Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates

Mathieu Gautier∗,§,1

∗INRA, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez, France,
§Institut de Biologie Computationnelle, 95 rue de la Galera, 34095 Montpellier, France

ABSTRACT In population genomics studies, accounting for the neutral covariance structure across population allele frequencies is critical to improve the robustness of genome-wide scan approaches. Elaborating on the BAYENV model, this study investigates several modeling extensions i) to improve the estimation accuracy of the population covariance matrix and all the related measures; ii) to identify significantly overly differentiated SNPs based on a calibration procedure of the $X^2$ statistics; and iii) to consider alternative covariate models for analyses of association with population-specific covariables. In particular, the auxiliary variable model allows to deal with multiple testing issues and, providing the relative marker positions are available, to capture some Linkage Disequilibrium information. A comprehensive simulation study was carried out to evaluate the performances of these different models. Also, when compared in terms of power, robustness and computational efficiency, to five other state-of-the-art genome scan methods (BAYENV2, BAYSCENV, BAYSCAN, FLK and LFMM) the proposed approaches proved highly effective. For illustration purpose, genotyping data on 18 French cattle breeds were analyzed leading to the identification of thirteen strong signatures of selection. Among these, four (surrounding the KITLG, KIT, EDN3 and ALB genes) contained SNPs strongly associated with the piebald coloration pattern while a fifth (surrounding PLAG1) could be associated to morphological differences across the populations. Finally, analysis of Pool–Seq data from 12 populations of Littorina saxatilis living in two different ecotypes illustrates how the proposed framework might help addressing relevant ecological issue in non–model species. Overall, the proposed methods define a robust Bayesian framework to characterize adaptive genetic differentiation across populations. The BAYPASS program implementing the different models is available at http://www1.montpellier.inra.fr/CBGP/software/baypass/.

KEYWORDS Genome Scan; Bayesian statistics; Association Studies; Linkage Disequilibrium; Pool–Seq

Contrasting patterns of local genetic variation over the whole genome represents a valuable strategy to identify loci underlying the response to adaptive constraints (Cavalli-Sforza 1966). As further noted by Lewontin and Krakauer (1973): “while natural selection will operate differently for each locus and each allele at a locus, the effect of breeding structure is uniform over all loci and all alleles”. Hence, genome scan approaches to detect footprints of selection aim at discriminating among the global effect of the demographic evolutionary forces (e.g., gene flow, inbreeding and genetic drift) from the local effect of selection (Vitalis et al. 2001; Balding and Nichols 1995). In practice, applications of these methods have long been hindered by technical difficulties in assessing patterns of genetic variation on a whole genome scale. However, the advent of next-generation sequencing and genotyping molecular technologies now allows to provide a detailed picture of the structuring of genetic variation across populations in both model and non–model species (Davey et al. 2011). As a result, in the population genomics era, a wide range of approaches have been developed and applied to detect selective sweeps using population data (see Vitti et al. 2013; Oleksyk et al. 2010, for reviews). Among these, population differentiation ($F_{ST}$) based methods still remain among the most popular,
particularly in non-model species since they do not require accurate genomic resources (e.g., physical or linkage maps) and experimental designs with only a few tens of genotyped individuals per population are generally informative enough. Also, $F_{ST}$-based methods are well suited to the analysis of data from Pool–Seq experiments that consist in sequencing pools of individual DNAs (Schlötterer et al. 2014) and provide cost-effective alternatives to facilitate and even improve allele frequency estimation at genome-wide markers (Gautier et al. 2013).

In practice, assuming the vast majority of the genotyped markers behave neutrally, overly differentiated loci that are presumably subjected to selection might simply be identified from the extreme tail of the empirical distribution of the locus-specific $F_{ST}$ (Akey et al. 2002; Weir et al. 2005; Flori et al. 2009). Even if such a model-free strategy does not rely on any arbitrary assumptions about the (unknown) demographic history of the sampled populations, it prevents from controlling for false positive (and negative) signals. Conversely, model-based approaches have also been developed and are basically conceived as locus-specific tests of departure from expectation under neutral demographic models (e.g., Gautier et al. 2010a). These include, for instance, demographic models under pure-drift (Gautier et al. 2010a; Nicholson et al. 2002) and at migration-drift equilibrium without (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler et al. 2008; Guo et al. 2009) or with selection (Vitalis et al. 2014). Although robust, to some extent, to more complex history (Beaumont and Nichols 1996; Beaumont 2005), these methods remain limited by the oversimplification of the underlying demographic models. In particular, hierarchically structured population history, as produced under tree-shaped phylogenies, have been shown to increase false positive rates (Excoffier et al. 2009).

To cope with these issues, two kinds of modeling extensions have recently been explored. They either rely on hierarchical island models, thus requiring a priori definition of the sampled population relationships (Foll et al. 2014; Gompert et al. 2010), or consist in estimating the correlation structure of allele frequencies across the populations that originates from their shared history (Bonhomme et al. 2010; Coop et al. 2010; Günther and Coop 2013).

Whatever the method used, the main limitation of the indirect genome scan approaches ultimately resides in the biological interpretation of the footprints of selection identified, i.e., to which adaptive constraints the outlier loci are responding. In species with functionally annotated reference genomes, the characterization of co-functional relationships among the genes localized within regions under selection might help gaining insights into the underlying driving physiological pathways (e.g., Flori et al. 2009). Although, following a “reverse ecology” approach (Li et al. 2008), they may further lead to the definition of candidate adaptive traits for validation studies, such interpretations remain prone to misleading storytelling issues (Pavlidis et al. 2012). Alternatively, prior knowledge about some characteristics discriminating the populations under study could provide valuable insights. Focusing on environmental gradients, several approaches have recently been proposed to evaluate association of ecological variables with marker genetic differentiation by extending $F_{ST}$-based models (Coop et al. 2010; de Villémureuil and Gaggiotti 2015; Frichot et al. 2013; Günther and Coop 2013; Guillot et al. 2014; Hancock et al. 2011, 2008; Joost et al. 2007; Poncet et al. 2010). The rationale is that environmental variables distinguishing the differentiated populations should be associated with allele frequencies differences at loci subjected to the selective constraints they impose (Coop et al. 2010). In principle, such population-based association studies may also be more broadly relevant to any quantitative or categorical population-specific covariable. More generally, as for the covariable-free genome–scan approaches, accounting for the neutral correlation of allele frequencies across populations is critical for these methods (de Villémureuil et al. 2014; De Mita et al. 2013).

Overall, the Bayesian hierarchical model proposed by Coop et al. (2010) and implemented in the BAYENV2 software represents a flexible framework to address these issues. It indeed allows to both identify outlier loci (Günther and Coop 2013) and to further annotate the resulting footprints of selection by quantifying their association with population-specific covariables (if available). A key parameter of the model is the (scaled) population covariance matrix across population allele frequencies. Although this matrix might be viewed as purely instrumental, it explicitly incorporates their neutral correlation structure and is in turn highly informative for demographic inference purposes (Lipson et al. 2013; Pickrell and Pritchard 2012). Elaborating on the BAYENV model (Coop et al. 2010; Günther and Coop 2013), the purpose of this paper is threefold. First, we introduce modeling modifications and extensions to improve the estimation accuracy of the population covariance matrix and the different related measures. Second, we propose a posterior checking procedure to identify markers subjected to adaptive differentiation based on a calibration of the XIX statistics (Günther and Coop 2013). Third, we investigate alternative modeling strategies and decision criteria to perform association studies with population-specific covariables. In particular, we introduce a model with a binary auxiliary variable to classify each locus as associated or not. Through the prior distribution on this latter variable, the approach deals with the problem of multiple testing (e.g. Riebler et al. 2008). In addition, providing information about marker positions is available, this modeling also allows to account for Linkage Disequilibrium (LD) between markers via an Ising prior. As a by–product of this study, a user–friendly and freely available program, named BAYPASS, was developed to implement inferences under the different models. To evaluate the accuracy of the methods, we further carry out comprehensive simulation studies. In addition, two real data sets were analyzed in more detail to illustrate the range of application of the methods. The first consists of 453 individuals from 18 French cattle breeds genotyped at 42,056 SNPs (Gautier et al. 2010b) and the second of a Pool–Seq data on 12 Littorina saxatilis populations from three distinct geographical regions and living in two different ecotypes (Westram et al. 2014).

Models

In the following we describe the different Bayesian hierarchical models considered in this study and implemented in the BAYPASS program. Consider a sample made of $J$ populations (sharing a common history) with a label, $j$, which varies from 1 to $J$. The data consist of $I$ SNP loci, which are biallelic markers with reference allele arbitrarily defined (e.g., by randomly drawing the ancestral or the derived state). Let $n_{ij}$ be the total number of genes sampled at the $i^{th}$ locus ($1 \leq i \leq I$) in the $j^{th}$ population ($1 \leq j \leq J$), that is, twice the number of genotyped individuals in a diploid population. Let $y_{ij}$ be the count of the reference allele at the $i^{th}$ locus in the $j^{th}$ sampled population. When considering allele count data, the $y_{ij}$’s (and the $n_{ij}$’s) are the observations while for Pool–Seq data, read count are observed instead. In this case, the $n_{ij}$’s correspond for all
the markers within a given pool to its haploid sample size \( n_i \)
(i.e., twice the number of pooled individuals for diploid species).
Let further \( c_{ij} \) be the (observed) total number of reads and \( r_{ij} \)
the (observed) number of reads with the reference allele. For
Pool-seq data, to integrate over the unobserved allele count, the
conditional distribution of the \( r_{ij} \) given \( c_{ij}, n_i \) and the (unknown)
yield is assumed binomial (Gautier et al. 2013; Günther and Coop
2013): \( r_{ij} | c_{ij}, n_i, y_{ij} \sim \text{Bin}(\frac{n_i}{n_j}, c_{ij}) \).

Assuming Hardy–Weinberg Equilibrium, the conditional dis-
tribution of \( y_{ij} \) given \( n_i, r_{ij} \) and the (unknown) allele frequency \( \pi_i \)
also assumed binomial:

\[
y_{ij} | n_{ij}, \pi_i \sim \text{Bin}(n_{ij}, \pi_i) \tag{1}
\]

Note that this corresponds to the first level (likelihood) of the hi-
archical model when dealing with allele count data and to the
second level (prior) for Pool-Seq data. As previously proposed
and discussed (Coop et al. 2010; Gautier et al. 2010a; Nicholson
et al. 2002), for each SNP \( i \) and population \( j \) an instrumental
variable \( a_{ij}^* \) taking value on the real line is further introduced
such that:

\[
a_{ij}^* = \min(1, \max(0, a_{ij}^*) \tag{2}
\]

where \( 1 \) is a all-one vector of length \( J \); the precision matrix \( \Lambda \)
is the inverse of the (scaled) covariance matrix \( \Omega = \Omega^{-1} \)
of the population allele frequencies; and \( \pi_i \) is the weighted
mean reference allele frequency that might be interpreted as the
ancestral population allele frequency (Coop et al. 2010; Pickrell
and Pritchard 2012). The \( \pi_i \) are assumed Beta distributed:

\[
\pi_i | \alpha_i, \beta_i \sim \text{Beta}(\alpha_i, \beta_i) \tag{3}
\]

In such models, the parameters \( \alpha_i \) and \( \beta_i \) are frequently
fixed. For instance in BAYENV2 (Coop et al. 2010), \( \alpha_i = \beta_i = 1 \) lead-
ing to a uniform prior on \( \pi_i \) over the (0,1) support. However,
these parameters may be easily estimated from the model by
specifying a prior distribution on the mean \( \mu_\rho = \frac{\alpha_i}{\alpha_i + \beta_i} \), and the
so-called ’sample size’ \( v_\rho = a_{ij}^* + b_{ij} \) (Kruschke 2014). Hence, a
uniform and an exponential prior distribution are respectively
considered for these two parameters:

\[
\mu_\rho = \frac{\alpha_i}{\alpha_i + \beta_i} \sim \text{Unif}(0,1) \tag{4}
\]

and

\[
v_\rho = a_{ij}^* + b_{ij} \sim \text{Exp}(1) \tag{5}
\]

Finally, a Wishart prior distribution is assumed for the the preci-
sion matrix \( \Lambda \):

\[
\Lambda | \rho \sim W_j \left( \frac{1}{\rho} I_j, \rho \right) \tag{6}
\]

i.e., \( \pi(\Lambda | \rho) = \left( \frac{\rho}{\pi(\Lambda)} \right)^{\frac{n}{2}} e^{-\frac{\rho}{2} \text{tr}(\Lambda)} \) (\( I_j \) being the identity
matrix of size \( J \)). For \( \rho \geq 1 \) this is strictly equivalent to the
parametrization introduced in Coop et al. (2010) who eventually
came to fix \( \rho = 1 \). Here, weaker informative priors are also
explored with \( 0 < \rho < 1 \) (e.g., Gelman et al. 2003, p. 581) leading
to so-called singular Wishart distributions. As will become ap-
parent, \( \rho = 1 \) appears as the best default choice. Note however
that inspection of the full conditional distribution of \( \Lambda \) (see File
S1) suggests the influence of the prior might become negligible
with increasing number of SNPs \( I \) and populations \( J \).

The standard covariate model (STD model)

The STD model represented in Figure 1B extends the core model
as Coop et al. (2010) proposed and allows to evaluate association
of SNP allele frequencies with a population–specific covariable
\( Z_j \). Note that \( Z_j \) is a (preferably scaled) vector of length \( J \)
containing for each population the measures of interest. Under the
STD model, the prior distribution of the vector \( a_{ij}^* \) is multivariate
Gaussian for each SNP \( i \):

\[
a_{ij}^* | \Lambda, \beta_i, \pi_i \sim N_j \left( \pi_i \beta_i + Z_j, \pi_i(1 - \pi_i) \Lambda^{-1} \right) \tag{7}
\]

The prior distribution for the correlation coefficients \( \beta_i \) is as-
sumed uniform:

\[
\beta_i | \text{Unif}(\beta_{\text{min}}, \beta_{\text{max}}) \tag{8}
\]

Unless stated otherwise, \( \beta_{\text{min}} = -0.3 \) and \( \beta_{\text{max}} = 0.3 \) instead of
\( \beta_{\text{min}} = -0.1 \) and \( \beta_{\text{max}} = 0.1 \) as in Coop et al. (2010).

The covariate model with auxiliary variable (AUX model)

The AUX model represented in Figure 1C is an extension of the
STD model that consists in attaching to each locus regression
coefficient \( \beta_i \), a Bayesian (binary) auxiliary variable \( \delta_i \). In
a similar population genetics context, this modeling was also
proposed by Riebler et al. (2008) to identify markers subjected
to selection in a genome-wide scan for adaptive differentiation
(under a \( \mathcal{F} \)-model). In the AUX model, the auxiliary variable
actually indicates whether a specific SNP \( i \) can be regarded as
associated with the covariable \( Z_j \) (\( \delta_i = 1 \) or not (\( \delta_i = 0 \)). As
a consequence, the posterior mean of \( \delta_i \) may directly be inter-
preted as a posterior probability of association of the SNP \( i \) with
the covariable, from which a Bayes Factor (BF) is straightforward
to derive (Gautier et al. 2009). Under the AUX model, the prior
distribution of the vector \( a_{ij}^* \) is multivariate Gaussian for each
SNP \( i \):

\[
a_{ij}^* | \Lambda, \beta_i, \delta_i, \pi_i \sim N_j \left( \pi_i \beta_i + \delta_i Z_j, \pi_i(1 - \pi_i) \Lambda^{-1} \right) \tag{9}
\]

Providing information about marker positions is available, the
\( \delta_i \) ‘s auxiliary variables also make it easy to introduce spatial
dependency among markers. In the context of high–throughput
genotyping data, SNPs associated to a given covariable might in-
deed cluster in the genome due to LD with the underlying (possibly
not genotyped) causal polymorphisms(s). To learn from such
positional information, the prior distribution of \( \delta = \{ \delta_i \}_{1 \ldots I} \),
the vector of SNP auxiliary variables, takes the general form of a
1D Ising model with a parametrization inspired from Dufretest et al. (2014):
\[
\pi (\delta \mid P, b_{\eta}) \propto P^{s_1}(1 - P)^{s_2}e^{b_{\eta}} \tag{10}
\]
where \( s_1 = \sum_{i=1}^{I} I_{b_i=1} \) (respectively \( s_2 = I - s_1 \)) are the number of SNPs associated (respectively not associated) with the covari-able, and \( \eta = \sum_{i,j} I_{b_i=b_j} \) is the number of pairs of consecutive markers (neighbors) that are in the same state at the auxiliary variable (i.e., \( \delta_i = \delta_{i+1} \)). The parameter \( P \) corresponds to the proportion of SNPs associated to the covariable and is assumed Beta distributed:
\[
P \sim \beta (a_P, b_P) \tag{11}
\]
Unless stated otherwise, \( a_P = 0.02 \) and \( b_P = 1.98 \). This amounts to assume that only a small fraction of the SNPs (\( \frac{a_P + b_P}{a_P} = 1\% \)) are associated to the covariable, but within a reasonably large range of possible values (e.g., \( P \mid P > 10\% \) = 2.8% a priori). Importantly, integrating over the uncertainty on the key parameter \( P \) allows to deal with multiple testing issues.

Finally, the parameter \( b_{\eta} \), called the inverse temperature in the Ising (and Potts) model literature, determines the level of spatial homogeneity of the auxiliary variables between neighbors. When \( b_{\eta} = 0 \), the relative marker position is ignored (no spatial dependency among markers). This is thus equivalent to assume a Bernoulli prior for the \( \delta_i \)'s: \( \delta_i \sim Ber (P) \) as in Riebler et al. (2008). Conversely, \( b_{\eta} > 0 \) leads to assume that the \( \delta_i \) with similar values tend to cluster in the genome (the higher the \( b_{\eta} \), the higher the level of spatial homogeneity). In practice, \( b_{\eta} = 1 \) is commonly used and values of \( b_{\eta} \leq 1 \) are recommended. Note that the overall parametrization of the Ising prior assumes no external field and no weight (as in the so-called compound Ising model) between the neighboring auxiliary variables. In other words, the information about the distances between SNPs is therefore not accounted for and only the relative position of markers is considered. Hence, marker spacing is assumed homogeneous.

### Material and Methods

#### MCMC sampler
To explore the different models and estimates, the full posterior distribution of the underlying parameters, a Metropolis–Hastings within Gibbs Markov Chain Monte Carlo (MCMC) algorithm was developed (see the File S1 for a detailed description) and implemented in a program called BAYPASS (for BAYesian Population ASSociation analysis). The software package containing the Fortran 90 source code, a detailed documentation and several example files is freely available for download at http://www1.montpellier.inra.fr/CGBP/software/baypass/. Unless otherwise stated, a MCMC chain first consists in 20 pilot runs of 1,000 iterations each allowing to adjust proposal distributions (for Metropolis and Metropolis–Hastings updates) with targeted acceptance rates lying between 0.2 and 0.4 to achieve good convergence properties (Gilks et al. 1996). Then MCMC chains are run for 25,000 iterations after a 5,000 iterations burn-in period. Samples are taken from the chain every 25 post-burn-in iterations to reduce autocorrelations using a so-called thinning procedure. To validate the BAYPASS sampler, an independent implementation of the core model was coded in the BUGS language and run in the OPENBUGS software (Thomas et al. 2009) as detailed in the File S2. Analyses of some (small) test data sets using both implementations gave consistent results (data not shown).

Finally, as a matter of comparison, in the analysis of prior sensitivity in \( \Omega \) estimation, the BAYENV2 ( Günther and Coop 2013 ) software was also used with default options except the total number of iterations that was set to 50,000.

#### Estimation and visualisation of \( \Omega \)
For BAYPASS analyses, point estimates of each elements of \( \Omega \) consisted of their corresponding posterior means computed over the sampled matrices. For BAYENV2 analyses, the first ten sampled matrices were discarded and only the 90 remaining ones were retained. As a matter of comparison, the frequentist estimate of \( \Omega \) as proposed by Bonhomme et al. (2010) and implemented in the FLK package was also considered. Briefly, the FLK relies on a neighbor–joining algorithm on the Reynolds pairwise population distances matrix to build a population tree from which the covariance matrix is deduced (after midpoint rooting of the tree).

For visualization purposes, a given \( \Omega \) estimate was transformed into a correlation matrix \( \tilde{\Omega} \) using the \texttt{cor2cor()} R function (R Core Team 2015). The graphical display of this correlation matrix was done with the \texttt{corrplot()} function from the R package corrplot (Wei 2013). In addition, hierarchical clustering of the underlying populations was performed using the \texttt{hclust()} R function considering \( 1 - \tilde{\rho}_{ij} \) as a dissimilarity measure between each pair of population \( i \) and \( j \). The resulting bifurcating tree was plotted with the \texttt{plot} and \texttt{phylo} functions from the R package ape (Paradis et al. 2004). Note that the latter representation reduces the correlation matrix into a block-diagonal matrix thus ignoring gene flow and admixture events.

#### Computation of the FMD metric to compare \( \Omega \) matrices
The metric proposed by Förstner and Moonen (2003) for covariance matrices and hereafter referred to as the FMD distance was used to compare the different estimates of \( \Omega \) and to assess estimation precision and robustness in the prior sensitivity analysis. Let \( \Omega_1 \) and \( \Omega_2 \) be two (symmetric positive definite) covariance matrices with rank \( f \), the FMD distance is defined as:
\[
\text{FMD} (\Omega_1, \Omega_2) = \sqrt{\sum_{j=0}^{f-1} \lambda_j^2 (\Omega_1, \Omega_2)} \tag{12}
\]
where \( \lambda_j (\Omega_1, \Omega_2) \) represent the \( j \)-th generalized eigenvalue of the matrices \( \Omega_1 \) and \( \Omega_2 \) that were all computed with the R package geigen (Hasselman 2015).

#### Computation and calibration of the \( \chi^2 \) statistic
Identification of SNPs subjected to adaptive differentiation relied on the \( \chi^2 \) differentiation measure introduced by Günther and Coop (2013). This statistic might be viewed as a SNP–specific relative field and no weight (as in the so-called compound Ising model) between the neighboring auxiliary variables. In other words, the information about the distances between SNPs is therefore not accounted for and only the relative position of markers is considered. Hence, marker spacing is assumed homogeneous.

The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.
estimated the posterior predictive distribution of this statistic under the null (core) model by analyzing pseudo-observed data sets (POD). PODs are produced by sampling new observations (either allele or read count data) from the core inference model with (hyper-)parameters $a_n$, $b_n$ and $\Lambda$ (the most distant nodes in the DAG of Figure 1) fixed to their respective posterior means obtained from the analysis of the original data. The sample characteristics are preserved by sampling randomly (with replacement) SNP vectors of $n_{ij}$’s (allele count data) or $c_{ij}$’s (for read count data) among the observed ones. For Pool–Seq data, haplotype sample sizes are set to the observed ones. The R (R Core Team 2015) function `simulate.bypass()` available in the BAYPASS software package was developed to carry out these simulations. The POD is further analyzed using the same MCMC parameters (number and length of pilot runs, burn-in, chain length, etc.) as for the analysis of the original data set. The XtX values computed for each simulated locus are then combined to obtain an empirical distribution. The quantiles of this empirical distribution are computed and are used to calibrate the XtX observed for each locus in the original data: e.g., the 99% quantile of the XtX distribution from the POD analysis provides a 1% threshold XtX value, which is then used as a decision criterion to discriminate between selection and neutrality. Note that this calibration procedure is similar to the one used in Vitalis et al. (2014) for the calibration of their SNP KLD.

Population Association tests and decision rules

Association of SNPs with population–specific covariates is assessed using Bayes Factors (BF) or what may be called “empirical Bayesian P–values” (eBP). Briefly, for a given SNP, BF compares models with and without association while eBP is aimed at measuring to which extent the posterior distribution of the regression coefficient $\beta_i$ excludes 0. Note that eBP’s are not expected to display the same frequentist properties as classical P–values.

Two different approaches were considered to compute BF’s. The first estimate (hereafter referred to as BF$_{S}$) relies on the Importance Sampling algorithm proposed by Coop et al. (2010) and uses MCMC samples obtained under the core model (see File S3 for a detailed description). The second estimate (hereafter referred to as BF$_{mc}$) is obtained from the posterior mean $\mu(\delta_i)$ of the auxiliary variable $\delta_i$ under the AUX model:

$$BF_{mc} = \frac{\mu(\delta_i)}{1 - \mu(\delta_i)} b_p$$

(14)

where $\mu(\delta_i)$ is the (estimated) posterior odds that the locus $i$ is associated to the covariable and $b_p$ is the corresponding prior odds (Gautier et al. 2009). Hereby, BF$_{mc}$ is only derived for the AUX model with $b_{mc} = 0$ (the prior odds being challenging to compute when $b_{mc} \neq 0$). In practice, to account for the finite MCMC sampled values $T$, $\mu(\delta_i)$ is set equal to $1 - 0.5 \Phi(0.5/T)$ (respectively $0.5 \Phi(0.5/T)$) when the posterior mean of the $\delta_i$ is equal to 1 (respectively 0). Note that, through the prior on $P$, the computation of BF$_{mc}$ explicitly accounts for multiple testing issues. BF’s are generally expressed in deciban units (dB) (via the transformation $10\log_{10}(BF)$). The Jeffreys’ rule (Jeffreys 1961) provide a useful decision criterion to quantify the strength of evidence (here in favor of association of the SNP with the covariable) using the following dB unit scale: "strong evidence" when $10 \times BF > 15$, "very strong evidence" when $15 \times BF > 20$ and "decisive evidence" when $BF > 20$.

For the computation of eBP’s, the posterior distribution of each SNP was approximated as a Gaussian distribution: $N(\mu(\beta_i), \sigma^2(\beta_i))$ where $\mu(\beta_i)$ and $\sigma(\beta_i)$ are the estimated posterior mean and standard deviation of the corresponding $\beta_i$. The eBP’s are further defined as:

$$eBP = -\log_{10} \left( 1 - 2 \times 0.5 \Phi \left( \frac{\mu(\beta_i)}{\sigma(\beta_i)} \right) \right)$$

(15)

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Roughly speaking, a value of $\beta$ might be viewed as “significantly” different from 0 at a level of $10^{-eBP}$. Two different approaches were considered to estimate the moments of the posterior distribution of the $\beta_i$’s. The first, detailed in the File S3, relies on an Importance Sampling algorithm similar to the one mentioned above and thus uses MCMC samples obtained under the core model. The resulting eBP’s estimates are hereafter referred to as $eBP_{mc}$. The second approach relies on posterior samples of the MCMC samples obtained under the STD model. The resulting eBP’s estimates are hereafter referred to as $eBP_{mc}$. Note finally, that for estimating BF$_{mc}$ (under the AUX model) and $eBP_{mc}$ (under the STD model), the value of $\Lambda$ was fixed to its posterior mean as obtained from an initial analysis carried out under the core model.

Simulation study

Simulation under the inference model. Simulation under the core or the STD inference models defined above (Figure 1) were carried out using the function `simulate.bypass()` available in the BAYPASS software package. Briefly, a simulated data set is specified by the $\Omega$ matrix, the parameters of the Beta distribution for the ancestral allele frequencies ($a_n$ and $b_n$) and the sample sizes. As a matter of expedience, ancestral allele frequencies below 0.01 (respectively above 0.99) were set equal to 0.01 (respectively 0.99) and markers that were not polymorphic in the resulting simulated data set were discarded from further analyses. For the generation of PODs (see above), the $n_{ij}$’s (or the $c_{ij}$’s for Pool–Seq data) were sampled (with replacement) from the observed ones and for the power analyses, these were fixed to $n_{ij} = 50$ for all the populations. To simulate under the STD model, the simulated $\beta_i$’s (SNP regression coefficients) were specified and the population covariable vector $Z$ was simply taken from the standard normal cumulative distribution function such that $z_j = \Phi \left( 0.01 + 0.98^{-1/j} \right)$ for the $j$th population (out of the J ones).

Individual–based simulations. Individual–based forward–in–time simulations under more realistic scenarios were carried out under the SimuPOP environment (Peng and Kimmel 2005) as described in de Villemereuil et al. (2014). Briefly, three scenarios corresponding to i) an highly structured isolation with migration model (HsIMM–C), ii) an isolation with migration model (IMM); and iii) a stepping stone scenario (SS) were investigated. For each scenario, one data set consisted of 320 individuals belonging to 16 different populations that were genotyped for 5,000 SNPs regularly spread along 10 chromosomes of 1 Morgan length. Polygenic selection acting on an environmental gradient (see de Villemereuil et al. 2014, for more details) was included in the simulation model by choosing 50 randomly distributed SNPs (among the 5,000 simulated ones) and affecting them a selection coefficient $s_i$ calculated as a logistic transformation of the corresponding population–specific environmental variable

```r
# Simulation R code

# Load libraries
library(simPop)

# Define parameters
populations <- 16
snps <- 5000
chromosomes <- 10
length <- 1

# Create populations
population <- simPopulation(populations, snps, chromosomes, length)

# Add selection
selection <- 0.01 + 0.98^{-1/j}

# Add environmental gradient
environment <- simEnviron(populations, snps, chromosomes, length, selection)

# Simulate data
data <- simulateData(population, environment, length)

# Analyze data
results <- analyzeData(data)

# Output results
print(results)
```

The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. All rights reserved. No reuse allowed without permission.
was also used. It is conceived as an extension of BAYPASS (see above), the HslMM-C, IMM and SS individual-based simulated data sets were analyzed with five other popular or recently developed genome scan approaches. These first include BAYESCAN (Foll and Gaggiotti 2008) which is a Bayesian covariate-free approach that identifies overly differentiated markers (with respect to expectation under a migration–drift equilibrium demographic model) via a logistic regression of the population-by-locus FSST on a locus-specific and population-specific effect. The decision criterion was based on a Bayes Factor that quantify the support in favor of a non-null locus effect. Second, the recently developed BAYSCENV (de Villemereuil and Gaggiotti 2015) model was also used. It is conceived as an extension of BAYESCAN incorporating environmental information by including a locus-specific regression coefficient parameter (noted $g$) in the above mentioned logistic regression. The decision criterion to assess association with the covariate was based on the estimated posterior probability of $g$ being non null. In practice, to limit computation burden for both BAYESCAN (version 2.1) and BAYSCENV, default MCMC parameter options of the programs were chosen except for the length of the pilot runs (set to 1,000), the length of the burn-in period (set to 10,000) and the number of sampled values (set 2500). A third and covariate-free approach consisted in computing the FLK statistics (which might be viewed as the frequentist counterpart of the $X^2$ described above) as described in Bonhomme et al. (2010). The fourth considered method relied on Latent Factor Mixed Models as implemented in the LFMM (version 1.4) software (Frichot et al. 2013) to detect association of allele frequencies differences with population-specific covariables while accounting for population structure via the so-called latent factors. Following de Villemereuil et al. (2014) that analyzed the same data sets, the prior number of latent factors requested by the program was set to $K = 15$. Note also that LFMM analyses were run on individual genotyping data rather than population allele frequencies, which were previously shown to display better performances (de Villemereuil et al. 2014). For each data set, the decision criterion to assess association of the SNP with the environmental covariable relied on a $P$-value that was computed based either on a single analysis (denoted as LFMM) or derived after combining Z-score from 10 independent analyses (denoted as LFMM–10rep) following the procedure described in the LFMM (version 1.4) manual. Finally, the data sets were also analyzed with BAYENV2 (Coop et al. 2010) following a two–steps procedure (as required by the program) that was similar to the one performed by de Villemereuil et al. (2014). For each data set, a first MCMC of 15,000 iterations was run under default parameter settings and the latest sampled covariance matrix was used as an estimate of $\Omega$. For each SNP in turn, an MCMC of 30,000 iterations was further run to estimate the corresponding $X^2$ and BF based on this latter matrix. To facilitate automation of the whole procedure, a custom shell script was developed.

Each analysis was run on a single node of the same computer cluster to provide a fair comparison of computation times. To further compare the performances of the different models, the actual i) True Positive Rates (TPR) or power i.e. the proportion of true positives among the truly selected loci; ii) False Positive Rates (FPR) i.e. the proportion of false positives among the non selected loci; and iii) False Discovery Rates (FDR) i.e. the proportion of false positives among the significant loci were computed from the results of each data set with the different methods for various thresholds covering the range of values of the corresponding decision criterion. From these estimates, both standard Receiver Operating Curves (ROC) plotting TPR against FPR, and Precision-Recall (PR) curves plotting (1-FDR) against TPR could then be drawn.

Real Data sets

The HSA_{snp} data set. This data set is the same as in Coop et al. (2010) and was downloaded from the BAYENV2 software web-page (http://gcbias.org/bayenv/). It consists of genotypes at 2,333 SNPs for 927 individuals from 52 human population of the HGDP panel (Conrad et al. 2006).

The BTA_{snp} data set. This data set is a subset of the data from Gautier et al. (2010b) and consists of 453 individuals from 18 French cattle breeds (from 18 to 46 individuals per breed) genotyped for 42,046 autosomal SNPs displaying an overall MAF>0.01. As detailed in File S4, two breed-specific covariables were considered for association analyses. The first covariable corresponds to a synthetic morphology score (SMS) defined as the (scaled) first principal component of breed average weights and wither heights for both males and females (taken from the French BRG website: http://www.brg.prd.fr/). The second covariable is related to coat color and corresponds to the piebald coloration pattern of the different breeds that was coded as 1 for pied breed (e.g., Holstein breed) and −1 for breeds with a uniform coloration pattern (e.g., Tarine breed).

The LSA_{snp} data set. This data set was obtained from whole transcriptomes of pooled Littorina saxatilis (LSA) individuals belonging to 12 different populations originating (Westram et al. 2014). These populations originate from three distinct geographical regions (UK, the United Kingdom; SP, Spain and SW, Sweden) and lived in two different ecotypes corresponding to the so-called "wave" habitat (subjected to wave action) and "crab" habitat (i.e., subjected to crab predation). The mpileup file with the aligned RNA–seq reads from the 12 pools (three countries $\times$ two ecotypes $\times$ two replicates) onto the draft LSA genome assembly was downloaded from the Dryad Digital Repository doi:10.5061/dryad.21pf0 (Westram et al. 2014). The mpileup file was further processed using a custom awk script to perform SNP calling and derive read counts for each alternative base (after discarding bases with a BAQ quality score <25). A position was considered as variable if i) it had a coverage of more than 20 and less than 250 reads in each population; ii) only two different
bases were observed across all the five pools and; iii) the minor allele was represented by at least one read in two different pool samples. Note that tri-allelic positions for which the two most frequent alleles satisfied the above criteria and with the third allele represented by only one read were included in the analysis as bi-allelic SNPs (after filtering the third allele as a sequencing error). The final data set then consisted of allele counts for 53,387 SNPs. As a matter of expediency, the haploid sample size was set to 100 for all the populations because samples consisted of pools of ca. 40 females with their embryos (from tens to hundreds per female) (Westram et al. 2014). To carry out the population analysis of association with ecotype and identify loci subjected to parallel phenotypic divergence, the habitat is considered as a binary covariable respectively coded as 1 for the “wave” habitat and −1 for the “crab” habitat.

Results

**Performance of the core model for estimation of the scaled population covariance matrix $\Omega$**

The scaled covariance matrix $\Omega$ of population allele frequencies represents the key parameter of the models considered in this study. To illustrate how prior parametrization might influence estimation of $\Omega$, we first analyzed the BTA$_{\text{ASS}}$ (with J=18 French cattle populations) and the HSA$_{\text{ASS}}$ (with J=52 worldwide human populations) data sets using both BAYPASS (under the core model represented in Figure 1A with $\rho = 1$) and BAYENV2 (in which $\rho = J$ and $a_\pi = b_\pi = 1$ according to Coop et al. (2010)). Note that the sampled populations in these two data sets have similar characteristics in terms of the overall $F_{ST}$ ($F_{ST} = 9.84\%$ and $F_{ST} = 10.8\%$ for the cattle and human sampled populations, respectively). The resulting estimated $\Omega$ matrices are hereafter denoted as $\Omega_{\text{BTA}}^{\text{bvas}}$ and $\Omega_{\text{BTA}}^{\text{benv}}$ respectively for the cattle data set and are represented in Figure 2. Similarly, for the human data set, the resulting $\Omega_{\text{HSA}}^{\text{bvas}}$ and $\Omega_{\text{HSA}}^{\text{benv}}$ are represented in Figure S1. For both data sets, the comparisons of the two different estimates of $\Omega$ reveal clear differences that suggest in turn some sensitivity of the model to the prior assumption. Analyses under three other alternative BAYPASS model parametrizations (i) $\rho = 1$ and $a_\pi = b_\pi = 1$; ii) $\rho = J$ and; iii) $\rho = J$ and $a_\pi = b_\pi = 1$) confirmed this intuition (Figure S2). For the human data set, the FMD between the different estimates of $\Omega$ varied from 1.73 (BAYPASS with $\rho = 1$ vs BAYPASS with $\rho = 1$ and $a_\pi = b_\pi = 1$) to 31.1 (BAYPASS with $\rho = 1$ vs BAYPASS with $\rho = 52$). However, for the cattle data set that contains about 20 times as many SNPs for 3 times less populations, the four BAYPASS analyses gave consistent estimates (pairwise FMD always below 0.5) that clearly depart from the BAYENV2 one (pairwise FMD always above 14). Note also that BAYPASS estimates were in better agreement with the historical and geographic origins of the sampled breeds (see Figure 2 and Gautier et al. (2010b) for further details).

Overall these contrasting results call for a detailed analysis of the sensitivity of the model to prior specifications on both $\Omega$ ($\rho$ value) and the $\pi$’s. Beta distribution parameters ($a_\pi$ and $b_\pi$), but also to data complexity (number and heterozygosity of SNPs). To that end we first simulated under the core inference model (Figure 1A) data sets for four different scenarios labeled SpsH1, SpsH2, SpsB1 and SpsB2. In SpsH1 and SpsH2 (respectively SpsB1 and SpsB2), the population covariance matrix was set equal to $\Omega_{\text{HSA}}^{\text{bvas}}$ (respectively $\Omega_{\text{BTA}}^{\text{bvas}}$) and in SpsH1 and SpsB1 (respectively SpsH2 and SpsB2) the $\pi$’s were sampled from a Uniform distribution over (0,1) (respectively a Beta(0.2,0.2) distribution). Note that the two different $\pi$ distributions lead to quite different SNP frequency spectrum, the Uniform one approaching (ascertained) SNP chip data (i.e., good representation of SNPs with an overall intermediate MAF) while the Beta(0.2,0.2) one is more similar to that obtained in whole genome sequencing experiments with an over-representation of poorly informative SNPs (see, e.g., results obtained on the LSA$_{\text{ps}}$ Pool–Seq data below). To assess the influence of the number of genotyped SNPs, data sets consisting of 1,000, 5,000, 10,000 and 25,000 SNPs were simulated for each scenario. For each set of simulation parameters, ten independent replicate data sets were generated leading to a total of 160 simulated data sets (10 replicates × 4 scenarios × 4 SNP numbers) that were each analyzed with BAYENV2 (Coop et al. 2010) and four alternative BAYPASS model parameterizations (i) $\rho = 1$; ii) $\rho = 1$ and $a_\pi = b_\pi = 1$; iii) $\rho = J$ and; iv) $\rho = J$ and $a_\pi = b_\pi = 1$. As a matter of comparisons, the FLK frequentist estimate (Bonhomme et al. 2010) of the covariance matrices was also computed. FMD distances (averaged across replicates) of the resulting $\Omega$ estimates from their corresponding true matrices are represented in Figure 3. Note that for a given simulation parameter set, the FMD distances remained quite consistent (under a given model parametrization) across the ten replicates (Figure S3).

[Figure 3 about here.]

Except for the BAYENV2 and FLK analyses, the estimated matrices converged to the true ones as the number of SNPs (and thus the information) increase. In addition, as observed above for real data sets, the BAYENV2 estimates were always quite different from those obtained with BAYPASS parametrized under the same model assumptions ($\rho = \text{pop}$ and $a_\pi = b_\pi = 1$). It should also be noticed that reproducing the same simulation study by using the $\Omega_{\text{BTA}}^{\text{bvas}}$ and $\Omega_{\text{BTA}}^{\text{benv}}$ matrices in the four different scenarios lead to similar patterns (Figure S4). Reasons for this behavior of BAYENV2 (possibly the result of some minor implementation issues) were not investigated further and we hereafter only concentrated on results obtained with BAYPASS.

As expected, the optimal number of SNPs also depends on their heterozygosity. Hence, when the simulated $\pi$’s were sampled from a Beta(0.2,0.2) (Figure 3B and D) instead of a Unif(0,1) distribution, a higher number of SNPs was required (compare Figures 3B and A; and Figures 3D and C, respectively) to achieve the same accuracy. Likewise, all else being equal, the estimation precision was found always lower for the SpsH1 (and SpsH2) than SpsB1 (and SpsB2) scenarios. This shows that the optimal number of SNPs is an increasing function of the number of sampled populations. One might also expect that more SNPs are required when population differentiation is lower (although this was not formally tested here). Regarding the sensitivity of the models to the prior definition, the parametrization with $\rho = 1$ clearly outperformed the more informative one ($\rho = J$), most particularly for smaller number of SNPs and more complex data sets. Naturally, estimating the parameters $a_\pi$ and $b_\pi$ compared to setting them to $a_\pi = b_\pi = 1$ had almost no effect in the estimation precision of $\Omega$ for the SpsH1 and SpsB1 scenarios, their resulting posterior means being slightly larger than one ($\simeq 1.1$ due probably to the simulation SNP ascertainment scheme as described in Material and Methods). Interestingly however, a substantial gain in precision was obtained for the SpsH2 and

BAYPASS
SpS2 data sets (for which \( \tau_{j}^{\text{sim}} \sim \text{Beta}(0.2, 0.2) \)). Hence, for the SpS2 data sets (Figure 3D), the FMD curves reached a plateau with the \( \hat{a}_{X} = \beta_{X} = 1 \) parametrization (for both \( \rho = 1 \) and \( \rho = 18 \)) as the number of SNPs increase whereas precision kept improving when \( \hat{a}_{X} \) and \( \beta_{X} \) were estimated.

We finally investigated to which extent estimation of \( \hat{a}_{X} \) and \( \beta_{X} \) might improve robustness to SNP ascertainment. To that end, ten additional independent data sets of 100,000 SNPs were simulated under both the SpH1 and SpB1 scenarios. For each of the twenty resulting data sets, six subsamples were constituted by randomly sampling 25,000 SNPs with an overall MAF > 0.01, > 0.025, > 0.05, > 0.075 and > 0.10 respectively. The 120 resulting data sets (2 scenarios \( \times 10 \) replicates \( \times 6 \) MAF thresholds) were analyzed with BayPass (assuming \( \rho = 1 \)) by either estimating \( \hat{a}_{X} \) and \( \beta_{X} \) or setting \( \hat{a}_{X} = \beta_{X} = 1 \). Although the estimation precision of \( \Omega \) was found to decrease with increasing MAF thresholds (Figure S5), estimating \( \hat{a}_{X} \) and \( \beta_{X} \) allowed to clearly improve accuracy in these examples. Note however, that the effect of the ascertainment scheme remained limited, in particular for small MAF thresholds (MAF<0.05).

**Performance of the X\( \mathbf{X} \) statistics to detect overly differentiated SNPs.**

To evaluate the performance of the X\( \mathbf{X} \) statistics to identify SNPs subjected to selection, data sets were simulated under the STD inference model (Figure 1B), i.e., with a population–specific covariable. This simulation strategy was mainly adopted to compare covariable-free X\( \mathbf{X} \) based decision (scan for differentiation) with association analyses (based on covariate models) as described in the next section. Obviously, the X\( \mathbf{X} \) is a covariable-free statistic that is powerful to identify SNPs subjected to a broader kind of adaptive constraints, as elsewhere demonstrated (Bonhomme et al. 2010; Günther and Coop 2013). Hence, two different (demographic) scenarios, labeled SpaH and SpaB, were considered.

In the scenario SpaH (respectively SpaB), \( \Omega_{\text{sim}}^{\text{amp}} \) was set equal to \( \Omega_{\text{HSA}}^{\text{spas}} \) (respectively \( \Omega_{\text{BTA}}^{\text{spas}} \)), and the \( \tau_{i} \)'s were sampled from a Uniform distribution. For each scenario, 25,600 SNPs were simulated of which 25,000 are neutral SNPs (i.e., with a regression coefficient \( \beta_{i} = 0 \)) and 600 are SNPs associated with a normally distributed population–specific covariable (see Material and Methods) and with regression coefficients \( \beta_{i} = -0.2 \) (n=100), \( \beta_{i} = -0.1 \) (n=100), \( \beta_{i} = -0.05 \) (n=100), \( \beta_{i} = 0.05 \) (n=100), \( \beta_{i} = 0.1 \) (n=100), \( \beta_{i} = 0.2 \) (n=100). For each scenario, ten independent replicate data sets, each with a randomized population covariable vector, were generated. The resulting 20 simulated data sets (10 replicates \( \times 2 \) scenarios) were then analyzed with four alternative BayPass model parameterizations corresponding to i) the core model (Figure 1A) with \( \rho = 1 \); ii) the core model by setting \( \Omega = \Omega_{\text{amp}}^{\text{sim}} \); iii) the STD model (Figure 1B) by setting \( \Omega = \Omega_{\text{amp}}^{\text{sim}} \) and; iv) the default AUX model (Figure 1C) i.e. with \( b_{h} = 0 \) and \( \Omega = \Omega_{\text{amp}}^{\text{sim}} \).

As expected, under the core model, the higher \( |\beta_{i}| \), the higher the estimated X\( \mathbf{X} \) on average (Figure S6). As a matter of expediency, for power comparisons, 1% POD threshold were further defined for each analysis using the X\( \mathbf{X} \) distribution obtained for SNPs with simulated \( \beta_{i} = 0 \). Note that the resulting thresholds were very similar to those obtained using independent data sets (e.g., SpH1 and SpB1) that lead to FPR close to 1%. As shown in Table 1, the power was optimal (> 99.9%) for strongly associated SNPs (\( |\beta_{i}| = 0.2 \)) in both scenarios but remained small (\(< 10\% \)) for weakly associated SNPs. In addition, power was always higher with the SpaH than with the SpaB data probably due to a more informative design (three times as many populations). Likewise, estimating \( \Omega \) (i.e., including information from the associated SNPs) slightly affected the performance of the X\( \mathbf{X} \)-based criterion when compared to setting \( \Omega = \Omega_{\text{amp}}^{\text{sim}} \) (see Table 1 and also the ROC analyses in Figure S7). Yet the resulting estimated matrices \( \Omega \) were close to the true simulated ones (FMD = 2.4 across the SpaH and FMD = 0.5 across the SpaB simulated data sets) suggesting in turn that the core model is also robust to the presence of SNPs under selection (at least in moderate proportion). Conversely, a misspecification of the prior \( \Omega \), as investigated here by similarly analyzing the SpaH (respectively SpaB) data sets under the core, the STD and the AUX models but setting \( \Omega = \Omega_{\text{HSA}}^{\text{amp}} \) (respectively \( \Omega = \Omega_{\text{BTA}}^{\text{amp}} \)), lead to an inflation of the X\( \mathbf{X} \) estimates (Figure S8). The X\( \mathbf{X} \) mean was in particular shifted away from \( J \) (number of populations) expected under neutrality (see also Figure 5 in Günther and Coop (2013)). As a consequence, the overall performances of the X\( \mathbf{X} \)-based criterion were clearly impacted (see Table S1 and ROC in Figure S7).

Interestingly, under both the STD and AUX models, the distribution of the X\( \mathbf{X} \) for SNPs associated to the population covariable was similar to the neutral SNP one, whatever the underlying \( \beta_{i} \) (Figure S6). Accordingly, the corresponding true positive rates were close to the nominal POD threshold in Table 1. This suggests that both covariable models allow to efficiently correct the X\( \mathbf{X} \) estimates for the ("fixed") covariable effect of the associated SNPs.

**Performance of the models to detect SNP associated to a population–specific covariable.**

The performances of the STD and AUX models to identify SNPs associated to a population–specific covariable were further evaluated using results obtained on the SpaH and SpaB data sets (see above). As shown in Figure 4, the Importance Sampling estimates of the \( \beta_{i} \) coefficients (computed from parameter values sampled under the core model) were found less accurate than posterior mean estimates obtained from values sampled under the STD or AUX models. For smaller \( |\beta_{i}| \) however, the introduction of the auxiliary variable (AUX model) tended to shrink the estimates towards zero in the SpaB data sets probably due, here also, to a less powerful design (three times less populations).

![Figure 4 about here.](http://dx.doi.org/10.1101/023721)

Accordingly, the BF’s estimated under the AUX model (BF\( _{\text{mc}} \)) had more power to identify SNPs associated to the population-specific covariables than the corresponding BF\( _{\text{is}} \) (Table 2 and Figure S9). Indeed, although constrained by construction to a maximal value (here 53.0 dB) that both depends on the number of MCMC samples (here 1,000) and on the prior expectation of \( P \) (here 0.01), at the "decisive evidence" threshold of 20 dB (Jeffreys 1961), the TPR for SNPs with a simulated \( |\beta_{i}| = 0.05 \) were for instance equal to 81.7% with BF\( _{\text{mc}} \) for the SpaH data compared to 31.9% with the BF\( _{\text{is}} \) based decision criterion (Table 2). For the SpaH data (but not for the SpaB ones) a similar trend was observed when comparing decision criteria based on the eBF\( _{\text{is}} \) (relying on Importance Sampling algorithm) and the eBF\( _{\text{mc}} \) as estimated under the STD model (see Table 2 and Figure S10). In addition, Table 2 shows that the intuitive, but still arbitrary, threshold of 3 on the eBF performed worse than the 20 dB threshold on the BF, particularly for the smallest \( |\beta_{i}| \). This suggests...
that a decision criterion rule relying on the BF_{mc} may be the most reliable in the context of these models.

[Table 2 about here.]

We next explored how a misspecification of the prior \( \Omega \) affected the estimation of the \( \beta \)'s and the different decision criteria. As in the previous section, we considered results obtained for the SpaH (respectively SpaB) data sets with analyses setting \( \Omega = \hat{\Omega}_{\text{SpaH}} \) (respectively \( \Omega = \hat{\Omega}_{\text{SpaB}} \)). Surprisingly, although the Importance Sampling estimates of the \( \beta \)'s obtained under the core model clearly performed poorer (particularly for the SpaB data), the estimates obtained under the STD and AUX models were not so affected (Figure S11). Nevertheless, if the resulting TPR and FPR were similar to the previous ones for the SpaH data, for the SpaB data the power to detect associated SNPs strongly decreased with both the BF_{pa} and eBP_{pa} criteria. Conversely, increased FPR were observed with the BF_{mc} (up to 22.5\%) and eBP_{mc} based decision criteria (see Table S2 and compare with Table 2). These results thus suggest that the influence of model misspecification, although unpredictable, may be critical for association studies under the STD and AUX covariate models.

**Comparison of the performances of BAYPASS with other genome–scan methods under realistic scenarios**

To compare the performances of the different approaches implemented in BAYPASS with other popular or recently developed methods, data sets simulated under three realistic scenarios were considered. Following de Villemereuil et al. (2014) (see Material and Methods), these correspond to i) an highly structured isolation with migration model (HsIMM–C); ii) an isolation with migration model (IMM); and iii) a stepping stone scenario (SS) with polygenic selection acting on an environmental gradient. In total, 300 data sets (100 per scenario), each consisting of genotyping data on 5,000 SNPs for 320 individuals belonging to 16 different populations were analyzed with BAYPASS under the core model (to estimate XIX, BF_{pa} and eBP_{pa}), the STD model (to estimate BF_{mc} and also the XIX corrected for the fixed covariable effect) and the AUX model (to estimate BF_{pa} and also the corrected XIX).

These data sets were also analyzed with five other programs (see Material and Methods), two of which, namely BAYEscan (Foll and Gaggiotti 2008) and FLK (Bonhomme et al., 2010), implementing (only) covariate–free approaches; and the three others namely BAYENV2 (Coop et al., 2010), LFMM (Frichot et al., 2013) and BAYESCVN (de Villemereuil and Gaggiotti 2015) allowing to test association with population–specific covariable.

For each scenario, average ROC and PR curves resulting from the analyses of the 100 simulated data sets are plotted for the different methods (and decision criteria) in Figure 5.

In addition, Area Under the ROC Curve (AUC) together with averaged computation times are detailed in Table 3. In agreement with previous studies (e.g., de Villemereuil et al., 2014), under such complex scenarios with polygenic selection, the association based methods clearly outperformed covariate–free approaches (BAYEscan, FLK and XIX–based criterion). For the latter however, the BAYPASS XIX (as estimated under the core model) always performed better than BAYEscan and FLK in all the three scenarios. Surprisingly, for the HsIMM–C and the SS scenarios, the BAYENV2 XIX–based criterion lead to higher AUC than its BAYPASS counterpart with a value close to the BAYENV2 BF association test (Table 3).

[Figure 5 about here.]

Among the association–based methods, BAYPASS was found to display similar performances (using the BF_{pa}, eBP_{pa} and eBP_{mc}) than LFMM–10rep, being even slightly better than single–run LFMM analyses for the IMM and SS scenarios. Both methods outperformed BAYESCVN and BAYENV2 in all scenarios (except SS scenario for the latter). It should be noted that LFMM–10rep analyses were based on individual genotyping data (and a balanced design) which represent the most favorable situation (de Villemereuil et al., 2014). The BF_{mc} criterion displayed similar performances in the PR analysis than the BAYPASS BF_{pa}, eBP_{pa} and eBP_{mc} criteria. Nevertheless, ROC AUC values were always found lower when considering BF_{mc} probably as a result of the inherent correction in the AUX model for multiple testing issues which, as expected, affects the power. Interestingly, as expected from previous results, the XIX calculated under the STD model (and to a lesser extent the AUX model) lead here to a worthless decision criterion (ROC AUC almost equal to 0.5) illustrating the efficiency of the correction for the fixed covariable effect (Table 3).

[Table 3 about here.]

Finally, under the parameter options chosen to run the different programs (see, Material and Methods), BAYPASS analyses were always among the most computational efficient approaches (Table 3). For instance, under the core model, BAYPASS was found to run 1.5 times faster than a single LFMM run.

**Performance of the Ising prior to account for SNP spatial dependency in association analyses**

To evaluate the ability of the AUX model Ising prior to capture SNP spatial dependency information, 100 data sets simulated under the HsIMMild–C scenario (see Material and Methods) were analyzed under the AUX model with three different parameterizations for the Ising prior i) b_{pa} = 0 (no spatial dependency); ii) b_{pa} = 0.5 and iii) b_{pa} = 1. For each data set, analyses with and without the causal variants were carried out and the required estimate of the covariance matrix was obtained from a preliminary analysis performed under the core model. As shown in Figure 6, increasing b_{pa} improved the mapping precision. Indeed, both a noise reduction at neutral position and a sharpening of the 95\% envelope (containing 95\% of the \( \delta \) posterior means across the 100 simulated data sets) around the selected locus can be observed (e.g., compare Figure 6A1 and A3). Interestingly, given the considered SNP density (and level of LD) excluding the causal variants had only marginal effect on the overall results.

[Figure 6 about here.]

**Analysis of the French cattle SNP data**

The XIX estimates were obtained for the 42,046 SNPs of the BTA_{snps} data (Figure S12) from the previous analysis under the core model with \( r = 1 \) (e.g., Figure 2). In agreement with above results, setting instead \( \Omega = \hat{\Omega}_{\text{BTA}} \) (the estimate of \( \Omega \) obtained in the latter analysis) gave almost identical XIX estimates (\( r = 0.995 \)). To calibrate the XIX’s, a POD containing 100,000 simulated SNPs was generated and further analyzed leading to a posterior estimate of \( \Omega \) very close to \( \hat{\Omega}_{\text{BTA}} \) (FMD = 0.098).

Similarly, the posterior means of \( \Omega \) obtained on the POD data set (\( \delta_{\text{FMD}} = 1.44 \) and \( b_{\text{FMD}} = 3.43 \), respectively) were almost equal to the ones obtained in the original analysis of the BTA_{snps} data (\( \delta_{\text{FMD}} = 1.43 \) and \( b_{\text{FMD}} = 3.44 \), respectively). This indicated that the POD faithfully mimics the real data set, allowing
the definition of relevant POD significance thresholds on XtX to identify genomic regions harboring footprints of selection.

To that end, the UMD3.1 bovine genome assembly (Liu et al. 2009) was first split into 5,400 consecutive 1-Mb windows (with a 500 kb overlap). Windows with at least two SNPs displaying XtX > 3.54 (the 0.1% POD threshold) were deemed significant and overlapping ‘significant’ windows were further merged to delineate significant regions. Among the 15 resulting regions, two regions were discarded because their peak XtX value was lower than 40.0 (the 0.01% POD threshold). As detailed in Table 4, the 13 remaining regions lie within or overlap with a Core Selective Sweep (CSS) as defined in the recent meta-analysis by Gutiérrez-Gil et al. (2015). This study combined results of 21 published genome-scans performed on European cattle populations using various alternative approaches. The proximity of the XtX peak allows to define positional candidate genes (Table 4) that have, for most regions, already been proposed (or demonstrated) to be either under selection or to control genes involved in traits targeted by selection (see Discussion).

To illustrate how information provided by population-specific covariables might help in formulating or even testing hypotheses to explain the origin of the observed footprints of selection, characteristics of the 18 cattle populations for traits related to morphology (SMS) and coat pigmentation (piebald pattern) were further analyzed within the framework developed in this study. An across population genome–wide association studies was thus carried out under both the STD and AUX models (with more consistent decisions (113 SNPs displaying both BF greater than 50kb) to the XtX peaks. Accordingly, the corresponding BF estimates decreased when estimated under the STD model, i.e. accounting for the covariables (Figure S14). For instance, the SNP under the XtX peak dropped from 76.3 to 50.3 (from 40.7 to 19.3) for region #3 (respectively region #8). Overall, the posterior means of the individual SNP βi regression coefficients estimated under the STD model ranged (in absolute value) from $2.2 \times 10^{-6}$ (respectively $1.0 \times 10^{-8}$) to 0.166 (respectively 0.233) for SMS (respectively piebald pattern). These estimates remained close to those derived from Importance Sampling algorithm, although the latter tended to be lower in absolute value (Figure S15). As expected from the above simulation studies, estimates obtained under the AUX model tended to be shrunk towards 0, which was particularly striking in the case of SMS (Figure S15).

![Figure 7 about here.](https://example.com/figure7)

Finally, analyses of association with SMS were conducted under the AUX model with three different Ising prior parametrizations ($b_{is} = 0$, $b_{is} = 0.5$ and $b_{is} = 1$) focusing on the 1,394 SNPs mapping to BTA14 (Figure 7). Under the $b_{is} = 0$ parametrization (equivalent to the AUX model analysis conducted above on a whole genome basis), four SNPs (all lying within region #12) displayed significant signals of association at the BF ≥ 20 dB threshold with a peak BFmc value of 28.5 dB at position 24.6 Mb (Figure 7A). These results, obtained on a chromosome-wide basis, provide additional support to the region #12 signal previously observed. They alternatively suggest that power of the BFmc as computed on a whole genome basis might have been altered by the small proportion of SNPs strongly associated to SMS due to multiple testing issues (which BFmc computation is not accounted for). Hence, for SMS mapping to BTA14, the BFmc estimated on the initial genome-wide analysis were almost identical to the BFis ($r = 0.993$) and highly correlated to the BFmc ($r = 0.805$) estimated in the chromosome-wide analysis. As expected from simulation results, increasing $b_{is}$ lead to refine the position of the peak toward a single SNP mapping about 400 kb upstream the PLAG1 gene (Figure 7B and C).

**Analysis of the Littorina saxatilis Pool–Seq data**

The LSAps Pool–Seq data set was first analyzed under the core model (with $\rho = 1$). In agreement with previous results (Westram et al. 2014), the resulting estimate of the population covariance matrix $\Omega$ confirmed that the 12 different Littorina populations cluster at the higher level by geographical location and then by ecotype and replicate (Figure 8A). This analysis also allowed to estimate the XtX for each of the 53,387 SNPs that were further calibrated by analyzing a POD containing 100,000 simulated SNPs to identify outlier SNPs (Figure 8). As for the cattle data analysis, the estimate of $\Omega$ on the POD was close to the matrix estimated on the original LSAps data set ($FMD = 0.516$) although the posterior means of $a_{pq}$ and $b_{pq}$ were slightly higher ($\hat{a}_{pq} = \hat{b}_{pq} = 0.370$ compared to $\tilde{a}_{pq} = \tilde{b}_{pq} = 0.214$ with the LSAps data set). In total, 169 SNPs subjected to adaptive divergence were found at the 0.01% POD significance threshold. To illustrate how the BayPASS models may help discriminating between parallel phenotype divergence from local adaptation, analyses of association were further conducted with ecotype (crab vs wave) as a categorical population–specific covariable. Among the 169 XTX outlier SNPs, 65 (respectively 75) displayed BFis > 20 dB (respectively BFmc > 20 dB) (Figure 8B). The two BF estimates resulted in consistent decisions (113 SNPs displaying both BFmc > 20 dB and BFmc > 20 dB), although at the 20 dB more SNPs were found significantly associated under the AUX model ($n = 176$) than with BFis ($n = 117$). Interestingly, several overly differentiated SNPs (high XtX value) were clearly not associated to the population ecotype covariable (small BF). These might thus be responding to other selective pressures (local adaptation) but
might also, for some of them, map to sex-chromosomes (Gautier 2014). As a consequence, SNP XIX estimated under the AUX model (i.e., corrected for the "fixed" ecotype effect) remained highly correlated with the XIX estimated under the core model (including for some XIX outliers) with the noticeable exception of the SNPs significantly associated to the ecotype. For the latter, the corrected XIX dropped to values generally far smaller than the 0.01% POD threshold (Figure 8C). Finally, Figure 8D gives the posterior mean of the SNP regression coefficients quantifying the strength of the association with the ecotype covariable. It shows that several SNPs displayed strong association signals

\[ |\beta_i| > 0.2 \] pointing towards candidate genes underlying parallel phenotype divergence. As observed above in the simulation study and in the analysis of the cattle data set, the AUX model estimates tended to be shrunken towards 0, except for the highest values (corresponding to SNPs significantly associated to the covariable) when compared to the estimates obtained under the core model (Figure S16A). A similar trend for the \( \beta_i \) estimates of the strongly associated SNPs was observed with the importance sampling estimates (Figure S16B).

[Figure 8 about here.]

**Discussion**

The main purpose of this study was to develop a general and robust Bayesian framework to identify genomic regions subjected to adaptive divergence across populations by extending the approach first described in Coop et al. (2010) and Günther and Coop (2013). Because of the central role played in the underlying models by the scaled population covariance matrix, a first objective was to improve the precision of its estimation. To that end, instead of defining an Inverse-Wishart prior on \( \Omega \) or on the CCP model, a Wishart prior defined on the precision matrix \( \Lambda = \Omega^{-1} \) was rather considered and equivalently parametrized with an identity scale matrix but varying number of degrees of freedom \( \rho \). As the extensive simulation study revealed, the most accurate estimates were obtained by setting \( \rho = 1 \) (instead of the number of populations which is equivalent to Coop et al. (2010)) leading to a weaker (and singular) informative Wishart prior. Although flexible, the purely instrumental nature of the \( \Omega \) prior parametrization considered in our models makes it difficult to incorporate prior and possibly relevant information about the populations under study. For instance, spatially (Guillot et al. 2014) or even phylogenetically explicit prior might represent in some context attractive alternatives, borrowing for the latter on population genetics theory to model the effect of the demographic history on the covariance matrix (Lipson et al. 2013; Pickrell and Pritchard 2012). Apart from investigating different \( \Omega \) prior specification, additional levels in the hierarchical models were also introduced to estimate the parameters of the \( (\beta, \gamma) \) prior distribution on the ancestral allele frequency. Interestingly, estimating these parameters improved robustness to the SNP ascertainment scheme, in particular when the allele frequency spectrum is biased towards poorly informative SNPs as generally obtained with data from whole genome sequencing experiment (e.g., Pool–Seq data). Simulation results on MAF filtered data sets also suggested that these additional levels might reduce sensitivity of the models to SNP ascertainment bias characterizing genotyping data obtained from SNP chip. Finally, inclusion of a moderate proportion of SNPs under selection did not significantly affect estimation of \( \Omega \). Overall, it can be concluded that the core model parametrized with a weakly informative Wishart prior \( \rho = 1 \) and that includes the estimation of the parameters \( \alpha, \beta \) and \( \gamma \) provides a general robust and accurate approach to estimate \( \Omega \) even with a few thousands of genotyped SNPs. It should also be noted that it outperforms previous implementations carried out under a similar hierarchical Bayesian framework, as in the BAYENV software (Coop et al. 2010), or relying on moment-based estimators Lipson et al. (2013); Bonhomme et al. (2010); Pickrell and Pritchard (2012) (see, e.g., Figure 3). As the latter are based on sample allele frequencies, they also remain more sensitive to sample size (and coverage for Pool–Seq data) and, more importantly, they do not allow combining estimation of both the ancestral allele frequencies and covariance matrix that represent a serious issue for small and/or unbalanced designs. Finally, as briefly sketched with visualizations based on correlation plot or hierarchical trees in the present study, the estimation procedure implemented in the BAYPASS core model might be quite relevant for demographic inference purposes since the matrix \( \Omega \) has already been shown to be informative about the population history Lipson et al. (2013); Pickrell and Pritchard (2012).

Accounting for \( \Omega \) renders the identification of SNPs subjected to selection less sensitive to the confounding effect of demography (Bonhomme et al. 2010; Günther and Coop 2013). To that end the XIX introduced by Günther and Coop (2013) provides a valuable differentiation measure for genome scan of adaptive divergence. While XIX might be viewed as a Bayesian counterpart of the FLK statistic (Bonhomme et al. 2010), its computation allows considering population histories more complex than bifurcating trees (i.e., including migration or ancestral admixture events) not to mention improved precision in the estimation of the underlying \( \Omega \). For practical purposes however, defining significance threshold for the XIX remains challenging. Indeed, although the XIX are expected under the neutral model to be \( \chi^2 \)-squared distributed (Günther and Coop 2013), the Bayesian (hierarchical) model based procedure leads to shrink the XIX posterior mean towards their prior mean (Gelman et al. 2003). As a consequence, an empirical posterior checking procedure, similar in essence to the one previously used in a similar context (Vitalis et al. 2014), was evaluated here. It represents a relevant alternative to an arbitrary threshold although it comes at a cost of some additional computational burden. The procedure indeed consists in analyzing (POD) data simulated under the inference model with hyperparameters \( \alpha, \beta \) and \( \gamma \) set equal to those estimated on the real data. Comparing the \( \Omega, \alpha, \beta \) and \( \gamma \) estimates obtained on the POD to the original ones ensures that the simulated data provide good surrogates to neutrally evolving SNPs under a demographic history similar to that of the sampled populations. More generally, given the efficiency of the simulation procedure, such simulated data sets might also be relevant to investigate the properties of other estimators of genetic diversity or to evaluate the robustness of various approaches to demographic confounding factors. In the context of this study, a better estimation of \( \Omega \) was hence shown to improve the performance of the XIX-based differentiation test and association studies with population–specific covariables under the STD and AUX covariate models.

Based on the STD model, Coop et al. (2010) relied on an importance sampling (\( BF_{\text{est}} \)) estimates of the BF to assess association of allele frequency differences with population–specific covariables. A major advantage of this algorithm stems from its computational efficiency, since only parameter samples drawn from the core model are required. However, the simulation study showed
As a proof of concept, analyses were carried out on real data sets from both model and non-model species. Results obtained for the French cattle data demonstrated the versatility of the approach and illustrated how association studies could give insights into the putative selective forces targeting footprints of selection. As a matter of expedience we only hereby focused on the thirteen strongest differentiation signals. As expected from the importance of coat pigmentation in the definition of breed standards, at least six genomic regions contained genes known to be associated to coat color and patterning variation, in agreement with previous genome scan for footprints of selection (see Gutiérrez-Gil et al. 2015, for review). These include MC1R (region #13) that corresponds to the locus Extension with three alleles identified to date in cattle responsible for the red, black (or combination of both) colors (Seo et al. 2007). Similarly, variants localized within the KIT (region #7) and PAX5 (region #10) genes were found highly associated to patterned pigmentation (proportion of black) in Holstein, accounting for respectively 9.4% and 6.0% of the trait variance (Hayes et al. 2010). Within region #7, KIT clusters with KDR (closest to the Xtx peak) and PDGFA, two other Tyrosine kinase receptor genes that have also been proposed as candidate coloration genes under selection in other studies (Flori et al. 2009; Gutiérrez-Gil et al. 2015; Qanbari et al. 2014). In region #11, the Xtx peak was less than 25 kb upstream EDN3 that is involved in melanocyte development and within which mutations were found associated to pigmentation defects in mouse, human and also chicken (Bennett and Lamoreux 2003; Dorshorst et al. 2011; Saldana-Caboverde and Kos 2010). Accordingly, Qanbari et al. (2014) recently found a variant in the vicinity of EDN3 strongly associated with coat spotting phenotype of bulls (measured as the proportion of their daughters without spotting) in the Fleckvieh breed. The
peak in region #2 was 100 kb upstream the KITLG gene which is involved in the roan phenotype (mixture of pigmented and white hairs) observed in several cattle breeds (Seitz et al. 1999). Mutations in this gene have also been found to underlie skin pigmentation diseases in human (Picardo and Cardinali 2011). Finally, region #5 contains the LEP1 gene (100 kb from the XIX peak) that has recently been demonstrated to be tightly involved in blond hair color in (human) Europeans (Guenther et al. 2014).

Three other regions contained genes that affect cattle body conformation. These include region #1 containing the myostatin gene (MSTN), one of the best known examples of economically important genes in farm animals since it plays an inhibitory role in the development and regulation of skeletal muscle mass (Stickens et al. 2011). MSTN is in particular responsible for the so-called double muscling phenotype in cattle (Grobet et al. 1997). Region #12 contains PLAG1 that has been demonstrated to influence bovine stature (Karim et al. 2011). Similarly, region #6 contain encompasses the NCPAG-LCORL cluster in which several polymorphisms have been found strongly associated to height in human (Allen et al. 2010), horse (Signer-Hasler et al. 2012) and cattle (Pryce et al. 2011). However, combining results from a genome-scan for adaptive selection with a comprehensive genome-wide association study with milk production traits in the Holstein cattle breed, Xu et al. (2015) proposed the LAP3 gene (within which the Xtx peak mapped as the main driver of a selective sweep overlapping with region #12). Regarding the four remaining regions (#2, #4, #8 and #9), the retained candidate genes corresponded to the gene within which the XIX peak is located (NUDCD3, RPS26 and VDAC1 for regions #2, #4 and #9 respectively) or is the closest (less than 15 kb from ALB for region #8). As for RPS26, although NUDCD3 has been highlighted in other studies (e.g., Flori et al. 2009; Xu et al. 2015), the poorly known function of these genes makes highly speculative any interpretation of the origin of the signals. Conversely, the various and important roles played by ALB (bovine serum albumin precursor) do not allow a clear hypothesis to be formulated about the trait underlying the region #8 signal. More presumably, due to the role of VDAC1 in male fertility (Kwon et al. 2013), the footprint of selection observed in region #9 might result from selection for a trait related to reproduction. Overall, association analyses carried out under the covariate models revealed strong association of SNPs within KITLG (region #3), KIT (region #7) and EDN3 (region #11) with variation in the piebald pattern across the populations thereby supporting the hypothesis of selection on coat coloration to be the main driver of the three corresponding signatures of selection. These results also confirm the already well known key role of these genes on coloration patterning. Interestingly, the observed association signals within ALB (region #8) also suggest that this gene might influence coat coloration in cattle which, to our knowledge, has not been previously reported. Finally, association studies on the SMS trait suggested that PLAG1 (region #12) has been under strong selection in European cattle and contribute to morphological differences across the breeds. Yet, the strongest association signals was 400 kb upstream PLAG1 suggesting the existence of some functional variants (possibly in regulatory regions) different from those already reported (Karim et al. 2011) although such results need to be confirmed with denser SNP data sets. Conversely, no association signal was found within the selection signature under region #6 adding more credits to selection for milk production (Xu et al. 2015) as the main underlying adaptive constraint rather than morphological trait as previously hypothesized (see above). Analysis of the Littorina saxatilis Pool-Seq data (Westram et al. 2014) illustrate how BAYPASS can be helpful to realize a typology of the markers relative to an ecological covariate in a non-model species. In agreement with the original results, several genes represent good candidate to underlie parallel phenotypic divergence in this organism and might deserve follow-up validation studies. From a practical point of view however, compared to combining several pairwise FST population tests (Westram et al. 2014), the approach proposed here greatly simplified the analyses and the biological interpretation of the results while allowing both an optimal use of the data and a better control for multiple testing issues.

Overall, the models described here and implemented in the software package BAYPASS provide a general and robust framework to better understand the patterns of genetic divergence across populations at the genomic level. They allow i) an accurate estimation of the scaled covariance matrix whose interpretation gives insights into the history of the studied populations; ii) a robust identification of overly differentiated markers by correcting for confounding demographic effects; and iii) robust analyses of association of SNP with population-specific covariates giving in turn insights into the origin of the observed footprints of selection. In practice, when compared to BAYENV2 , BAYPASS lead to a more accurate and robust estimation of the matrix Ω (and the related measures) and thus improved the performances of the different tests. In addition, various program options were developed to investigate the different modeling extensions, including analyses under the STD and AUX models and exploration of the Ising prior parameters to incorporate LD information. More generally, as demonstrated by the analysis of individual-based simulated data sets, the method developed in this study was found to be among the most efficient in terms of power, robustness and computational cost when compared to the other state-of-the-art or recently developed genome scan methods. Moreover, as opposed to most of the currently available approaches, the different decision measures (XtX, eBP and BF) can be computed for both allele (from standard individual genotyping experiments) and read (from Pool-Seq experiments) count data (while also accommodating missing data). Finally, although computation times scale roughly linearly with the data set complexity (number of populations × number of markers), for very large data sets, several strategies might be efficient to reduce computational burden. For instance, because estimation of Ω was found robust to moderate ascertainment bias, one may filter low polymorphic markers (e.g., overall MAF<0.01) since those are not informative for genome scan purposes, and/or consider sub-sampling of the initial data set (e.g., chromosome-wise analyses).

Acknowledgments

I wish to thank Anja Westram for providing early access to the Littorina data and Pierre de Villemeureuil for providing the polygenic data sets simulated under the HsIMM, IMM and SS scenarios (and information about the underlying Python/SimuPOP scripts). I also wish to thank the two anonymous reviewers for their valuable comments. I am finally grateful to the genotoul computing resources. This work was partly funded by the ERA–bioinformatics platform Toulouse Midi–Pyrenees for providing their valuable comments. I am finally grateful to the genotoul narios (and information about the underlying Python/SimuPOP extensions, including analyses under the STD and AUX models to realize a typology of the markers relative to an ecological co-variate in a non–model species. In agreement with the original results, several genes represent good candidate to underlie parallel phenotypic divergence in this organism and might deserve follow-up validation studies. From a practical point of view however, compared to combining several pairwise FST population tests (Westram et al. 2014), the approach proposed here greatly simplified the analyses and the biological interpretation of the results while allowing both an optimal use of the data and a better control for multiple testing issues.

Overall, the models described here and implemented in the software package BAYPASS provide a general and robust framework to better understand the patterns of genetic divergence across populations at the genomic level. They allow i) an accurate estimation of the scaled covariance matrix whose interpretation gives insights into the history of the studied populations; ii) a robust identification of overly differentiated markers by correcting for confounding demographic effects; and iii) robust analyses of association of SNP with population-specific covariates giving in turn insights into the origin of the observed footprints of selection. In practice, when compared to BAYENV2, BAYPASS lead to a more accurate and robust estimation of the matrix Ω (and the related measures) and thus improved the performances of the different tests. In addition, various program options were developed to investigate the different modeling extensions, including analyses under the STD and AUX models and exploration of the Ising prior parameters to incorporate LD information. More generally, as demonstrated by the analysis of individual-based simulated data sets, the method developed in this study was found to be among the most efficient in terms of power, robustness and computational cost when compared to the other state-of-the-art or recently developed genome scan methods. Moreover, as opposed to most of the currently available approaches, the different decision measures (XtX, eBP and BF) can be computed for both allele (from standard individual genotyping experiments) and read (from Pool-Seq experiments) count data (while also accommodating missing data). Finally, although computation times scale roughly linearly with the data set complexity (number of populations × number of markers), for very large data sets, several strategies might be efficient to reduce computational burden. For instance, because estimation of Ω was found robust to moderate ascertainment bias, one may filter low polymorphic markers (e.g., overall MAF<0.01) since those are not informative for genome scan purposes, and/or consider sub-sampling of the initial data set (e.g., chromosome-wise analyses).

Acknowledgments

I wish to thank Anja Westram for providing early access to the Littorina data and Pierre de Villemeureuil for providing the polygenic data sets simulated under the HsIMM, IMM and SS scenarios (and information about the underlying Python/SimuPOP scripts). I also wish to thank the two anonymous reviewers for their valuable comments. I am finally grateful to the genotoul bioinformatics platform Toulouse Midi–Pyrenees for providing computing resources. This work was partly funded by the ERA–Net BiodivERsA2013–48 (EXOTIC), with the national funders FRB, ANR, MEDDE, BELSPO, PT-DLR and DFG, part of the 2012–2013 BiodivERsA call for research proposals.
Literature Cited


Hasselman, B., 2015 geigen: Calculate Generalized Eigenvalues of a Matrix Pair. R package 1.5.


Marin, J.-M. and C. P. Robert, 2014 Bayesian essentials with R.


Wei, T., 2013 corrplot: Visualization of a correlation matrix. R package version 0.73.


List of Figures

1 Directed Acyclic Graphs of the different hierarchical Bayesian models considered in the study and implemented in the BAYPASS software. See the main text for details about the underlying parameters and modeling assumptions. ......... 18

2 Representation of the scaled covariance matrices \( \Omega \) among 18 French cattle breeds \( \Omega^{\text{bews}}_{\text{BTA}} \) (A and C) as estimated from BAYENV2 (Coop et al. 2010) and \( \Omega^{\text{bpas}}_{\text{BTA}} \) (B and D) as estimated from BAYPASS under the core model with \( \rho = 1 \). Both estimates are based on the analysis of the BTA\text{bpas} data consisting of 42,036 autosomal SNPs (see the main text). Breed codes (and branches) are colored according to their broad geographic origins (see File S4 and Gautier et al. (2010b) for further details) with populations in red, blue and green originating from South-Western and Central France, North-Western France, and Eastern France (e.g. Alps) ......... 19

3 FMD distances (Förster and Moonen 2003) between the matrices used to simulate the data sets and their estimates. Simulation scenarios are defined according to the matrix \( \Omega_{\text{sim}} \) used to simulated the data (\( \Omega_{\text{sim}} = \Omega^{\text{bews}}_{\text{HSA}} \) in A and B; and \( \Omega_{\text{sim}} = \Omega^{\text{bpas}}_{\text{BTA}} \) in C and D) and the sampling distribution of the \( \eta_i \)'s (Unif(0,1) in A and C and Beta(0.2,0.2) in B and D). For each scenario, ten independent data sets of 1,000, 5,000, 10,000, and 25,000 markers were simulated (160 data sets in total) and analyzed with BAYENV2 (Coop et al. 2010) and four alternative BAYPASS model parameterizations (i) \( \rho = 1 \); ii) \( \rho = 1 \) and \( \eta_i = b_{\eta_i} = 1 \); iii) \( \rho = f \) and \( \eta_i = b_{\eta_i} = 1 \); iv) \( \rho = f \) and \( \eta_i = b_{\eta_i} = 1 \). As a matter of comparisons, the FLK frequentist estimate (Bonhomme et al. 2010) of the covariance matrices was also computed. Each point in the curves is the average of the ten pairwise FMD distances between the underlying \( \Omega_{\text{sim}} \) and each of the \( \Omega \) estimated in the ten corresponding simulation replicates. ......... 20

4 Distribution of the estimated SNP regression coefficients \( \hat{\beta}_i \) as a function of their simulated values obtained from analyses under three different model parameterizations with \( \Omega = \Omega^{\text{bews}}_{\text{HSA}} \) (for SpaH data) and \( \Omega = \Omega^{\text{bpas}}_{\text{BTA}} \) (for SpaB data). For a given scenario (SpaH and SpaB), results from the ten replicates are combined. ......... 21

5 Comparison of the performances of BAYPASS with other genome--scan methods based on data simulated under three different scenarios (HsIMM–C, IMM and SS) that include polygenic selection. For each scenario, ROC and PR curves corresponding to the different approaches (and decision criteria) were plotted from the actual TPR, FPR and FDR estimates averaged over the results of 100 independent data sets. ......... 22

6 Comparison of the performances of three different Ising prior parameterizations for the AUX model (\( b_{\eta_i} = 0 \), \( b_{\eta_i} = 0.5 \) and \( b_{\eta_i} = 1 \) ) on the HsIMMmid-C simulated data sets with (A1, A2 and A3) and without (B1, B2 and B3) the causal variants. Each panel summarizes the distribution at each SNP position (x-axis) of the \( \delta_i \) (auxiliary variable) posterior means over the 100 independent simulated data with the median values in green and the 95% envelope in orange. Each simulated data set consisted of 5,000 SNPs spread on 5 chromosomes of 4 cM. In the middle of the third chromosome (indicated by an arrow), a locus with a strong effect on individual fitness was defined by two consecutive SNPs strongly associated with the environmental covariable. ......... 23

7 Results of the BTA14 chromosome–wide association analyses with SMS under three different Ising prior parameterizations of the AUX model (A \( b_{\eta_i} = 0 \); B \( b_{\eta_i} = 0.5 \) and C \( b_{\eta_i} = 1 \) ). Plots give, for each SNP, the posterior probability of being associated (\( P [ \hat{\beta}_i = 1 | \text{data} ] \)) according to their physical position on the chromosome. The main figure focuses on the region surrounding the candidate gene PLAG1 (positioned on the vertical dotted line) while results over the whole chromosome are represented in the upper left corner. In A), the horizontal dotted line represents the threshold for decisive evidence (corresponding to \( BF = 20 \) dB). ......... 24

8 Analysis of the LSA\text{bps} Pool–Seq data. A) Inferred relationship among the 12 Littorina populations represented by a correlation plot and a hierarchical clustering tree derived from the matrix \( \Omega \) estimated under the core model (with \( \rho = 1 \)). Each population code indicates its geographic origin (SP for Spain, SW for Sweden and UK for the United Kingdom), its ecotype (crab or wave) and the replicate number (1 or 2). B) SNP \( \text{XtX} \) (estimated under the core model) as a function of the \( \text{BF}_{\eta} \) for association with the ecotype population covariable. The vertical dotted line represents the 0.1% POD significance threshold (\( \text{XtX}=28.1 \)) and the horizontal dotted line represents the 20 dB threshold for BF. The point symbol indicates significance of the different \( \text{XtX} \) values, \( \text{BF}_{\eta} \) and \( \text{BF}_{\text{mc}} \) (AUX model) estimates. C) SNP \( \text{XtX} \) corrected for the ecotype population covariable (estimated under the STD model) as a function of \( \text{XtX} \) estimated under the core model. The vertical and horizontal dotted lines represent the 0.1% POD significance threshold (\( \text{XtX}=28.1 \)). Point symbols follow the same nomenclature as in B). D) Estimates of SNP regression coefficients (\( \hat{\beta}_i \)) on the ecotype population covariable (under the AUX model) as a function of \( \text{XtX} \). Point symbols follow the same nomenclature as in B).
Figure 1 Directed Acyclic Graphs of the different hierarchical Bayesian models considered in the study and implemented in the BAYPASS software. See the main text for details about the underlying parameters and modeling assumptions.
Figure 2 Representations of the scaled covariance matrices $\Omega$ among 18 French cattle breeds $\hat{\Omega}_B$ (A and C) as estimated from BAYENV2 (Coop et al. 2010) and $\hat{\Omega}^\text{bpas}_B$ (B and D) as estimated from BAYPASS under the core model with $\rho = 1$. Both estimates are based on the analysis of the BTA.snp data set consisting of 42,036 autosomal SNPs (see the main text). Breed codes (and branches) are colored according to their broad geographic origins (see File S4 and Gautier et al. (2010b) for further details) with populations in red, blue and green originating from South-Western and Central France, North-Western France, and Eastern France (e.g. Alps).
Figure 3 FMD distances ( Förstner and Moonen 2003) between the matrices used to simulate the data sets and their estimates. Simulation scenarios are defined according to the matrix $\Omega_{\text{sim}}$ used to simulated the data ($\Omega_{\text{sim}} = \hat{\Omega}^{\text{b pas}}_{\text{HSAD}}$ in A and B; and $\Omega_{\text{sim}} = \hat{\Omega}^{\text{b pas}}_{\text{BTA}}$ in C and D) and the sampling distribution of the $\pi_i$’s (Unif(0,1) in A and C and Beta(0.2,0.2) in B and D). For each scenario, ten independent data sets of 1,000, 5,000, 10,000 and 25,000 markers were simulated (160 data sets in total) and analyzed with BayENV2 (Coop et al. 2010) and four alternative BAYPASS model parameterizations (i) $\rho = 1$; ii) $\rho = 1$ and $a_{\pi} = b_{\pi} = 1$; iii) $\rho = J$ and; iv) $\rho = J$ and $a_{\pi} = b_{\pi} = 1$). As a matter of comparisons, the FLK frequentist estimate (Bonhomme et al. 2010) of the covariance matrices was also computed. Each point in the curves is the average of the ten pairwise FMD distances between the underlying $\Omega_{\text{sim}}$ and each of the $\hat{\Omega}$ estimated in the ten corresponding simulation replicates.
Figure 4 Distribution of the estimated SNP regression coefficients $\beta_i$ as a function of their simulated values obtained from analyses under three different model parameterizations with $\Omega = \hat{\Omega}_{\text{HSA}}$ (for SpaH data) and $\Omega = \hat{\Omega}_{\text{BTA}}$ (for SpaB data). For a given scenario (SpaH and SpaB), results from the ten replicates are combined.
Figure 5 Comparison of the performances of BAYPASS with other genome-scan methods based on data simulated under three different scenarios (HsIMM–C, IMM and SS) that include polygenic selection. For each scenario, ROC and PR curves corresponding to the different approaches (and decision criteria) were plotted from the actual TPR, FPR and FDR estimates averaged over the results of 100 independent data sets.
Figure 6 Comparison of the performances of three different Ising prior parameterizations for the AUX model (β = 0, β = 0.5 and β = 1) on the HsIMMld-C simulated data sets with (A1, A2 and A3) and without (B1, B2 and B3) the causal variants. Each panel summarizes the distribution at each SNP position (x-axis) of the $\delta_i$ (auxiliary variable) posterior means over the 100 independent simulated data with the median values in green and the 95% envelope in orange. Each simulated data set consisted of 5,000 SNPs spread on 5 chromosome of 4 cM. In the middle of the third chromosome (indicated by an arrow), a locus with a strong effect on individual fitness was defined by two consecutive SNPs strongly associated with the environmental covariable.
Figure 7 Results of the BTA14 chromosome–wide association analyses with SMS under three different Ising prior parameterizations of the AUX model (A) $b_{\text{is}} = 0$; B) $b_{\text{is}} = 0.5$ and; C) $b_{\text{is}} = 1$). Plots give, for each SNP, the posterior probability of being associated ($P[\delta_i = 1 | \text{data}]$) according to their physical position on the chromosome. The main figure focuses on the region surrounding the candidate gene PLAG1 (positioned on the vertical dotted line) while results over the whole chromosome are represented in the upper left corner. In A), the horizontal dotted line represents the threshold for decisive evidence (corresponding to $BF = 20$ dB).
Figure 8 Analysis of the LSAPs Pool–Seq data. A) Inferred relationship among the 12 Littorina populations represented by a correlation plot and a hierarchical clustering tree derived from the matrix $\Omega$ estimated under the core model (with $\rho = 1$). Each population code indicates its geographical origin (SP for Spain, SW for Sweden and UK for the United Kingdom), its ecotype (crab or wave) and the replicate number (1 or 2). B) SNP $\Xi X$ (estimated under the core model) as a function of the $BF_{\text{is}}$ for association with the ecotype population covariable. The vertical dotted line represents the 0.1% POD significance threshold ($\Xi X=28.1$) and the horizontal dotted line represents the 20 dB threshold for BF. The point symbol indicates significance of the different $\Xi X$ values, $BF_{\text{is}}$ and $BF_{\text{mc}}$ (AUX model) estimates. C) SNP $\Xi X$ corrected for the ecotype population covariable (estimated under the STD model) as a function of $\Xi X$ estimated under the core model. The vertical and horizontal dotted lines represent the 0.1% POD significance threshold ($\Xi X=28.1$). Point symbols follow the same nomenclature as in B). D) Estimates of SNP regression coefficients ($\beta$) on the ecotype population covariable (under the AUX model) as a function of $\Xi X$. Point symbols follow the same nomenclature as in B).
List of Tables

1 True Positive Rates (TPR) at the 1% POD threshold as a function of the simulated $|\beta_i|$ values for four different model parameterizations. TPR are given in % and were computed by combining results over the ten replicate data sets for each SpaH (and SpaB given in parenthesis) scenarios. ................................................................. 27

2 True (TPR) and False (FPR) Positive Rates as a function of the decision criterion and the model parametrization (with $\Omega = \Omega_{\text{HSA}}$ for the SpaH and $\Omega = \Omega_{\text{BTA}}$ for the SpaB data sets respectively). The thresholds are set to 20 dB for both the BF$_{is}$ and BF$_{mc}$ Bayes Factors; and to 3 for both the eBP$_{is}$ and eBP$_{mc}$ (empirical) Bayesian P-values. The true and false positive rates (given in %) are computed by combining results over the ten replicate data sets from the SpaH and SpaB (given in parenthesis) scenarios. ............................................ 28

3 Computation times and Area Under the ROC Curves (AUC in %) for the analyses of the HsIMM-C, IMM and SS data sets using the different genome-scan approaches. Computation times are averaged over the 300 analyses (100 data sets times 3 scenarios) ........................................................................................................................................... 29

4 Regions harboring footprints of selection based on the XtX measure of differentiation and association of the underlying SNPs with SMS (morphology related trait) and piebald coloration differences across the 18 French cattle breeds. For each region, the table gives the peak XtX value (and position in Mb) and the peak BF$_{is}$ and BF$_{mc}$ values in dB units (and positions in Mb) for each traits if the evidence for association is decisive (n.s. if BF < 20). The Table also gives the overlapping Core Selective Sweeps (CSS) regions (with their corresponding sizes and the number of supporting studies) from the meta-analysis by Gutiérrez-Gil et al. (2015). Finally, putative underlying candidate genes (and associated candidate functions) are proposed (see the main text). .................................................. 30
Table 1 True Positive Rates (TPR) at the 1% POD threshold as a function of the simulated $|\beta_i|$ values for four different model parameterizations. TPR are given in % and were computed by combining results over the ten replicate data sets for each SpaH (and SpaB given in parenthesis) scenarios.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>core model</th>
<th>core model with $\Omega = \Omega^{\text{sim}}$</th>
<th>STD model with $\Omega = \Omega^{\text{sim}}$</th>
<th>AUX model with $\Omega = \Omega^{\text{sim}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\beta_i</td>
<td>= 0.05$</td>
<td>2.90 (2.30)</td>
<td>9.15 (3.35)</td>
</tr>
<tr>
<td>$</td>
<td>\beta_i</td>
<td>= 0.1$</td>
<td>36.3 (13.5)</td>
<td>82.6 (22.3)</td>
</tr>
<tr>
<td>$</td>
<td>\beta_i</td>
<td>= 0.2$</td>
<td>100 (86.3)</td>
<td>100 (96.4)</td>
</tr>
<tr>
<td>Criterion</td>
<td>BF(_{ls})</td>
<td>BF(_{mc})</td>
<td>eBP(_{ls})</td>
<td>eBP(_{mc})</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>FPR</td>
<td>0.01 (0.02)</td>
<td>0.39 (0.11)</td>
<td>0.00 (2.03)</td>
<td>0.17 (0.01)</td>
</tr>
<tr>
<td>TPR (</td>
<td>(\beta_i</td>
<td>= 0.05)</td>
<td>31.9 (4.35)</td>
<td>81.7 (13.0)</td>
</tr>
<tr>
<td>TPR (</td>
<td>(\beta_i</td>
<td>= 0.1)</td>
<td>98.5 (47.0)</td>
<td>99.9 (64.1)</td>
</tr>
<tr>
<td>TPR (</td>
<td>(\beta_i</td>
<td>= 0.2)</td>
<td>100 (99.9)</td>
<td>100 (99.9)</td>
</tr>
</tbody>
</table>
Table 3 Computation times and Area Under the ROC Curves (AUC in %) for the analyses of the HsIMM-C, IMM and SS data sets using the different genome–scan approaches. Computation times are averaged over the 300 analyses (100 data sets times 3 scenarios).

<table>
<thead>
<tr>
<th>Method</th>
<th>Criterion</th>
<th>Mean (median) Computation Time in min</th>
<th>HsIMM-C</th>
<th>IMM</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAYESCAN</td>
<td>BF</td>
<td>529 (469)</td>
<td>60.13</td>
<td>53.81</td>
<td>62.05</td>
</tr>
<tr>
<td>FLK</td>
<td>FLK</td>
<td>0.16 (0.16)</td>
<td>58.92</td>
<td>61.63</td>
<td>62.17</td>
</tr>
<tr>
<td>BayEnv2</td>
<td>XiX</td>
<td>660 (358)</td>
<td>70.45</td>
<td>61.00</td>
<td>72.16</td>
</tr>
<tr>
<td></td>
<td>BF</td>
<td>70.58</td>
<td>73.84</td>
<td>81.96</td>
<td></td>
</tr>
<tr>
<td>BAYPASS (core model)</td>
<td>XiX</td>
<td>22.6 (22.2)</td>
<td>61.66</td>
<td>61.88</td>
<td>65.33</td>
</tr>
<tr>
<td></td>
<td>Bfis</td>
<td>74.36</td>
<td>78.91</td>
<td>82.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eBPis</td>
<td>74.33</td>
<td>78.78</td>
<td>82.22</td>
<td></td>
</tr>
<tr>
<td>BAYPASS (STD model)</td>
<td>XiX</td>
<td>21.4 (17.8)</td>
<td>49.85</td>
<td>49.16</td>
<td>47.72</td>
</tr>
<tr>
<td></td>
<td>eBPmc</td>
<td>74.15</td>
<td>78.76</td>
<td>82.22</td>
<td></td>
</tr>
<tr>
<td>BAYPASS (AUX model)</td>
<td>XiX</td>
<td>45.3 (44.9)</td>
<td>60.60</td>
<td>59.82</td>
<td>61.08</td>
</tr>
<tr>
<td></td>
<td>Bfmc</td>
<td>58.30</td>
<td>65.24</td>
<td>70.51</td>
<td></td>
</tr>
<tr>
<td>BAYESCENV</td>
<td>Post. Prob.</td>
<td>510 (478)</td>
<td>66.93</td>
<td>62.34</td>
<td>70.36</td>
</tr>
<tr>
<td>LFMMb</td>
<td>P–value</td>
<td>33.0 (30.4)</td>
<td>75.58</td>
<td>78.29</td>
<td>81.98</td>
</tr>
<tr>
<td>LFMM–10repb</td>
<td>P–value</td>
<td>310 (248)</td>
<td>76.27</td>
<td>79.37</td>
<td>82.56</td>
</tr>
</tbody>
</table>

* not accounting for the time required to estimate the covariance matrix (obtained here after running BAYPASS under the core model)

b Analyses were carried out using individual genotyping data rather than (population) allele count which provides the best performances (see, e.g., de Villemereuil et al. 2014)

c Not accounting for the time required to estimate the number of latent factor K (set here to K=15)
Table 4: Regions harboring footprints of selection based on the XtX measure of differentiation and association of the underlying SNPs with SMS (morphology related trait) and piebald coloration differences across the 18 French cattle breeds. For each region, the table gives the peak XtX value (and position in Mb) and the peak BF\textsubscript{is} and BF\textsubscript{mc} values in dB units (and positions in Mb) for each traits if the evidence for association is decisive (n.s. if BF < 20). The Table also gives the overlapping Core Selective Sweeps (CSS) regions (with their corresponding sizes and the number of supporting studies) from the meta-analysis by Gutiérrez-Gil et al. (2015). Finally, putative underlying candidate genes (and associated candidate functions) are proposed (see the main text).

<table>
<thead>
<tr>
<th>ID</th>
<th>Region</th>
<th>Overlapping CSS\textsuperscript{a}</th>
<th>XtX</th>
<th>BF\textsubscript{is}-BF\textsubscript{mc} for morphology</th>
<th>BF\textsubscript{is}-BF\textsubscript{mc} for piebald</th>
<th>Candidate Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>BTA02:4.17-8.64</td>
<td>CSS–32</td>
<td>58.4</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>MSTN:6.214-6.220</td>
<td>conformation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.47</td>
<td>(6.70)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#2</td>
<td>BTA04:76.7-78.6</td>
<td>CSS–99</td>
<td>45.8</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>NUDCD3: 77.599-77.670</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.93</td>
<td>(77.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#3</td>
<td>BTA05:18.0-19.5</td>
<td>CSS–103</td>
<td>76.3</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>KITLG:18.318-18.377</td>
<td>pigmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.50</td>
<td>(18.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#4</td>
<td>BTA05:54.7-58.6</td>
<td>CSS–109</td>
<td>54.7</td>
<td>26.6-n.s.</td>
<td>n.s.-n.s.</td>
<td>RPS26:57.604-57.607</td>
<td>pigmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.93</td>
<td>(57.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#5</td>
<td>BTA06:17.6-19.2</td>
<td>CSS–117</td>
<td>63.2</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>LEF1:18.335-18.451</td>
<td>unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.54</td>
<td>(18.2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#6</td>
<td>BTA06:37.8-40.2</td>
<td>CSS–123</td>
<td>69.4</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>LAP3:38.575-38.600</td>
<td>(conformation/dairy traits)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.40</td>
<td>(38.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#7</td>
<td>BTA06:65.5-74.9</td>
<td>CSS–130</td>
<td>55.6</td>
<td>n.s.-n.s.</td>
<td>37.42-26.45</td>
<td>KIT:71.796-71.917</td>
<td>conformation/dairy traits</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9.38</td>
<td>(72.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#8</td>
<td>BTA06:89.6-90.6</td>
<td>CSS–130</td>
<td>40.7</td>
<td>n.s.-n.s.</td>
<td>52.07-38.76</td>
<td>ALB:90.233-90.251</td>
<td>conformation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.02</td>
<td>(90.2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#9</td>
<td>BTA07:46.4-47.8</td>
<td>CSS–141</td>
<td>46.5</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>VDAC1:47.248-47.273</td>
<td>reproduction?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.48</td>
<td>(47.3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#10</td>
<td>BTA08:61.4-63.3</td>
<td>CSS–162</td>
<td>49.8</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>PAX5:61.400-61.580</td>
<td>pigmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.94</td>
<td>(61.8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#11</td>
<td>BTA13:56.6-58.6</td>
<td>CSS–248</td>
<td>71.6</td>
<td>23.7-n.s.</td>
<td>26.58-n.s.</td>
<td>EDN3:57.571-57.597</td>
<td>pigmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.98</td>
<td>(57.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#12</td>
<td>BTA14:22.1-28.8</td>
<td>CSS–254</td>
<td>52.0</td>
<td>35.7-n.s.</td>
<td>n.s.-n.s.</td>
<td>PLAG1:25.007-25.009</td>
<td>conformation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6.76</td>
<td>(24.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#13</td>
<td>BTA18:13.3-16.0</td>
<td>CSS–297</td>
<td>51.8</td>
<td>n.s.-n.s.</td>
<td>n.s.-n.s.</td>
<td>MC1R:14.757-14.759</td>
<td>pigmentation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.75</td>
<td>(14.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a} Full descriptions of the CSS (including references to the original studies) are provided in Table S2 by Gutiérrez-Gil et al. (2015)