci with a non selective (spurious) environmental variable. In this case, BayEnv and the linear regression methods yielded a stronger "spurious power" than LFMM. Also LFMM tended to associate the spurious variable with the selected loci more often than with the neutral loci.

We finally investigated the influence of the data specification (population allele frequencies or individual genotypes) for the linear regression and LFMM methods. The population allele frequencies data specification allowed for better performance in terms of error rate and most of the time in terms of power, at least under our simulated scenarios. This can be due to the fact that using genotypic data involved a larger sample size, which led to a higher rate of null model rejection due to slight violations of its underlying neutral hypotheses (higher power, but higher error rate). Note that, for polygenic selection, LFMM was less sensitive to the data specification. More puzzling, the genotypes specification sometimes led to a lower power.

Characteristics of the methods and comparison to previous studies Overall, we see that methods using an environmental variable have generally more power than genome-only based methods. Notably, Bayescan

		Regression	BayEnv	LFMM	Bayescan
HsIMM-U	FDR	⋆⋆	⋆⋆	⋆⋆⋆⋆	⋆
	FPR	⋆	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆
	Power	⋆⋆⋆⋆	⋆	⋆⋆⋆	⋆⋆
HsIMM-C	FDR	⋆	⋆⋆⋆⋆	⋆⋆	⋆⋆
	FPR	⋆	⋆⋆⋆⋆	⋆⋆	⋆⋆⋆
	Power	⋆⋆⋆⋆	⋆	⋆⋆⋆	⋆⋆
IMM	FDR	⋆⋆⋆	⋆⋆	⋆⋆⋆⋆	⋆
	FPR	⋆⋆⋆	⋆	⋆⋆⋆⋆	⋆⋆
	Power	⋆⋆⋆	⋆⋆⋆⋆	⋆	⋆⋆
SS	FDR	⋆	⋆⋆⋆⋆	⋆	⋆⋆⋆
	FPR	⋆	⋆⋆⋆⋆	⋆	⋆⋆⋆
	Power	⋆⋆⋆	⋆	⋆⋆⋆⋆	⋆⋆

Table 2: Summary of the properties of each genome scan methods, under the different scenarios focusing on the polygenic case. FDR: False Discovery Rate; FPR: False Positive Rate. Methods are ranked from the best (⋆⋆⋆⋆) to the worst (⋆). All properties are compared against the α thresholds below 0.05. When the ranking of the method was ambivalent, they were both assigned the same rank.

was always less powerful than at least one of the other methods. This is expected, since the method is not taking advantage of as much information as the others. One has to note, however, that sometimes it may not be possible to identify the environmental variable that should be considered, in which case a "blind" genome scan method must be used. Although Bayescan has been shown to perform quite well under Island and Stepping Stone scenarios (Foll and Gaggiotti, 2008; Narum and Hess, 2011; Vilas *et al.*, 2012; De Mita *et al.*, 2013), it did not perform very well under our Isolation with Migration (IMM) model and polygenic selection. This is a potentially common scenario so the results of our study differ from those of previous ones in that they suggest caution when using F_{ST}-based genome scans. Note, however, that the low power under the IMM scenario was only severe for the polygenic case.

Regarding LFMM and BayEnv, the two methods have much in common: both approaches employ mixed models in which environmental variables are introduced as fixed effects whereas population structure is introduced using unobserved variables or hidden factors. Yet, there are two main differences between the two methods. First, whereas BayEnv is a two-step procedure, estimating first the covariance structure of the population allele frequencies, and only then testing for association with an environmental variable, LFMM uses hidden factors to capture the part of genetic variation that cannot be explained by the set of measured environmental variables, all at once. This variation could include unknown demographic history, IBD patterns or environmental gradients not accounted for in the study. Second, the PCA-related nature of LFMM would *a priori* allow the method to take into account more complex scenarios. In particular, BayEnv has already been shown to perform poorly when confronted to hierarchical structure, and perform quite well in an island model (De Mita *et al.*, 2013). In our study, on the other hand, we also included LFMM and observed that this method over-corrected under the low-structure IMM scenario, leading to a very low FDR, but also a lower power.

Since the regression is not correcting for any population structure, we would expect it to yield more false positives, most likely accompanied by higher power. The regression is indeed the most error-prone method for

all scenarios, except for the IMM one (which is the least structured scenario). Note that, when increasing the number of selected loci (i.e. from the monogenic to the polygenic case), the compromise between false positives and power gets better for the regression model (see Fig. III and IV). This could be caused by the fact that the regression is a more sensitive method, which is less 'reluctant' to identify loci as selected. Thus, when many loci are selected, each with a small effect, we can expect this method to yield better power.

The results of this study differ from the previous ones (e.g. Pérez-Figueroa *et al.*, 2010; Narum and Hess, 2011; Vilas *et al.*, 2012; De Mita *et al.*, 2013) in several aspects. First, we used the same metric (the *q-value*) for all methods, which allows for a fair comparison. Second, while other studies investigated polygenic selection (Narum and Hess, 2011; Vilas *et al.*, 2012), they only considered up to 10 loci, and only investigated F_{ST}-based methods. Third, we used more complex models with strong hierarchical structure.

General issues and properties of genome-scan methods The results about polygenic selection tell us that assessing methods for monogenic scenarios only is not sufficient, especially because we expect the polygenic case to be the norm rather than the exception *in natura* (Pritchard and Di Rienzo, 2010). Of course, we have assumed a model of small locus effects, which could be one of the most difficult for genome scan methods. All methods may perform better under an L-shaped distribution of locus effects (see an example in Kulwal *et al.*, 2003), where a few loci have strong effects among numerous small effect loci. Yet, although there is evidence for the L-shaped architecture in the context of local adaptation (Yeaman and Whitlock, 2011), there is also evidence that some phenotypic traits are under the control of many small-effect loci (reviewed in Stranger *et al.*, 2011; Rockman, 2012).

Another important issue concerns methods that can consider both population- and individual-level data. In principle one expect that individual based data (genotypes) should lead to better performance, however, this is not necessarily the case. The type of data used has a large effect on the rate of false positives and consequently the FDR. We here illustrated this fact using LFMM and the linear regression models. Although we did not test it for BayEnv because the current implementation does not allow it, the results should be similar. This result is due to the simple fact that using the individual genotypes instead of allele frequencies (by frequencies here, we mean allele count data) increases the number of observations. This has the desirable property of increasing the power, but also leads to the undesirable increase in number of false positives, because the null models are essentially false. Indeed no model is a perfect description of the data; there will always be a discrepancy with the underlying processes that lead to the data (because of non linearity of effects, small differences between the potentially assumed and real demographic history, non-uniform mutation rates, etc.) and increasing the number of observations lead to the rejection of the null model for most loci instead of only the outlier ones (c.f. Raftery, 1995). Using population frequencies instead of genotypes is then a more conservative method. Yet it is not always possible to use frequencies, because of non homogeneous sample sizes, pooled sampling or the use of dominant data (e.g. AFLP). In those cases, one has to be aware that the statistical methods are not that

robust to departures from the underlying model, and the more observation points there are, the higher is the overall false positive rate (Raftery, 1995). Note that this is true for the total number of sampled individuals and the way they are implemented in the models, but not for the number of loci, which does not *a priori* increase the false positive rate.

Finally high FDR, especially in the case of monogenic selection, corresponds to an acceptable (though still inflated) false positive rate (FPR). For example, a FDR of 75% for the monogenic case corresponds to a FPR of 6.10^{-4} (see Eq. 4 in SI). For the polygenic case though, and assuming a power of 20%, it will correspond to a FPR of 6.10^{-3}. The fact that the methods tend to disagree might seem like a drawback, but it is in fact advantageous, because they tend to agree more on true positives than on false positives. Thus, by using all 4 methods together, we obtained FDRs between 0% and 40%, which are by far more acceptable.

Perspectives & Conclusion The results of our study pointed out two main directions in which statistical genomic studies should direct attention. First, we need more general and robust likelihood models that would be flexible enough to accommodate for strong departures from classical models. LFMM is an attempt in this direction, because its likelihood does not depend on a particular population model (Frichot *et al.*, 2012).

Second, we need methods better adapted to polygenic selection scenarios. The *q-value* framework allows to control for false discovery rate (Storey and Tibshirani, 2003; Storey *et al.*, 2004), which allow for test statistics that balance power and false positive rate. Another direction would be to develop a test that is suitable for polygenic selection. The difficulty in this case is that it would require to infer the genetic architecture of the trait(s) under selection, a very difficult task especially in absence of any phenotypic data.

Since polygenic selection and complex spatial population structures are likely to be quite common in the wild, it is important to tackle these two issues in order to develop reliable genome scan methods that can be applied to new NGS data from non-model species.

Acknowledgement We thank K. Csilléry and the anonymous reviewers for their useful insights. PdV was supported by a doctoral studentship from the French *Ministère de la Recherche et de l'Enseignement Supérieur*. OEG was supported by French ANR grant No 09 GENM 017 001 and by the Marine Alliance for Science and Technology for Scotland (MASTS). EF and OF were supported by a grant from *la Région Rhône-Alpes*. OF was further supported by Grenoble INP.

References

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

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings: Biological Sciences*, **263**(1377), 1619–1626.

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

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**(4), 1411 –1423.

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

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

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

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**(2), 977 –993.

Frichot E, Schoville S, Bouchard G, François O (2012) Landscape genomic tests for associations between loci and environmental gradients. *arXiv*, **1205.3347**.

Hermisson J (2009) Who believes in whole-genome scans for selection? *Heredity*, **103**(4), 283–284.

Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, Taberlet P (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**(18), 3955–3969.

Kulwal P, Roy J, Balyan H, Gupta P (2003) QTL mapping for growth and leaf characters in bread wheat. *Plant Science*, **164**(2), 267–277.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**(1), 175–195.

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.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

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

Pritchard JK, Di Rienzo A (2010) Adaptation – not by sweeps alone. *Nature Reviews Genetics*, **11**(10), 665–667.

Pérez-Figueroa A, García-Pereira MJ, Saura M, Rolán-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, **23**(10), 2267–2276.

Raftery AE (1995) Bayesian model selection in social research. *Sociological methodology*, **25**, 111–164.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**(1), 1–17.

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

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1), 187–205.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–9445.

Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait. *Genetics*, **187**(2), 367–383. PMID: 21115973.

Vilas A, Pérez-Figueroa A, Caballero A (2012) A simulation study on the performance of differentiation-based methods to detect selected loci using linked neutral markers. *Journal of Evolutionary Biology*, **25**(7), 1364–1376.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**(4), 1811–1823.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under Migration–Selection balance. *Evolution*, **65**(7), 1897–1911.