Genome Scans for Detecting Footprints of Local Adaptation Using a Bayesian Factor Model

Nicolas Duforet-Frebourg,[1] Eric Bazin,[2] and Michael G.B. Blum*,[1]

[1]Laboratoire TIMC-IMAG, UMR 5525, Centre National de la Recherche Scientifique, Université Joseph Fourier, Grenoble, France

[2]Laboratoire d'Ecologie Alpine, UMR 5553, Centre National de la Recherche Scientifique, Université Joseph Fourier, Grenoble, France

***Corresponding author:** E-mail: michael.blum@imag.fr.

Associate editor: John Novembre

Abstract

There is a considerable impetus in population genomics to pinpoint loci involved in local adaptation. A powerful approach to find genomic regions subject to local adaptation is to genotype numerous molecular markers and look for outlier loci. One of the most common approaches for selection scans is based on statistics that measure population differentiation such as F_{ST}. However, there are important caveats with approaches related to F_{ST} because they require grouping individuals into populations and they additionally assume a particular model of population structure. Here, we implement a more flexible individual-based approach based on Bayesian factor models. Factor models capture population structure with latent variables called factors, which can describe clustering of individuals into populations or isolation-by-distance patterns. Using hierarchical Bayesian modeling, we both infer population structure and identify outlier loci that are candidates for local adaptation. In order to identify outlier loci, the hierarchical factor model searches for loci that are atypically related to population structure as measured by the latent factors. In a model of population divergence, we show that it can achieve a 2-fold or more reduction of false discovery rate compared with the software *BayeScan* or with an F_{ST} approach. We show that our software can handle large data sets by analyzing the single nucleotide polymorphisms of the Human Genome Diversity Project. The Bayesian factor model is implemented in the open-source PCAdapt software.

Key words: **F_{ST}, population structure, landscape genetics, population genomics, selection scans.**

Introduction

With the development of sequencing and genotyping technologies, there is a considerable impetus in population genomics to pinpoint loci involved in local adaptation (Akey et al. 2002; Bonin et al. 2006). A working hypothesis of population genomics is that most loci have neutral patterns of variation that are similarly affected by demographic processes, whereas loci targeted by natural selection have atypical patterns (Luikart et al. 2003). Although this working hypothesis is fiercely debated (Sella et al. 2009), it has led to a protocol for finding genomic regions subject to local adaptation which consists of genotyping numerous molecular markers and looking for outlier loci. Different regions of the genome are expected to exhibit highly variable levels of genetic differentiation between populations ranging from genomic regions exhibiting little differentiation to regions where genetic divergence is extremely pronounced. Although high levels of differentiation can be explained by various causes, adaptation of individuals to their local environment is a prominent explanation such that patterns of differentiation for adaptive loci exceed neutral expectations (Nosil and Buerkle 2010). Measures of genetic differentiation between populations such as F_{ST} have been commonly used to find outlier loci, although there are many alternative approaches (Oleksyk et al. 2010). A proof of concept for approaches based on genetic differentiation was provided when studying human adaptation to altitude; the most differentiated variants between a Tibetan population living in a hypoxic environment and a lowland Han Chinese population were found in hypoxia-inducible transcription factors (Yi et al. 2010; Xu et al. 2011).

Genome scans based on F_{ST} were proposed by Lewontin and Krakauer (1973) and have been considerably expanded since (Beaumont and Nichols 1996; Vitalis et al. 2001; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler et al. 2008; Guo et al. 2009; Bazin et al. 2010; Bonhomme et al. 2010; Gompert and Buerkle 2011; Fariello et al. 2013). They are not limited to two populations as in the adaptation-to-altitude example and can be used with multiple populations. One possibility is to compute an overall F_{ST} measure of genetic differentiation and to determine a threshold at which the null hypothesis of neutral evolution can be rejected (Beaumont and Nichols 1996). Another possibility is to adopt a model-based perspective by implementing the multinomial-Dirichlet model or F model, which is parameterized by population-specific F statistics (Beaumont and Balding 2004). The F statistics can be interpreted as measures of divergence from a common immigrant gene pool (Wright 1931) or as a divergence from an initial and hypothetical ancestral population (Nicholson et al. 2002). The Bayesian approach for distinguishing between neutral or adaptive evolution offers the opportunity to assign a probability to each of the two evolutionary models at each locus (Foll and Gaggiotti 2008; Riebler et al. 2008). In the following, approaches based

Article

on F_{ST} or on population-specific F statistics are referred as genome scans based on F statistics. There are many software-implementing genome scans based on F statistics (e.g., *BayeScan*, *DetSel*, *fdist2*, and *Lositan*), and they contribute to the popularity of this approach in population genomics (Beaumont and Nichols 1996; Vitalis et al. 2003; Antao et al. 2008; Foll and Gaggiotti 2008).

However, a major issue with genome scans based on F statistics is that they can generate a high rate of false positives for both biological and statistical reasons (Bierne et al. 2013). Here, we propose to address the statistical and computational problems that arise with F statistics. The first problem arises because F statistics have been derived under Wright's F model of population subdivision, which assumes a particular covariance structure for gene frequencies among populations (Bierne et al. 2013; Fourcade et al. 2013). When spatial structure departs from Wright's island model of population subdivision, genome scans based on F statistics produce many false positives (Bierne et al. 2013). Alternative statistical measures that account for population structure have recently been proposed (Bonhomme et al. 2010; Günther and Coop 2013). A second potential problem concerns the computational burden of some Bayesian approaches, which can become an obstacle with a large number of single nucleotide polymorphisms (SNPs) (Lange et al. 2014). The last intrinsic problem of genome scans based on F statistics is that individuals should be grouped into populations. However, it has been advocated in landscape genetics to rather work at the scale of individuals because it avoids potential bias in identifying populations in advance and it offers the opportunity to conduct studies at a finer scale (Manel and Holderegger 2013).

To tackle the aforementioned problems, we propose a statistical method based on a Bayesian factor model (West 2003) to pick outlier loci involved in local adaptation. With factor models, we seek to jointly determine population structure and outlier loci. Factor models are strongly related to principal component analysis (PCA) because they both approximate the matrix of individual genotypes by a product of two lower-rank matrices, albeit using different constraints and priors for the lower-rank matrices (Engelhardt and Stephens 2010). One of the two matrices encodes population structure using latent factors, whereas the second matrix measures to what extent each individual SNP is related to the pattern of population structure. The proposed factor model seeks for loci that are atypically related to population structure. To show the potential of factor models for genome scans, we consider two examples. First, we consider a model of population divergence. In this example, we compare false discovery rates (FDRs) obtained from the proposed factor model with *BayeScan* and with a genome scan based on F_{ST}. The second example is a model of isolation-by-distance with selection. It is an instance of how factor models can be used to detect local adaptation when it would be arbitrary to group individuals into populations. Finally, we analyze the HGDP human data set (Li et al. 2008) to provide an example of how factor models can be used to detect local adaptation with a large number of SNPs.

New Approaches

We denote the $n \times p$ matrix of centered allele counts by $\mathbf{Y}$, where n is the number of individuals and p is the number of loci. The elements $Y_{i\ell}$, $i = 1,\ldots,n$, $\ell = 1,\ldots,p$ correspond to the centered allele counts of the ith individual at locus ℓ. Before centering, the allele counts belong to {0, 1} or {0, 1, 2} for haploid and diploid species, respectively. After centering, each column of the matrix $\mathbf{Y}$ has a mean of 0.

Factor models assume that the matrix of column-centered genotypes $\mathbf{Y}$ can be written as a product of two lower-rank matrices

$$\mathbf{Y} = \mathbf{UV} + \epsilon, \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are of dimension $(n \times K)$ and $(K \times p)$, respectively, K is an hyper parameter that is much smaller than n and p, and ϵ is the matrix of residuals. In the following, the column vectors of $\mathbf{U}$ referred to as "factors" or "latent factors" are denoted by $\mathbf{U}_1, \ldots, \mathbf{U}_K$. Factor models assume that the vector—of size n—of centered allele counts $\mathbf{Y}_\ell$ can be obtained as follows:

$$\mathbf{Y}_\ell = \sum_{k=1}^{K} \mathbf{U}_k V_{k\ell} + \epsilon_\ell,\ \ell = 1, \ldots, p, \tag{2}$$

where ϵ_ℓ is a vector containing n independent Gaussian residuals of variance σ^2 and $V_{k\ell}$ are the elements of the matrix $\mathbf{V}$. Assuming that the K factors are known, then the elements $V_{k\ell}$ of the matrix $\mathbf{V}$ are the regression coefficients—sometimes called *factor loadings*—obtained after regressing the vector of centered allele counts $\mathbf{Y}_\ell$ by the K factors $\mathbf{U}_1, \ldots, \mathbf{U}_K$. As candidates for local adaptation, we consider the loci ℓ that have large (in absolute value) regression coefficients $V_{k\ell}$ for one of the factors $\mathbf{U}_1, \ldots, \mathbf{U}_K$. In factor models, the K factors $\mathbf{U}_1, \ldots, \mathbf{U}_K$ are in fact unknown and have to be estimated; they are parameters of the model and represent population structure (Engelhardt and Stephens 2010). In our proposed framework, outlier loci are loci that are excessively related to population structure, the latter being measured by the K latent factors. After statistical inference, the factors $\mathbf{U}_1, \ldots, \mathbf{U}_K$ are ordered by decreasing variances $\sigma_1^2 > \ldots > \sigma_K^2$, where σ_k^2 measures the variance of the regression coefficients $V_{k\ell}$, $\ell = 1 \ldots p$, for the kth factor.

To provide a concrete example of how factors represent population structure, we consider a model of population divergence. We assume that an initial population splits into two populations A and B that diverged according to neutral evolution. The initial neutral divergence of duration T is followed by two concomitant splits where each daughter population A and B splits into two subpopulations (A_1, A_2) and (B_1, B_2). In contrast to the initial divergence which is purely neutral, the second phase of divergence between populations assumes some local adaptation with a small proportion of SNPs conferring selective advantage (fig. 1). We fit the factor models with $K = 3$ and we display the three factors in figure 1. The first factor discriminates individuals according to the initial split and the second and third factors discriminate individuals according to the subsequent splits which separate

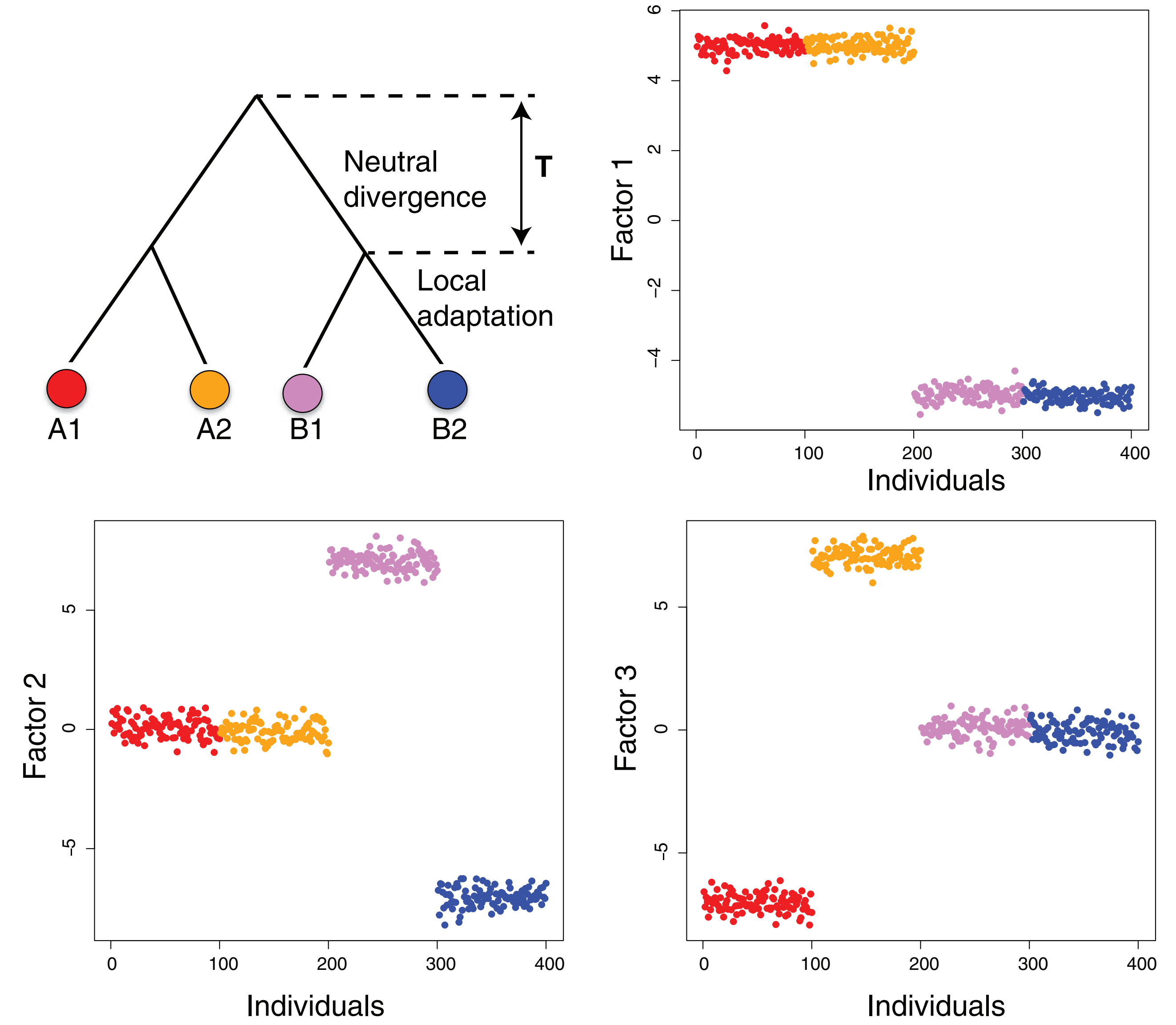

Fig. 1. Values of the $K = 3$ factors for a population divergence model with four populations. The upper-left panel shows the model of population divergence. The other panels show the values of the first three factors and each dot corresponds to one individual. As candidates for local adaptation, the factor model with $K = 3$ looks for SNPs whose variation is atypically well explained by one of the three factors. For the simulations, the effective population size is $N_e = 1{,}000$ diploid individuals in each population, 50 individuals are sampled in each population, and the neutral divergence time is $T = 200$ generations and is twice as long as the second phase during which there is adaptation. We assume that 400 SNPs, among a total of 10,000 SNPs, can confer selective advantage. The 400 adaptive SNPs are split into four sets of 100 SNPs, each set corresponding to adaptation in one of the four evolutionary lineages. We consider a selection coefficient $s = 0.1$ for homozygotes carrying two adaptive alleles and $s = 0.05$ for heterozygotes.

subpopulation B_1 from B_2 (second factor) and subpopulation A_1 from A_2 (third factor).

We now specify how we measure the degree of outlyingness for each locus using a Bayesian criterion. To account for outlier and non-outlier loci, we assume that, at a given locus ℓ, the vector of regression coefficients $\mathbf{V}_\ell = (V_{1\ell},\ldots,V_{K\ell})$ comes from a mixture of two different distributions. We introduce a vector $\mathbf{z}$ of indicator variables $(z_1,\ldots,z_p)$ whose elements are equal to 0 for non-outlier loci and take values $1,\ldots,K$ for outlier loci. For any locus ℓ, either outlier or nonoutlier, we assume that the vector $\mathbf{V}_\ell = (V_{1\ell},\ldots,V_{K\ell})$ is composed of independent Gaussian random variables. The model for non-outlier loci is a product of Gaussian distributions

$$V_{k\ell} \mid z_\ell = 0 \sim \mathcal{N}(0,\sigma_k^2),\ k = 1,\ldots,K, \qquad (3)$$

where $\mathcal{N}(m,\sigma^2)$ denotes the Gaussian distribution with mean m and variance σ^2. To model outlier loci, we consider a variance–inflation model which assumes an inflated variance to account for outlier loci (Box and Tiao 1968; Devlin and Roeder 1999). The model for outlier loci is itself a mixture model with K components of equal weights, where the kth component assumes an inflated variance for the kth regression coefficient but not for the other ones. Denoting the variance–inflation parameter for factor k by c_k^2 ($c_k^2 > 1$), the k_0th component of the mixture model for outlier assumes a product of Gaussian distributions

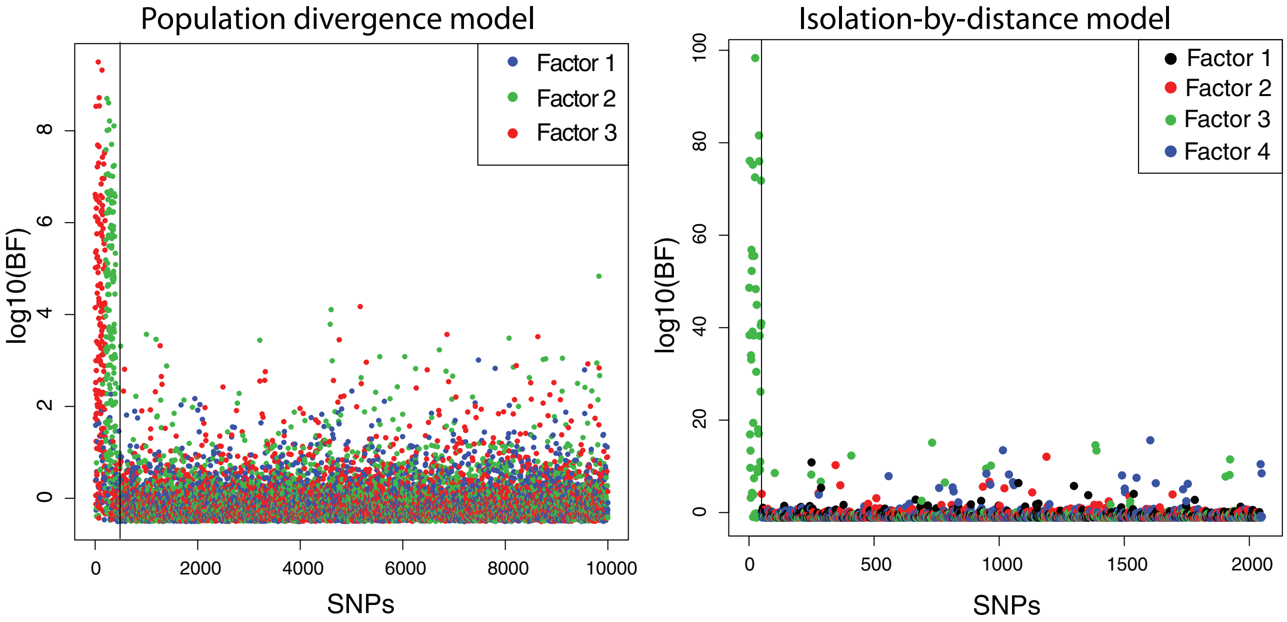

FIG. 2. Bayes factors for the two different simulation examples. The SNPs under selection are located on the left-hand side of the vertical bar.

$$
\begin{aligned}
V_{k\ell} \mid z_\ell &= k \sim \mathcal{N}(0, c_k^2 \sigma_k^2),\ k = k_0 \\
V_{k\ell} \mid z_\ell &= k \sim \mathcal{N}(0, \sigma_k^2),\ k \neq k_0
\end{aligned}
\qquad (4)
$$

The model for outliers has been chosen for the sake of interpretability. Each outlier locus can be related to one of the K factors because outlier loci should be atypically explained by one of the K factors. To measure the strength of evidence for outlyingness for each locus, we compute the Bayes factor of the outlier model against the non-outlier model. If a locus is considered as an outlier, the factor with which there is an atypical correlation is found by computing the posterior probabilities of each of the K components of the outlier mixture model. To account for linkage disequilibrium, we additionally consider a Potts model that encourages outlier loci to be clustered in the genome (see Materials and Methods).

When fitting the factor model with $K = 3$ to data simulated under the scenario of population divergence depicted in figure 1, the outlier model of equation (4) assumes three different types of outlier loci: loci that have large genetic differentiation when comparing the pair of subpopulations (A_1,A_2) with the pair (B_1,B_2) (large values of $|V_{1\ell}|$), loci that have large genetic differentiation when comparing subpopulation B_1 with B_2 (large values of $|V_{2\ell}|$), and loci that have large genetic differentiation when comparing subpopulation A_1 with A_2 (large values of $|V_{3\ell}|$). Because the simulation assumes that the initial period of divergence is purely neutral, the first types of outliers (large values of $|V_{1\ell}|$) are in fact false positives.

Results

Simulation Study

Population Divergence Model

The first simulation study investigates to what extent factor models better account for population structure than methods based on F statistics. We consider the model of population divergence depicted in figure 1. An initial neutral divergence is followed by adaptive divergence where 4% of the 10,000 simulated SNPs are involved in local adaptation. The set of adaptive SNPs is split in four equal parts and each subset of SNP confers a selective advantage in only one of the four populations. When the initial neutral divergence time T is null, the population tree is star-like, and the assumption of the F model is valid. As the initial neutral divergence time T increases, the departure from the F model increases. The neutral divergence time T is scaled so that $T = 1$ means that the neutral and adaptive phases are of same duration.

First, we present results using the factor model with $K = 3$ factors that is optimal because there are four populations in the divergence model (Patterson et al. 2006). We consider a long-enough divergence time $T = 2$ so that the first factor corresponds to the initial and neutral divergence, whereas the second and third factors correspond to the subsequent divergence events during which biological adaptation took place (fig. 1). The SNPs that have been truly involved in biological adaptation are usually associated with the correct factor because, among the 400 truly adaptive SNPs, 81% are associated with the second and third factor and this proportion raises to 98% (respectively 92%) when considering the 195 (respectively 305) adaptive SNPs with Bayes factors larger than 10 (respectively 1) (fig. 2).

Then, we compare the FDRs of three different approaches including the proposed factor model, *BayeScan* (version 2.1), and genome scans based on the F_{ST} statistic. For both *BayeScan* and the proposed factor model, we use Bayes factors for ranking SNPs, whereas we use F_{ST} values for the last method. More precisely, we use the q values for ranking SNPs with *BayeScan*, but, by definition of the q value, it provides the same ranking as the Bayes factors. To determine a threshold above which SNPs are considered as outliers, we enlarge the lists of top-ranked SNPs, provided by each method, until each

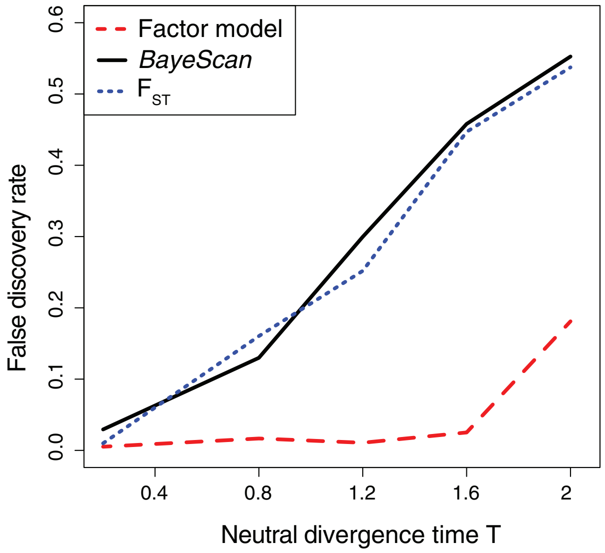

Fig. 3. False discovery rate as a function of the initial divergence time T in the population divergence model of figure 1. For the proposed factor model, Bayes factors are used for ranking SNPs, whereas we use q values with *BayeScan* and F_{ST} values for the genome scan based on F_{ST} values. To determine a threshold above which SNPs are considered as outliers, we constrain the lists of SNPs provided by each method to contain 50% of the 400 SNPs truly involved in local adaptation. The neutral divergence time T is scaled so that $T = 1$ means that the neutral and adaptive phases are of same duration, which is of 100 generations. For the simulations, the effective population size is $N_e = 1{,}000$ diploid individuals in each population and 50 individuals are sampled in each population. We assume that 400 SNPs, among a total of 10,000 SNPs, can confer selective advantage. The 400 adaptive SNPs are split into four sets of 100 SNPs, each set corresponding to adaptation in one of the four evolutionary lineages. We consider a selection coefficient $s = 0.1$ for homozygotes carrying two adaptive alleles and $s = 0.05$ for heterozygotes.

of them contains 50% of the 400 truly adaptive SNPs. This procedure amounts to setting the sensitivity to 50% (the sensitivity is also called recall rate in machine learning). Figure 3 shows that for all methods, the FDR is below 5% when the population tree is almost star-like ($T = 0.04$) but it increases with the initial neutral divergence time T. Although the FDR always increases with T, the FDRs of the factor model are always smaller than FDRs obtained with F_{ST} and *BayeScan*. For instance, when $T = 1$, the FDRs obtained with F_{ST} and *BayeScan* are between 20% and 30%, whereas it is smaller than 5% with the factor model. Instead of using a threshold of 50%, we also constrain the lists of SNPs to contain 25% or 75% of the truly adaptive SNP (i.e., setting the sensitivity to 25% or 75%). As for the 50% threshold, all methods have small FDR for sufficiently small initial divergence time T, and, as T increases, FDR increases at a slower rate for the factor model (supplementary fig. S1, Supplementary Material online). In summary, the FDR increases as the model of divergence deviates from a star-like phylogeny, but compared with other methods, the factor model reduces the proportion of false discoveries by a factor of 2 or more when there is a strong-enough deviation from the star-like assumption ($T > 0.8$).

The results presented so far were obtained using the factor model with $K = 3$ factors. By increasing the values of K from 1 to 6, we find that, compared with $K = 3$, the FDR drastically increases for underspecification of K ($K < 3$) but is almost insensitive to overspecification of K ($K > 3$, supplementary fig. S2, Supplementary Material online). We also compute the mean squared error (MSE) of equation (2) for different values of K to determine if the MSE can be a guide for choosing K. The MSE decreases from $K = 1$ to $K = 3$ before staying almost constant as K continues to grow (supplementary fig. S3, Supplementary Material online). In this example of population divergence, the MSE suggests choosing $K = 3$, but choosing a more complex model ($K > 3$) would provide comparable FDRs.

Isolation-By-Distance Model

The second simulation study provides an example of how to search for biological adaptation when there is isolation-by-distance. Approaches based on F statistics would require to group individuals into populations, and we want to avoid that. On a two-dimensional 10 × 10 grid, we simulate a stepping-stone model with selection acting on individuals located in the lower-right corner of the grid. We sample 10 diploid individuals at each of the 100 demes. A total of 50 out of 2,050 SNPs confer selective advantage in this region and the selection coefficient decreases gradually when moving away from the point where selection is maximal.

With the factor model, the selection gradient is reflected in a different factor depending on the value of the selection coefficients (results not shown). Here, we choose the intensity of selection so that the selection gradient becomes visible in the third factor (fig. 4). The other factors have spatial patterns that are typical for isolation-by-distance models (Novembre and Stephens 2008; Engelhardt and Stephens 2010). We choose $K = 4$ because the MSE decreases from $K = 1$ to $K = 4$ before being almost constant (supplementary fig. S3, Supplementary Material online). In terms of FDR, this choice of K is not optimal because $K = 3$ would provide smaller FDR (supplementary fig. S4, Supplementary Material online). However, as in the first example, smaller values of K, compared with the optimal value ($K < 3$), increase FDR drastically, whereas too large values of K ($K > 3$) increase the optimal FDR more moderately (supplementary fig. S4, Supplementary Material online). With $K = 4$, the FDR is equal to 0% when considering the top 25 SNPs, which corresponds to a sensitivity of 50%. However, when setting the sensitivity to 75%, the FDR increases to 30%, which corresponds to 38 true positive SNPs among a list of 54 SNPs. The 50 truly adaptive SNPs are all correctly associated with the factor corresponding to biological adaptation, which is the third factor here (fig. 2). When decreasing the number of sampled individuals from 10 to 1, the FDR, obtained with a sensitivity of 50%, increases considerably from 0% to 91% (supplementary fig. S5, Supplementary Material online).

We also investigate to what extent scaling the data matrix Y such that all columns have unit variance affects the results. For both models of population divergence and isolation-by-distance, we compare the FDR obtained when fitting the

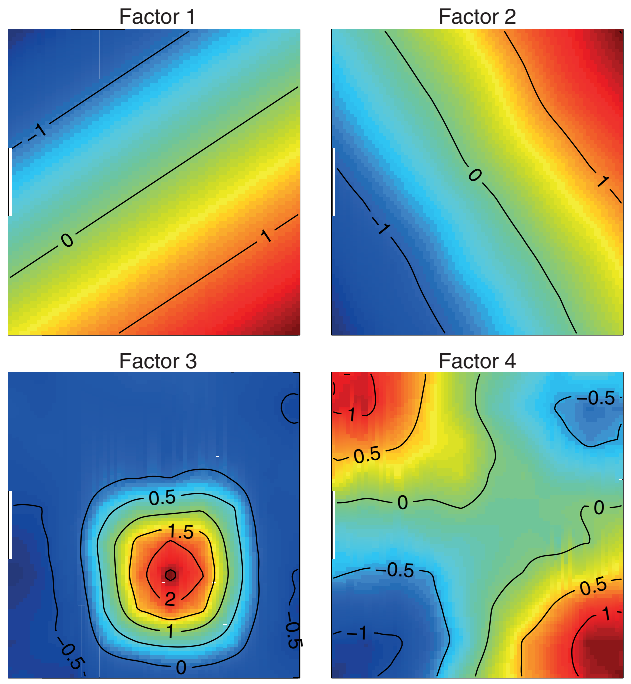

Fig. 4. Values of the $K = 4$ factors for the two-dimensional isolation-by-distance model. The different panels show the values of the first four factors when projected onto the two-dimensional space. As candidates for local adaptation, the factor model with $K = 4$ looks for SNPs whose variation is atypically well explained by one of the four factors. Spatial interpolation of the factors is obtained using the *Krig* function that is available from the *fields* R package. Each of the 100 populations in the grid has an effective population size of $N_e = 1{,}000$ diploid individuals with 10 sampled individuals in each population. After simulating an equilibrium stepping-stone model with a migration rate of $4N_em = 8$ per generation, we impose a selection gradient where 50 SNPs confer selective advantage during 100 generations. The selection coefficient $s = 0.1$ is maximal in population 64, which is located in the lower-right quarter of the grid. It is divided by 2 in the eight neighboring populations and again divided by 2 for the second layer of neighbors.

factor model with the scaled or unscaled data matrix *Y*. We do not find univocal evidence in favor or against the scaling. For the model of population divergence, no scaling is preferable, whereas the results are more complex for the isolation-by-distance model: with a sensitivity of 25%, no scaling is again preferable but scaling the data decreases the FDR when the sensitivity is larger than 50% (supplementary fig. S6, Supplementary Material online). For the model of isolation-by-distance, we compute the minimum allele frequency of the SNPs that are part of the list with largest Bayes factors (setting the sensitivity at 25%). We find that scaling the data matrix increases the proposition of low-frequency variants among the list of top SNPs, and this phenomenon is much more pronounced when looking at the false positives only in the top list (supplementary fig. S7, Supplementary Material online).

Analysis of Human SNP Data

The HGDP data set contains 644,199 SNPs, after removal of the SNPs on the sex chromosomes and on the mitochondrion, which have been typed for 1,043 individuals coming from 53 different populations (Li et al. 2008). First, we fit the factor model to the unscaled matrix *Y* of SNPs using different values of *K*. In contrast with the two previous examples, there is no value of *K* at which the MSE stops to decrease (supplementary fig. S3, Supplementary Material online). By looking at the different factors (fig. 5 and supplementary fig. S8, Supplementary Material online), we decide to consider a model that captures genetic differentiation between, but not within, continents. When considering $K = 8$, we find that factors 5–8 capture population structure within continents (supplementary fig. S8, Supplementary Material online) and we choose not to consider evolutionary processes acting at this scale. Using the fact that we are interested in genetic differentiation between continents, we consider a factor model with $K = 4$. The first factor mostly contrasts African

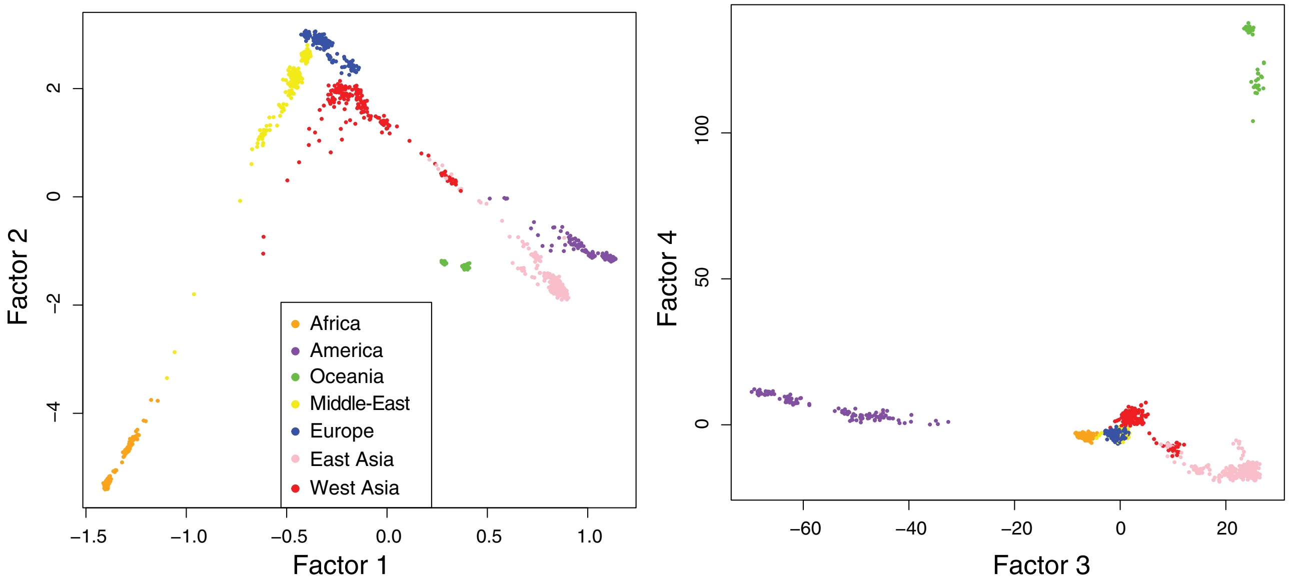

Fig. 5. Values of the $K = 4$ factors for the HGDP data set.

from Asiatic and Native American individuals and the second factor mainly discriminates African from Middle-Eastern and Western Eurasian individuals. The third factor distinguishes Native American individuals—coming from Central and South America—from the rest of the sample, whereas the last factor separates individuals from Oceania from the rest of the sample (fig. 5). We additionally check that genomic regions of strong linkage disequilibrium do not distort the latent factors. Using window sizes of 100 SNPs and a step size of 10 SNPs, we prune with *PLINK* the SNPs so that there are no more pairs of SNPs with a r^2 larger than 0.2 resulting in a total of 160,257 SNPs (Purcell et al. 2007). Comparing the four latent factors with and without pruning, we find r^2 values larger than 0.93 indicating that LD does not distort the ascertainment of population structure.

We choose to restrict our analysis to the 5,000 top-hit SNPs (supplementary table S1, Supplementary Material online). Their values of the Bayes factors range from 1.03 to 5.05 on a $\log_{10}$ scale. The two SNPs with the largest Bayes factors (rs1834640 and rs2250072) are correlated with the second factor. They are located on chromosome 15 and the closest gene is *SLC24A5*, which is located at 20–30 kb from the SNPs. Among the 5,000 SNPs with largest Bayes factors, 851 are related to factor 1, 844 with factor 2, 1,982 with factor 3, and 1,323 with factor 4. For each of the four sublists, we further provide information for the ten SNPs with the largest Bayes factors (table 1):

- For the first factor, although we consider ten different SNPs, only two genomic regions are found. One of the two genomic regions is located on chromosome 10 and downstream of the oncogene *CYP26A1* whose expression is enhanced in sunlight-damaged human skin (Osanai and Lee 2011; Mallick et al. 2013). The other SNPs were found in the *SM6* gene which is implicated in the structural maintenance of chromosome protein 6 and which has already been picked as a candidate for selection in another scan with the HGDP sample (Hao et al. 2013). For all the ten SNPs, we investigate the worldwide repartition of allele frequencies with the ALFRED database (Rajeevan et al. 2012). East Asiatic and Native American populations have allele frequencies that are different from the rest of the sample (supplementary table S2, Supplementary Material online) as can be predicted when looking at the geographic repartition of the first factor (fig. 5).
- For the outlier SNPs associated with the second factor, the allelic frequencies were mostly different when comparing Western Eurasian individuals with the rest of the sample (supplementary table S2, Supplementary Material online). In addition to the SNPs close to the *SLC24A5* gene that is associated with light skin in Western Eurasia (Canfield et al. 2013), we also find four other regions located close to the following genes: *EDAR* in chromosome 2 which has been associated with various traits including hair thickness and sweating (Kamberov et al. 2013), *SLC35F3* in chromosome 1, *KIF3A* in chromosome 5, *RABGAP1* and *STRBP* in chromosome 9 with the latter being involved in spermatogenesis, and *MYO5C* and *DUT* in chromosome 15.
- For the third factor, eight of the ten SNPs with the largest Bayes factor are found in a 1-Mb region of chromosome 22, which encompasses many different genes (table 1). For the SNPs in this large region of chromosome 22, the allelic frequencies mostly differ between Native Americans and the rest of the sample. For sub-Saharan African populations, allelic frequencies of these SNPs are intermediate with Pygmies populations having frequencies that are often the most similar to the Native Americans (supplementary table S2, Supplementary Material online).
- The allele frequencies of the SNPs that are the most associated with the fourth factor mostly differ between individuals from Oceania (Papuan and Melanesian) and the rest of the sample with Native Americans and Pygmies population having, for some SNPs, allele frequencies that are the most similar to the Oceanians (supplementary table S2, Supplementary Material online). Among the ten outlier SNPs, four SNPs are located in chromosome 8 and four SNPs are located in chromosome 17. Among the 1,323 SNPs associated with the fourth factor, there is an excess of outlier SNPs in chromosome 8 (supplementary fig. S9, Supplementary Material online) pointing to a prominent role of its genes in adaptation to the local conditions of Oceania. There are different genomic regions with large Bayes factors in chromosome 8 and one of these genomic regions encompasses *RP1L1*, a gene often found in selection scans (Barreiro et al. 2008) and related to eye diseases (Davidson et al. 2013).

We also perform a Gene Ontology (GO) enrichment analysis on human genes using the 5,000 SNPs with the largest Bayes factors. We find significant enrichment of biological processes for each of the four factors (supplementary table S3, Supplementary Material online). Some interesting instances of the enriched gene ontologies include three different GO terms related to regulation of hormone secretion for the first factor, enrichment of homophilic cell adhesion for the third factor, and aging for the fourth factor. Finally, we look at a catalog of published GWAS (Welter et al. 2014) to search for enrichment of outlier SNPs related to a particular phenotype (supplementary table S4, Supplementary Material online). The traits that are the most associated with the outlier SNPs are height (6 SNPs), obesity and weight (5 SNPs), and Crohn's disease (5 SNPs).

Discussion

Based on a Bayesian factor model, we provide a new approach to detect loci subject to local adaptation. The hierarchical factor model considers the SNPs that are atypically related to population structure as outliers and candidates for local adaptation. Population structure is captured by a set of *K* latent factors. Simulations showed that latent factors can adequately describe clustering of individuals into populations (fig. 1), isolation-by-distance patterns, and gradients of selection (fig. 4). Similarly to approaches based on measures of genetic differentiation, the hierarchical factor model is a statistical outlier approach where local adaptation is not

Table 1. List of the Ten SNPs with Largest Bayes Factors for Each of the Four Factors Obtained with the HGDP Data Set.

Chromosome	rs Identifier	Physical Position[a]	Closest Gene	Dist.[b]	Log$_{10}$(BF)[c]	Factor No.
chr 10	4918664	94911055	*CYP26A1*	83	2.8	1
chr 10	10882168	94919424	*CYP26A1*	92	2.7	1
chr 10	7091054	95008434	*MYOF*	48	2.7	1
chr 10	11187300	94910281	*CYP26A1*	83	2.6	1
chr 2	7556886	17771611	*SMC6*	0	2.6	1
chr 10	12220128	94965001	*MYOF*	91	2.5	1
chr 10	6583859	94883463	*CYP26A1*	56	2.5	1
chr 10	4918924	94966946	*MYOF*	89	2.4	1
chr 2	1834619	17764966	*SMC6*	0	2.4	1
chr 2	4578856	17716869	*SMC6*	0	2.4	1
chr 15	1834640	46179457	*SLC24A5*	21	5.1	2
chr 15	2250072	46172199	*SLC24A5*	29	4.2	2
chr 2	260714	108928927	*EDAR*	0	3.2	2
chr 1	7531501	232404926	*SLC35F3*	0	2.9	2
chr 15	11637235	46420445	*DUT*	0	2.9	2
chr 9	10760260	124753347	*RABGAP1*	0	2.9	2
chr 9	2416899	125054924	*STRBP*	0	2.8	2
chr 5	2406410	132093779	*KIF3A*	0	2.8	2
chr 15	3751631	50321636	*MYO5C*	0	2.8	2
chr 9	618746	124777344	*RABGAP1*	0	2.8	2
chr 22	139553	40517145	*MEI1*	0	4.1	3
chr 22	5996039	40311903	*PMM1*	0	4.0	3
chr 22	8139993	40325281	*DESI1*	0	4.0	3
chr 22	126092	40508387	*MEI1*	0	4.0	3
chr 22	1005402	39621676	*XPNPEP3*	0	3.6	3
chr 22	8137373	40059162	*ZC3H7B*	0	3.6	3
chr 22	133074	39408419	*MCHR1*	0	3.6	3
chr 22	9611613	40291777	*CSDC2*	0	3.5	3
chr 20	2424641	24665867	*SYNDIG1*	71	3.5	3
chr 14	2600814	46054248	*LINC00871*	13	3.5	3
chr 8	16892216	120271074	*MAL2*	19	2.8	4
chr 8	6990312	110671493	*SYBU*	0	2.8	4
chr 17	9908046	50918781	*MMD*	64	2.7	4
chr 17	575873	39055489	*MEOX1*	18	2.6	4
chr 4	4691075	164468935	*NPY1R*	0	2.6	4
chr 17	4471745	50923883	*MMD*	70	2.5	4
chr 14	12891534	80069114	*CEP128*	0	2.5	4
chr 8	6988341	110653602	*SYBU*	2	2.5	4
chr 8	12216712	9933221	*MSRA*	16	2.5	4
chr 17	11869714	45942306	*MYCBPAP*	0	2.5	4

[a] The positions are given for the NCBI36 assembly.

[b] Dist. is the distance from the closest gene and is measured in kilobase pairs.

[c] Log$_{10}$(BF) is the logarithm (in base 10) of the Bayes factor.

modeled explicitly. In contrast, another recent statistical approach for selection scans provides a generative model for local adaptation based on a diffusion approximation (Vitalis et al. 2014).

However, there are also important differences with approaches based on measures of genetic differentiation. Compared with the software *BayeScan* or genome scans based on F_{ST}, the factor model does not assume a particular model of population structure. In a model of population divergence, we show that removing the assumptions of the *F* model considerably reduces the FDR. To explain why the factor model generates fewer false discoveries, we introduce the notions of mechanistic and phenomenological models (Hilborn and Mangel 1997). Mechanistic models aim to mimic the biological processes that are thought to have given rise to the data, whereas phenomenological models seek only to best describe the data using a statistical model. In the spectrum between mechanistic and phenomenological models, the *F* model would stand close to mechanistic models, whereas factor models would be closer to the phenomenological ones. Mechanistic models are appealing because they provide quantitative measures that can be related to biologically meaningful parameters. For instance, F statistics measure genetic drift which can be related to migration rates, divergence times, or population sizes. In contrast, phenomenological models work with mathematical abstractions such as latent factors that can be difficult to interpret biologically. The downside of mechanistic models is that violation of the modeling assumption can invalidate the proposed framework and generate many false discoveries in the context of selection scans. The F model assumes a particular covariance matrix between populations which is found with star-like population trees for instance (Bonhomme et al. 2010). However, more complex models of population structure can arise for various reasons including non-instantaneous divergence and isolation-by-distance, and they will violate the mechanistic assumptions and make phenomenological models preferable.

Although PCA or the related factor model is generally used to investigate population structure, there have already been several attempts at performing selection scans based on these statistical approaches. The first idea is to compute F_{ST} values between pairs of populations that contain the top and bottom individuals for each principal component (Abdellaoui et al. 2013). This approach provides a list of outliers that are specific to each principal component in the same way as the hierarchical factor model of equations (2)–(4) provides outliers that are related to one of the K factors. A second proposition involves new interpretations of PCA related to F statistics, which provide statistical measures to detect local adaptation (Laloë and Gautier 2011). A last and recent proposition called “logistic factor analysis” adds a logistic link function to the factor model (eq. 2) in order to guarantee that the predicted values can be interpreted as frequencies because they lie between 0 and 1 (Hao et al. 2013). Loci involved in biological adaptation were scanned using a deviance statistic (Hao et al. 2013). These related approaches are built on the success of PCA and factor models to capture population structure with a small number of variables.

Choosing the dimension K of the statistical model that ascertains population structure is a recurrent problem. One possibility is to use an *objective* approach based on a quantitative criterion. Examples of such objective criteria include the Δ_K measure to detect the number of clusters using the software STRUCTURE (Evanno et al. 2005) or the Tracy–Widom statistic to choose the dimension K in PCA (Patterson et al. 2006). Another possibility is to adopt a *subjective* approach and to choose a value of K such that increasing K would provide results that are considered to be of too little interest. With the proposed Bayesian factor model, we implemented both approaches. For the simulations, choosing K based on the MSE of equation (2) works well because the MSE stops to decrease when K increases beyond a certain value. However, for the human data, the choice is more complex because the MSE decreases regularly as K increases. We chose $K = 4$ because we were only interested in biological adaptation that is related to genetic differentiation between continents, but we acknowledge that major adaptive processes also occur within continents (Jarvis et al. 2012). The subjective choice of K is related to the choice of pairwise comparisons when using F_{ST} between pairs of populations for genome scans (Nosil et al. 2008). Choosing which populations should be compared or which latent factors should be kept pertains to the biological questions addressed by the genome scan. To provide recommendations for choosing K, we suggest to fit the hierarchical factor model with different values of K in order to investigate if there is a value of K, at which the MSE stops or almost stops to decrease. If not, the choice of K can be based on subjective arguments where the latent factors of too little interest can be discarded.

Although the proposed factor model provides fewer false discoveries than approaches based on F statistics, there remain caveats and possible improvements. First, one of our objectives was to propose a method for selection scans that avoids the computational burden of some Bayesian approaches, which can become a serious obstacle when analyzing large-scale SNP data. However, this objective is only partly fulfilled. The downside of our approach is that it relies on a Markov chain Monte Carlo (MCMC) algorithm that quickly grinds to a halt under the sheer mass of SNP data (Lange et al. 2014). Fortunately, the MCMC algorithm is based on a Gibbs sampler that alternates the computation of least square solutions, which are fast to compute. For the HGDP data set (644,199 SNPs), the run-time ranges from 13 to 16 h using a single computer processor (2.4 GHz, 64-bit Intel Xeon) when K increases from 1 to 8. Second, we considered a particular outlier model (eq. 4) for the sake of interpretability which assumes that outliers should be atypically explained by one latent factor. However, other outlier models would be possible, for instance, by imposing a sparsity prior that forces most of the factor loadings to be null (Carvalho et al. 2008) or by assuming that outliers should be atypically related to the K latent factors and not to only one of them. For handling the two mentioned caveats, we are currently working on the development of a faster version of our software where various outlier models are evaluated. Third, we investigate the effect of scaling the data matrix Y and find that it can increase or decrease the FDR because of two opposing effects. The advantage of the scaling is that it makes the regression coefficients (the factor loadings) more comparable because all centered allele frequencies are at the same scale. The downside of scaling is that it gives more importance to the low-frequency variants (supplementary fig. S7, Supplementary Material online) for which the statistical estimates of population differentiation are highly variable (Meirmans and Hedrick 2011). The fourth caveat occurs if local adaptation has acted in directions that do not align with the underlying population structure. For example, in investigating worldwide convergent evolution of high-altitude adaptation in humans, it is unlikely that a latent factor will separate high-altitude populations (Tibet, Andes, Ethiopia) from lowland populations. An approach incorporating an environmental variable, that is, a fixed factor, corresponding to altitude would be more appropriate (Foll et al. 2014). The fifth caveat concerns the choice of the threshold for the Bayes factor. When analyzing the HGDP data, we use a threshold of 10 which corresponds to strong evidence for the outlier model according to the Jeffrey's scale of evidence for Bayes factors, but other choices would have also been possible (Kass and Raftery 1995). The last caveat concerns the inability to pinpoint the geographic location at which adaptation took place. Although outlier SNPs are related to latent factors, the factor model does not provide the population or group of individuals concerned by adaptation, whereas alternative statistical models can estimate selection coefficients for each population separately (Vitalis et al. 2014).

Although the purpose of our article is to present a novel method and not to provide an in-depth analysis of local adaptation in the HGDP, fitting the factor model to the HGDP data provides interesting results. The first two latent factors mainly measure differentiation between Africa, Western Eurasia, and East Asia and some outliers related with these two factors are involved in morphological traits, which are often reported to be enriched with genes having signatures of positive selection (Barreiro et al. 2008). In the list of outliers, we found the genes *SLC24A5* and *EDAR*, which are often reported as top hits in selection scans (Pickrell et al. 2009; Hao et al. 2013) and are related to skin pigmentation and hair thickness, respectively (Kamberov et al. 2013; Mallick et al. 2013). We also found that SNPs close to the oncogene *CYP26A1*, whose expression is enhanced in sunlight-damaged human skin (Osanai and Lee 2011), are part of the top list for outliers. The third and fourth factors correspond to genetic differentiation between Native Americans, individuals from Oceania, and the rest of the sample. There are many regions in chromosome 8 enriched with outlier SNPs. One of these regions encompasses the gene *RP1L1* that is associated with retinal diseases and which has already been reported to have one of the strongest signatures of positive selection along with other genes related to sensory functions (Barreiro et al. 2008). Many outlier SNPs strongly related to the third and fourth factors have allele frequencies that are similar between Southern Native Americans and Pygmies (third factor) or between individuals from Oceania, Southern Native Americans, and Pygmies (fourth factor). Because these individuals all live in tropical rain forests and have similar diet consisting of roots and tubers, our findings support the importance of diet, climate, and potentially pathogen load to explain human adaptation (Hancock et al. 2010; Fumagalli et al. 2011). The SNPs with similar allele frequencies in different geographic regions are good candidates for convergent evolution and would deserve further analysis.

Factor models are enriching the toolbox of population genetic methods. The main principle is to model population structure via latent variables called factors. Factors models have already been proposed to ascertain population structure (Engelhardt and Stephens 2010) and to account for population structure when testing for gene-environment association (Frichot et al. 2013). We showed that factor models also provide a convenient individual-based framework to find loci that have atypical patterns of genetic differentiation. A major argument supporting the proposed hierarchical factor model is that it produces fewer false discoveries compared with genome scans based on F_{ST}.

Materials and Methods

Hierarchical Bayesian Modeling

We provide the prior distributions for the latent variables of the hierarchical factor model defined by equations (2)–(4). To account for linkage disequilibrium in the genome and encourage outlier loci to be clustered along the genome, we consider a Potts model with an external field (Winkler 2003)

$$p(z_1,\ldots,z_p) \propto (1-\pi)^{p_0}\pi^{(p-p_0)}e^{\beta\sum_{i\sim j}1_{z_i=z_j}}, \qquad (5)$$

where the sum in the exponential ranges over all pairs of neighboring loci. We consider that each locus has two neighbors except at the beginning and at the end of the chromosome where a locus has only one neighbor. In equation (5), the variable p_0 is the number of loci such that $z_i = 0$, 1 is the indicator function, β is the parameter of the Potts model and is set to $\beta = 1$, and π is the prior proportion of outliers. To model the proportion of outliers, we consider a uniform prior on the $\log_{10}$ scale reflecting that we are interested in the order of magnitude of the proportion of outlier loci (Guan and Stephens 2011). In the following, we consider -4 and -1 for the lower and upper bound of the uniform prior, respectively. The proportion of loci under selection is a priori expected to be between 10^{-1} and 10^{-4} reflecting the working hypothesis that most loci are neutral and that loci under selection are rare. For the variance parameters σ_k^2, $k = 1,\ldots,K$, that are specific to each factor (eqs. 3 and 4), we consider the parameterization $\sigma_k^2 = \sigma^2\rho_k^2$, where σ^2 is the residual variance in equation (2) (Oba et al. 2003). We consider the non-informative prior for the variance parameters $p(\sigma^2) \propto 1/\sigma^2$ and $p(\rho_k^2) \propto 1/\rho_k^2$, $k = 1,\ldots,K$. For the variance–inflation parameters c_k^2 of equation (4), we consider uniform priors with 1 and 10 for the lower and upper bounds, and we find that a slight variation of the upper bound does not change the ranking of the SNPs (results not shown).

With the factor model of equation (2) as with many models with latent structure (Allman et al. 2009), there is a well-known issue of identifiability because identical likelihood values can be obtained from a solution (**U**,**V**) after using orthogonal rotations (West 2003). To add constraints to the model, we consider a prior with unit variance for each of the factors

$$\mathbf{U}_k \sim \mathcal{N}(0,\mathbf{I}_\mathbf{n}),$$

where $\mathbf{I}_\mathbf{n}$ is the squared $n \times n$ identity matrix (Oba et al. 2003). To further prevent the MCMC algorithm to produce alternative rotations of the factors (Engelhardt and Stephens 2010), we consider the solution of the singular value decomposition as starting values for the factors $\mathbf{U}_1,\ldots,\mathbf{U}_K$ in the MCMC algorithm.

To evaluate the strength of evidence for outlyingness at each locus, we compute the Bayes factor on a $\log_{10}$ scale. The Bayes factor is defined as the ratio of the posterior odds $p(z_\ell > 0\,|\,\mathbf{Y})/p(z_\ell = 0\,|\,\mathbf{Y})$ and the prior odds

$p(z_\ell > 0)/p(z_\ell = 0) = \pi/(1-\pi)$. The description of the MCMC algorithm and the computation of the Bayes factor is given in the Supplementary Material online.

Simulation of the Four-Population Divergence Model

The first simulation scenario is a divergence model with four populations. These populations have constant effective population sizes of $N_e = 1{,}000$ diploid individuals, with 50 individuals sampled in each population. The genotypes consist of 10,000 independent SNPs. The simulations are performed in two steps. In the first step, we use the software *ms* to simulate a neutral divergence model (Hudson 2002). When looking backward in time, we instantly merge population A_1 with A_2 and population B_1 with B_2, then after waiting a number $T = 20$, 80, 120, 160, 200 of generations, we merge the two remaining populations *A* and *B*. We keep only variants with a minor allele frequency larger than 5% at the end of this first step. The second step is performed with the software *SimuPOP* (Peng and Kimmel 2005). To run *SimuPOP*, we provide the allele frequencies in each of the four populations that have been generated with *ms*. Looking forward in time, we simulate 100 generations after the 2 concomitant divergence events. We assume no migration between populations. In each evolutionary lineage, we assume that 100 SNPs confer selective advantage using a selection coefficient of $s = 0.1$ for homozygotes carrying two adaptive alleles. In both simulation schemes, we assume an additive model for selection.

Simulation of the Stepping-Stone Model

The second simulation scenario is a two-dimensional stepping-stone model with a 10 × 10 grid. Each of the 100 populations has an effective population size of $N_e = 1{,}000$ diploid individuals. We sample ten individuals in each population and there are 2,050 independent SNPs. We also consider a two-step procedure for the simulations. First, we simulate an equilibrium stepping-stone model with the software *ms*. Neighboring populations exchange migrants with a rate of $4N_e m = 8$ per generation. Then we superimpose a selection gradient using *SimuPOP*. During 100 generations, we consider that 50 SNPs confer selective advantage. The selection coefficient $s = 0.1$ is maximal in population 64, which is located in the lower-right quarter of the grid. In the eight neighboring populations, the selection coefficient is $s = 0.05$, and in the second layer of neighbors, the selection coefficient is $s = 0.025$. The selection coefficient is equal to 0 for the rest of the grid.

False Discovery Rate and Sensitivity

To compare the performances of the different methods for selection scans, we compute the FDR for fixed values of the sensitivity. Denoting the number of false positives by FP, the number of false negatives by FN, and the number of true positives by TP, the FDR is defined as FP/(FP + TP) and the sensitivity (also called recall rate) is defined as TP/(TP + FN).

Gene Ontology Analysis

We perform a GO enrichment analysis with the software *Gowinda* (Kofler and Schlötterer 2012). The list of genes is built by considering all genes which contain outlier SNPs with a tolerance of 5,000 bp upstream and downstream. We use a threshold of 0.05 for the FDR, and we remove GO terms that are shared by less than 10 genes or more than 1,000 genes. We consider the *–snp* flag in *Gowinda* that assumes independence of SNPs within a gene. If we rather use the *–gene* flag, which assumes complete dependence of SNPs within a gene, there is no GO discovery with an FDR smaller than 5%. For each factor, we consider the ten GO terms with the smallest FDRs, and we only report GO terms that are related to biological processes.

Software Availability

The computer program PCAdapt for fitting the factor model is available from the authors' websites (http://membres-timc.imag.fr/Michael.Blum/, http://membres-timc.imag.fr/Nicolas.Duforet-Frebourg/, last accessed June 12, 2014).

Supplementary Material

Supplementary material, figures S1–S9, and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

This work has been supported by the French National Research Agency (DATGEN project, ANR-2010-JCJC-1607-01) and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

References

Abdellaoui A, Hottenga J-J, de Knijff P, Nivard MG, Xiao X, Scheet P, Brooks A, Ehli EA, Hu Y, Davies GE, et al. 2013. Population structure, migration, and diversifying selection in the Netherlands. *Eur J Hum Genet.* 21:1277–1285.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* 12:1805–1814.

Allman ES, Matias C, Rhodes JA. 2009. Identifiability of parameters in latent structure models with many observed variables. *Ann Stat.* 37: 3099–3132.

Antao T, Lopes A, Lopes R, Beja-Pereira A, Luikart G. 2008. Lositan: a workbench to detect molecular adaptation based on a F_{ST}-outlier method. *BMC Bioinformatics* 9:323.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet.* 40:340–345.

Bazin E, Dawson KJ, Beaumont MA. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* 185:587–602.

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol.* 13:969–980.

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London B Biol Sci.* 263: 1619–1626.

Bierne N, Roze D, Welch JJ. 2013. Pervasive selection or is it. . .? Why are FST outliers sometimes so frequent? *Mol Ecol.* 22 2061–2064.

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

Bonin A, Taberlet P, Miaud C, Pompanon F. 2006. Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol Biol Evol.* 23: 773–783.

Box GE, Tiao GC. 1968. A Bayesian approach to some outlier problems. *Biometrika* 55:119–129.

Canfield VA, Berg A, Peckins S, Wentzel SM, Ang KC, Oppenheimer S, Cheng KC. 2013. Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3 (Bethesda)* 3:2059–2067.

Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. 2008. High-dimensional sparse factor modeling: applications in gene expression genomics. *J Am Stat Assoc.* 103:1438–1456.

Davidson AE, Sergouniotis PI, Mackay DS, Wright GA, Waseem NH, Michaelides M, Holder GE, Robson AG, Moore AT, Plagnol V, et al. 2013. RP1L1 variants are associated with a spectrum of inherited retinal diseases including retinitis pigmentosa and occult macular dystrophy. *Hum Mutat.* 34:506–514.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004.

Engelhardt BE, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* 6:e1001117.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* 14:2611–2620.

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193:929–941.

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

Foll M, Gaggiotti OE, Daub JT, Excoffier L. 2014. Hierarchical Bayesian model of population structure reveals convergent adaptation to high altitude in human populations. arXiv preprint arXiv:1402.4348.

Fourcade Y, Chaput-Bardy A, Secondi J, Fleurant C, Lemaire C. 2013. Is local selection so widespread in river organisms? Fractal geometry of river networks leads to high bias in outlier detection. *Mol Ecol.* 8: 2065–2073.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol.* 30:1687–1699.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admettla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.

Gompert Z, Buerkle CA. 2011. A hierarchical Bayesian model for next-generation population genomics. *Genetics* 187:903–917.

Guan Y, Stephens M. 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat.* 5:1780–1815.

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

Guo F, Dey DK, Holsinger KE. 2009. A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *J Am Stat Assoc.* 104:142–154.

Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G, et al. 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A.* 107:8924–8930.

Hao W, Song M, Storey JD. 2013. Probabilistic models of genetic variation in structured populations applied to global human studies. arXiv preprint arXiv:1312.2041.

Hilborn R, Mangel M. 1997. The ecological detective: confronting models with data, Vol. 28. Princeton (NJ): Princeton University Press.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, Froment A, Bodo J-M, Beggs W, Hoffman G, et al. 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet.* 8:e1002641.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al. 2013. Modeling recent human evolution in mice by expression of a selected *EDAR* variant. *Cell* 152:691–702.

Kass RE, Raftery AE. 1995. Bayes factors. *J Am Stat Assoc.* 90:773–795.

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

Laloë D, Gautier M. 2011. On the genetic interpretation of between-group PCA on SNP data. HAL hal-00661214.

Lange K, Papp JC, Sinsheimer JS, Sobel EM. 2014. Next-generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. *Ann Rev Stat Appl.* 1:279–300.

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

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet.* 4:981–994.

Mallick CB, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SY, Romero IG, Crivellaro F, et al. 2013. The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet.* 9:e1003912.

Manel S, Holderegger R. 2013. Ten years of landscape genetics. *Trends Ecol Evol.* 28:614–621.

Meirmans PG, Hedrick PW. 2011. Assessing population structure: Fst and related measures. *Mol Ecol Resour.* 11:5–18.

Nicholson G, Smith AV, Jónsson F, Gústafsson.Ó, Stefánsson K, Donnelly P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J R Stat Soc Series B Stat Methodol.* 64:695–715.

Nosil P, Buerkle A. 2010. Population genomics. *Nat Educ Knowl.* 3:8.

Nosil P, Egan SP, Funk DJ. 2008. Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. *Evolution* 62:316–336.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet.* 40:646–649.

Oba S, Sato M, Takemasa I, Monden M, Matsubara K, Ishii S. 2003. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19:2088–2096.

Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci.* 365: 185–205.

Osanai M, Lee G-H. 2011. Enhanced expression of retinoic acid-metabolizing enzyme *CYP26A1* in sunlight-damaged human skin. *Med Mol Morphol.* 44:200–206.

Patterson N, Price AL, Reich D. 2006. Population structure and eigen analysis. *PLoS Genet.* 2:e190.

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

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19:826–837.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81:559–575.

Rajeevan H, Soundararajan U, Kidd JR, Pakstis AJ, Kidd KK. 2012. ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Res.* 40:D1010–D1015.

Riebler A, Held L, Stephan W. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178:1817–1829.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the drosophila genome? *PLoS Genet.* 5:e1000495.

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

Vitalis R, Dawson K, Boursot P, Belkhir K. 2003. DetSel 1.0: a computer program to detect markers responding to selection. *J Hered.* 94:429–431.

Vitalis R, Gautier M, Dawson KJ, Beaumont MA. 2014. Detecting and measuring selection from gene frequency data. *Genetics* 196:799–817.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42:D1001–D1006.

West M. 2003. Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian Stat.* 7:723–732.

Winkler G. 2003. Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction. Vol. 27. Berlin: Springer.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.

Xu S, Li S, Yang Y, Tan J, Lou H, Jin W, Yang L, Pan X, Wang J, Shen Y, et al. 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol.* 28:1003–1011.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen JE, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.