Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models

Eric Frichot,[1] Sean D. Schoville,[1] Guillaume Bouchard,[2] and Olivier François*,[1]

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

[2]Xerox Research Center Europe, Meylan, France

***Corresponding author:** E-mail: olivier.francois@imag.fr.

Associate editor: Asger Hobolth

Abstract

Adaptation to local environments often occurs through natural selection acting on a large number of loci, each having a weak phenotypic effect. One way to detect these loci is to identify genetic polymorphisms that exhibit high correlation with environmental variables used as proxies for ecological pressures. Here, we propose new algorithms based on population genetics, ecological modeling, and statistical learning techniques to screen genomes for signatures of local adaptation. Implemented in the computer program "latent factor mixed model" (LFMM), these algorithms employ an approach in which population structure is introduced using unobserved variables. These fast and computationally efficient algorithms detect correlations between environmental and genetic variation while simultaneously inferring background levels of population structure. Comparing these new algorithms with related methods provides evidence that LFMM can efficiently estimate random effects due to population history and isolation-by-distance patterns when computing gene-environment correlations, and decrease the number of false-positive associations in genome scans. We then apply these models to plant and human genetic data, identifying several genes with functions related to development that exhibit strong correlations with climatic gradients.

Key words: **local adaptation, environmental correlations, genome scans, latent factor models, population structure.**

Introduction

Local adaptation through natural selection plays a central role in shaping the variation of natural populations (Darwin 1859; Williams 1966) and is of fundamental importance in evolution, conservation, and global-change biology (Joost et al. 2007; Manel et al. 2010; Barrett and Hoekstra 2011; Jay et al. 2012; Schoville et al. 2012). The intensity of natural selection commonly varies in space and can result in gene-environment interactions that have measurable effects on fitness (Storz and Wheat 2010). This can lead to local adaptation if populations maintain locally advantageous traits despite gene flow with neighboring populations.

In principle, identifying chromosomal regions involved in adaptive divergence can be achieved by scanning genome-wide patterns of DNA polymorphism (Nielsen 2005; Storz 2005). Usually, the aim of screening procedures is to detect locus-specific signatures of positive selection. In populations inhabiting spatially distinct environments, loci that underlie adaptive divergence can be detected by comparing relative levels of differentiation among large samples of unlinked markers (Beaumont and Nichols 1996; Beaumont and Balding 2004) and by using empirical tests to compare levels of differentiation with the genomic background (Kelley et al. 2006; Akey 2009; Novembre and Di Rienzo 2009).

An alternative way to investigate signatures of local adaptation, especially when beneficial alleles have weak phenotypic effects, is by identifying polymorphisms that exhibit high correlation with environmental variables (Joost et al. 2007; Hancock et al. 2008; Coop et al. 2010; Poncet et al. 2010; Pritchard et al. 2010). In natural populations, quantitative traits that exhibit continuous geographic variation are often associated with specific ecological variables reflecting selective pressures acting on individual phenotypes (Endler 1977). This type of variation is then reflected in geographic clines or in sympatric populations that exploit different ecological niches (Haldane 1948; Berry and Kreitman 1993; Prugnolle et al. 2005; Young et al. 2005). Evidence for local adaptation to continuous environments can be detected if there is highly significant association with the environmental variables at some loci compared with the background genomic variation.

A major difficulty is that the geographical basis of both environmental and genetic variation can confound interpretation of the tests (Eckert et al. 2010), as local adaptation can be hindered by gene flow (Lenormand 2002), and can be difficult to distinguish from the effects of genetic drift and demographic history (Novembre and Di Rienzo 2009). The main problem is that without corrections for the effect of population structure or isolation-by-distance (IBD), the underlying null distribution may be insufficient to account for the demographic history of the study organism. As a result, tests for associations between loci and environmental variables using classical regression models will be prone to high rates of false positives (FP) (Meirmans 2012). Recent studies have used the background patterns of allele frequencies to build a null model that accounts for the effects of drift and demographic history (Hancock et al. 2008; Coop et al. 2010; Fumagalli et al. 2011; Hancock et al. 2011). To correct for

Article

population stratification, Hancock et al. (2008) used an empirical approach that estimates the covariance of allele frequencies among populations. These authors assessed the evidence for local adaptation of each allele by testing whether environmental variables explained more variance than a null model with this particular covariance structure.

A drawback of empirical tests is the need to identify selectively neutral loci from the genomic background before testing for associations with environmental factors. The need to identify such a list a priori implies that tests based on empirical estimates of relatedness can lack power to reject neutrality, which is an important limitation for data sets where all loci are potentially under selection. For example, single nucleotide polymorphism (SNP) data sets derived from expressed sequences are often used to study local adaptation in nonmodel organisms (Eckert et al. 2010). Choosing a subset of markers not only reduces the size of such a data set but could also arbitrarily bias downstream statistical tests if only certain subsets of data (e.g., synonymous sites) are chosen as the neutral markers. After all, putatively neutral sites can be linked to loci under selection over large physical distances (Thibert-Plante and Hendry 2010). In this study, we address this problem by introducing statistical models called latent factor mixed models (LFMM).

Using these models, we test correlations between environmental and genetic variation while estimating the effects of hidden factors that represent background residual levels of population structure. To perform parameter estimation, we extend probabilistic principal component analysis (PCA) and recent statistical learning approaches (Tipping and Bishop 1999; Salakhutdinov and Mnih 2008; Engelhardt and Stephens 2010; Frichot et al. 2012). Based on low rank approximation of the residual covariance matrix, we implement algorithms to deal with hundreds of thousands of polymorphisms with rapid computing times. We show that our algorithms control for random effects due to population history and spatial autocorrelation when estimating gene-environment association, and we provide examples of how our approach can be used to detect local adaptation in plants and humans.

New Approaches

Consider the data matrix, $(G_{i\ell})$, where each entry records the allele frequency for individual i at the genomic locus ℓ, $1 \leq i \leq n$, $1 \leq \ell \leq L$, and n and L represent the total sample size and number of loci, respectively. For simplicity, we assume our loci are bi-allelic, for example, SNPs. In this case, for each marker, there is an ancestral and a derived allele, and $G_{i\ell}$ is the number of derived alleles for locus ℓ and individual i. For diploid data, $G_{i\ell}$ is thus equal to 0, 1, or 2, and corresponds to the genotype at locus ℓ. In addition to the genotypic data, we have a vector of d geographic and environmental variables, (X_i), for each individual. The vector of covariates could include latitude and longitude, habitat and other ecological information, climatic variables, and so forth, which serve as proxies for unknown environmental pressures (Hancock et al. 2008; Eckert et al. 2010).

Model

To evaluate associations between allele frequencies and environmental variables while correcting for background levels of population structure, we regard the matrix G as being a response variable in a regression mixed model

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell}\,, \tag{1}$$

where μ_ℓ is a locus specific effect, β_ℓ is a d-dimensional vector of regression coefficients, U_i and V_ℓ are scalar vectors with K dimensions that model latent factors and their scores $(1 \leq K \leq n)$. The residuals $\epsilon_{i\ell}$ are statistically independent Gaussian variables of mean zero and variance σ^2.

We refer to the earlier-mentioned statistical model as a LFMM (see Materials and Methods). Similar models, termed factor regression models, have been considered earlier in biostatistics in the inference of molecular pathways from gene expression data (West 2003; Carvalho et al. 2008).

In LFMMs, environmental variables are introduced as fixed effects while population structure is modeled using latent factors. In the model, the matrix term $U^T V$ models the part of genetic variation that cannot be explained by the environmental pressures. Note that the use of factorization methods is closely related to estimating population structure by singular value decomposition, a well-established technique for identifying scores and loadings in PCA (Jolliffe 1986). Recently, matrix factorization methods have been generalized to include probabilistic PCA (Tipping and Bishop 1999) and probabilistic matrix factorization algorithms (Salakhutdinov and Mnih 2008), which have proven useful in analyzing population genetic data (Engelhardt and Stephens 2010). To clarify the connection between LFMM and PCA, assume that no environmental variable is available. In this case, we set $\beta_\ell = 0$ for all locus ℓ. In matrix factorization algorithms, a data matrix G with n rows and L columns can be decomposed into a product of two matrices U and V, where U has n rows and K columns, and V is a $K \times L$ matrix. Following Patterson et al. (2006), we assume that the genotypic data are centered. We consider the matrix $Y_{i\ell} = G_{i\ell} - \bar{G}_{\cdot\ell}$, where we have subtracted the mean value of each column, $\bar{G}_{\cdot\ell} = \sum_{i=1}^{n} G_{i\ell}/n$. For each individual i and locus ℓ, the decomposition is as follows:

$$Y_{i\ell} = U_i^T V_\ell = \sum_{k=1}^{K} U_{ik} V_{k\ell}\,. \tag{2}$$

To estimate the factor vectors U_i and V_ℓ, the squared error is minimized on the set of observed data

$$\min_{U,V} \sum_{k=1}^{K} (Y_{i\ell} - U_{ik} V_{k\ell})^2\,. \tag{3}$$

With $K = L$, this approach is similar to computing PCA loadings and scores (Jolliffe 1986). The number of components K can, however, be chosen much lower than the number of loci or individuals. In simulations, we based our choice of K on Tracy–Widom theory (Patterson et al. 2006). In real applications, this choice of K may be replaced by

estimates of population genetic structure obtained with clustering algorithms like STRUCTURE (Pritchard et al. 2000) or TESS (Chen et al. 2007). When values of K are low our algorithm is essentially a low-rank approximation of the covariance structure (Eckart and Young 1936), which leads to computationally fast estimation algorithms. To estimate the LFMM parameters, we implemented a Gibbs sampler algorithm (Materials and Methods and supplementary file S1, Supplementary Material online). We computed $|z|$-scores for all environmental effects, and we tested the significance of these effects using the standard Gaussian distribution and Bonferroni correction for multiple testing.

Incorporating population genetic structure using estimates of principal components or ancestry coefficients is common in genome-wide association studies (Price et al. 2006; Yu et al. 2006; Zhou and Stephens 2012), and in tests based on empirical approaches (Coop et al. 2010; Poncet et al. 2010). In this paragraph, we explain the distinction between LFMM and tests based on empirical covariance matrices. Suppose that we start by computing PCA scores from the matrix Y for all individuals, and denote by $\tilde{U}_i$ the PCA scores for individual i. The product matrix $\tilde{U}\tilde{U}^T$ is thus equal to the empirical covariance matrix

$$\tilde{U}\tilde{U}^T = YY^T/n. \quad (4)$$

Now using the scores as covariates in a Bayesian regression model, we obtain

$$G = \mu + \beta^T X + \tilde{U}^T V + \epsilon. \quad (5)$$

By a change of variables, this is equivalent to fitting the model

$$G = \mu + \beta^T X + \tilde{\epsilon}, \quad (6)$$

where the distribution of $\tilde{\epsilon}$ is a multivariate Gaussian distribution of the covariance matrix equal to $\sigma^2 \mathrm{Id} + \sigma_V^2 YY^T/n$. Here, Id is the n-dimensional identity matrix, and σ_V^2 is the variance of factor coordinates. Setting $\sigma_V^2 = 1$ and considering small values of the scaling parameter σ^2, the model defined in equation (6) is nearly equivalent to the model implemented in empirical approaches. In a Bayesian Gaussian regression framework, incorporating PCA scores as covariates in an association model is equivalent to modeling residuals as Gaussian vectors with covariance depending on the empirical covariance matrix of the genotypic data. Thus, a major difference between methods is that the factor matrix U and the regression coefficients β are estimated by a two-stage procedure in empirical approaches, whereas it requires a single step in LFMMs.

Results

We designed experiments based on simulated data to answer the following questions: 1) Are tests based on LFMMs conservative or liberal? 2) How does the LFMM algorithm perform compared with existing methods such as logistic or standard regression models (Joost et al. 2007), principal component regression model (PCRM), partial Mantel tests (PMTs) (Fumagalli et al. 2011), standard linear mixed models (Zhou and Stephens 2012), and Bayesian mixed models (Coop et al. 2010)?

Distribution of P Values under the Null Hypothesis

To evaluate the calibration of P values, we used equation (1) with $\beta = 0$ to simulate data under a null hypothesis of no association with any environmental variable (Materials and Methods). Figure 1 reports the empirical cumulative distribution function (ECDF) of P values for $K = 5$ and $K = 20$. Plots for other values of K are shown in supplementary figure S1, Supplementary Material online. P values are well calibrated when their ECDF is close to the uniform distribution, represented by the bisector line. Below the line, the test is conservative. For values of K less than 5, the ECDF was close to a uniform distribution, and P values were correctly calibrated. For $K = 20$, the tests were slightly conservative. Thus, for moderate and for large values of the number of latent factors, the tests produced small numbers of FP associations.

Next, we used equation (1) to simulate data exhibiting various levels of population structure and association with a randomly generated environmental variable, and we compared the distributions of statistical errors for the following three estimation approaches: 1) LFMM, 2) a standard linear regression model, and 3) a PC regression model (Materials and Methods).

Figure 2 reports the quantiles of absolute errors for LFMM, the standard linear regression, and PC regression models. For LFMM, absolute errors ranged between 0 and 0.6 for $K = 2 - 20$, and between 0 and 1.0 for $K = 100$. Mean squared errors indicated that the bias and variance of estimates were small (table 1). Compared with LFMMs, the relative errors of the linear and PC regression estimates increased with the rank of the hidden factor matrix. The absolute errors of these algorithms ranged between 0 and 1.4 for $K = 2$, between 0 and 3.2 for $K = 20$, and between 0 and 9.2 for $K = 100$. When linear or PC regression models were fitted to the data, the quantiles of errors shifted to values $\approx$ 1.74-fold higher for $K = 2$, $\approx$ 3.8- to 4.1-fold higher for $K = 20$, and $\approx$ 5.5- to 7.7-fold higher for $K = 100$. Mean squared errors provided additional evidence of relatively poor performances of the linear regression and PC regression estimates when the levels of underlying structure increased (table 1).

Spatial Coalescent Simulations

In another series of experiments, we compared the LFMM estimation algorithm against two methods that do not correct for population stratification, and against methods that use the empirical covariance matrix to correct for population stratification. The first set of methods include a linear model (LM) and generalized linear model (GLM) (Joost et al. 2007), and the second set of methods included three empirical methods: a PCRM, PMTs (Smouse et al. 1986; Legendre and Legendre 2012), and the mixed models implemented in BAYENV and GEMMA (Coop et al. 2010; Zhou and Stephens 2012). In PMTs, the relationship between population genetic distances at each SNP and a matrix of environmental variable

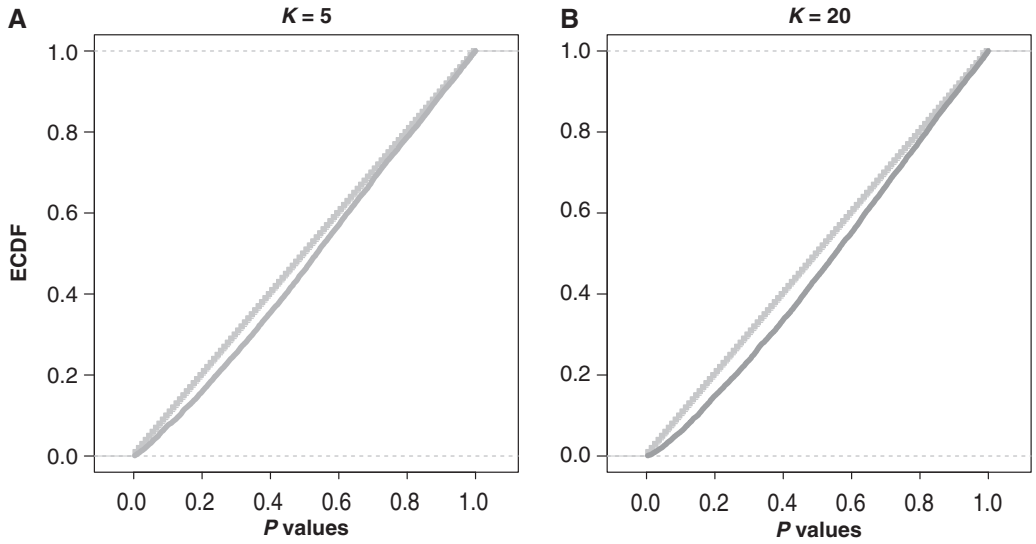

Fig. 1. Simulations from the null model. ECDF of *P* values for LFMM tests for simulations from a latent factor model using (*A*) $K = 5$ and (*B*) $K = 20$ latent factors.

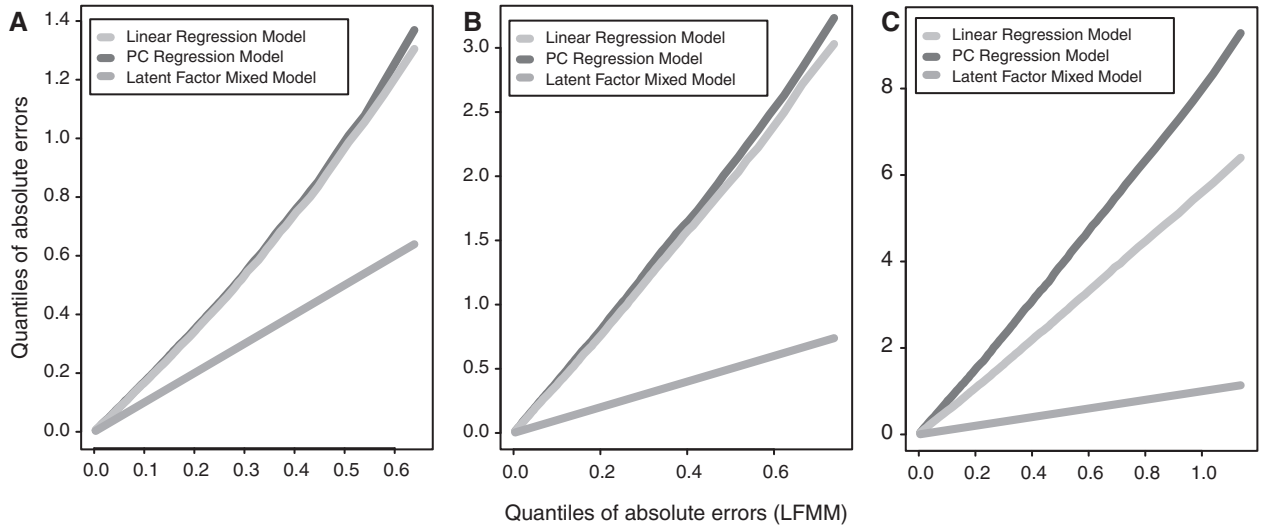

Fig. 2. Generative model simulations. Quantiles of absolute errors for the standard linear regression, PC regression, and LFM models using simulations from latent factor models using (*A*) $K = 2$, (*B*) $K = 20$, and (*C*) $K = 100$ latent factors.

Table 1. Mean Squared Errors for Estimates of Environmental Effects.

K	LM	PCRM	LFMM
2	0.20	0.21	0.15
20	1.27	1.42	0.08
100	6.13	12.41	0.20

distance was evaluated using a correction for correlations in genome-wide allele frequencies. With GEMMA, we implemented a standard linear mixed model in which a single environmental variable is explained by SNP genotype, and where relatedness is introduced by a random effect (see Materials and Methods for a description of all methods).

To examine the outcome of tests when genetic variation is neutral at all loci, we computed the distributions of *P* values under LM, GLM, PMT, PCRM, GEMMA, and LFMM with different values for the number of latent factors (*K* ranging from 1 to 20). The distributions of *P* values for tests based on LM and GLM showed a strong departure from the uniform distribution (fig. 3*A* and *B*). In those cases, the tests were too liberal and produced a large number of FP results. For GLM, using population allele frequencies instead of individual genotypes reduced the number of FP associations but the tests based on these models remained liberal. The ECDF for PMTs showed an excess of low and high *P* values, but the curve was closer to a uniform distribution than with LM tests (fig. 3*C*). Using $K = 7$ PCs in PCRM, *P* values for those tests and for GEMMA were well calibrated (fig. 3*D*–*F*). Choosing *K* based on Tracy–Widom theory ($K = 7$) and on Bayesian clustering algorithms ($K = 5$) led to slightly conservative tests for LFMMs (fig. 3*E*). ECDFs for all values of *K* are shown in supplementary figures S2 and S3, Supplementary Material online, respectively.

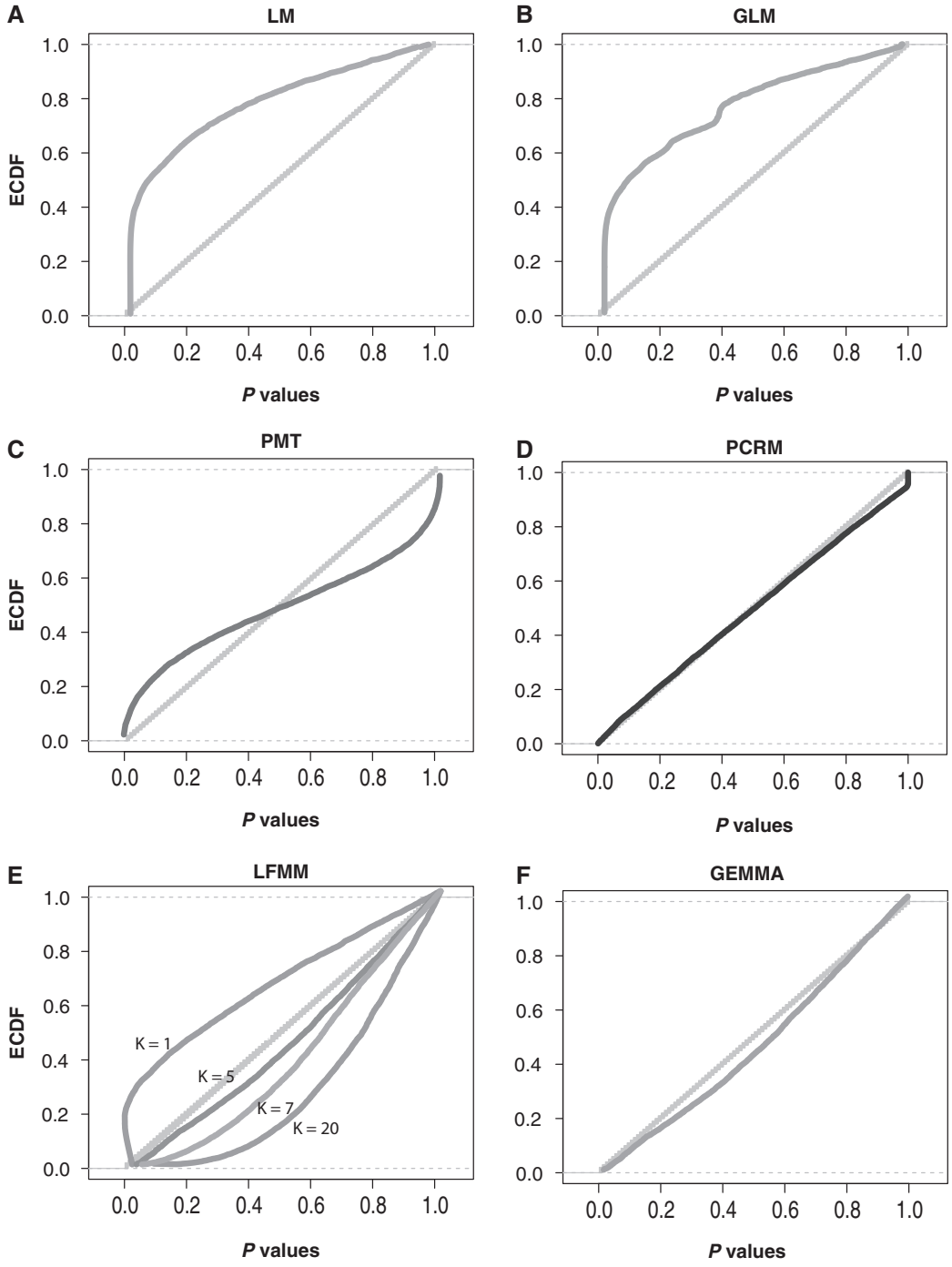

Fig. 3. Spatial neutral coalescent simulations. ECDF of *P* values for (*A*) the linear regression model (LM), (*B*) the GLM, (*C*) PMTs using Nei's genetic distance and the empirical correlation matrix for correction, (*D*) the PC regression model using $K = 7$ principal components (PCRM), (*E*) the LFM model using $K = 1$, 5, 7, and 20 latent factors (LFMM) where the value $K = 5$ corresponds to the estimate of the number of clusters obtained from Bayesian clustering algorithms, and the value $K = 7$ is a Tracy–Widom estimate, and (*F*) the standard linear mixed model implemented in GEMMA.

Next, we evaluated the ability of LFMMs to detect loci exhibiting correlations with particular environmental gradients and compared tests based on LFMMs with methods based on linear models. An environmental variable, *x*, was defined for each population as the geographic identifier of the population in the linear stepping-stone model. Following Haldane (1948), we chose a sigmoid function to represent the shape of a selected allele frequency cline through geographic space. Under strong selection ($\theta = 0.2$), we expect that tests produce low rates of FP associations while still preserving reasonable power.

For all simulated data sets, we evaluated the rates of false negative (FN) and of FP tests based on LM, GLM, PMT, PCRM, GEMMA, and LFMM for two values of the type I error (table 2). In the case of strong selection, we found that standard linear models exhibited high rates of FP. In contrast, tests that include corrections for population structure—based on PMTs, PCRM, and standard linear mixed

Table 2. Rates of FN and FP Association for Tests Based on LM, PCRM, Standard Linear Mixed Models (GEMMA), PMT Correlations, and LFMM.

FN (FP)	LM	GLM	PCRM	GEMMA	PMT	LFMM
Type I error						
$\alpha = 0.001$	0% (33%)	0% (24%)	100% (3%)	100% (2%)	99% (6.8%)	4% (5%)
$\alpha = 0.0001$	0% (27%)	0% (19%)	100% (0%)	100% (0%)	100% (3.4%)	14% (3%)

models—exhibited low rates of FP. But PMT, PCRM, and GEMMA exhibited large rates of FN, and these tests had low power to reject neutrality. These results provide evidence that the standard methods may fail to identify selected loci from the genomic background even though association with the environment is strong. In this context, tests based on LFMM produced low rates of FP and had reasonable power to reject neutrality (table 2).

To perform comparisons with the program BAYENV (Coop et al. 2010), we wanted to evaluate whether the program was able to detect weak selection. Thus, we set the intensity of selection through space to a low level ($\theta = 0.1$, Materials and Methods). As BAYENV returns Bayes factors instead of *P* values, we considered ranked lists recording the *M* loci corresponding to the strongest associations (*M* between 1 and $L = 1{,}050$). Figure 4 reports the number of true positives (TP) as a function of the number of FP. Considering the rates of TP and FP, the mean area under the receiver-operating characteristic curve (AUC) for tests based on LFMMs with $K = 5$–7 factors were approximately 0.95–0.96, whereas the AUC for BAYENV was equal to 0.88. In the linear stepping stone model simulations, the tests based on LFMM performed better than BAYENV for all values of *K* (fig. 4).

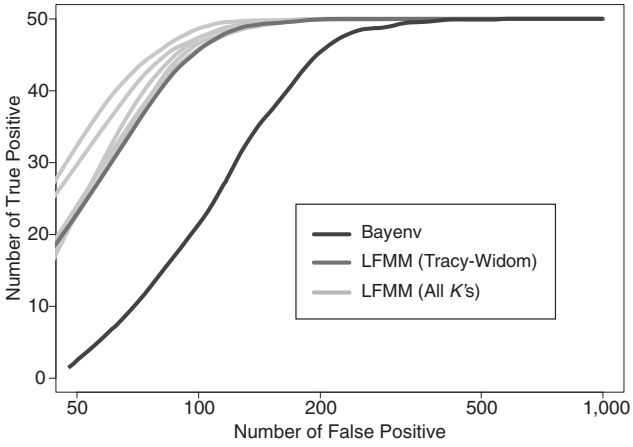

Fig. 4. Spatial coalescent simulations with loci under selection. Number of true positive associations for BAYENV and for LFMM for $K = 5{,}7$ (STRUCTURE and Tracy–Widom values) and for $K = 1$, 3, 10, and 20 for spatial coalescent simulations including 1,050 loci with 50 SNPs under selection.

Loblolly Pine

To illustrate the application of LFMMs, we analyzed genomic data of Loblolly pines (*Pinus taeda*, Pinaceae, Eckert et al. 2010). The Loblolly pine is distributed throughout the Southeastern United States, ranging from the arid Great Plains to the humid Eastern Temperate Forest ecoregion. These data consisted of 1,730 SNPs selected in expressed sequence tags (ESTs) for 682 individuals (Eckert et al. 2010).

We applied LFMM to the Loblolly pine data, testing 5 environmental variables representing the 5 first components of a PCA for 60 climatic variables (data from Eckert et al. 2010). A total of 392, 113, and 30 SNPs obtained $|z|$-scores greater than 3, 4, or 5 for at least one environmental variable, respectively. On the basis this result, we considered that a SNP effect was significant when its $|z|$-score was greater than 4 (two-sided test). The cutoff $|z| > 4$ corresponds to *P* values $P < 10^{-5}$ obtained after applying a Bonferroni correction for a type I error $\alpha = 0.01$ and $L \approx 10^3$ loci. Among the 50 loci with the highest $|z|$-scores, 17 were shared with those detected by Eckert et al. (2010) using BAYENV. Seven of the 10 SNPs with Bayes factors greater than 10^3 were confirmed by the LFMM analysis. For the first and second environmental variables, the two SNPs which obtained the highest Bayes factors using BAYENV were recovered by the LFMM analysis. Table 3 provides a list of SNPs associated with climatic gradients and their functional annotation. The LFMM analysis discovered new significant and interesting associations with climatic gradients not identified in the analysis of Eckert et al. (2010), such as the chloroplast lumen 19 kDA protein involved in photosynthesis ($|z| = 6.42$), a pentatricopeptide repeat protein involved in oxidative stress and salt stress ($|z| = 5.90$), and the heat shock transcription factor hsf5 ($|z| = 5.60$) involved in regulation of transcription and response to temperature stress (table 3 and supplementary table S1, Supplementary Material online).

Human Data Analysis

We applied LFMM to a worldwide sample of genomic DNA from 1,043 individuals in 52 populations, referred to as the Human Genome Diversity Project – Centre d'Etude du Polymorphisme Humain (HGDP–CEPH) Human Genome Diversity Cell Line Panel (hagsc.org/hgdp/). We extracted climatic data for each of the 52 population samples using the WorldClim data set at 30 arcsecond (1 km^2) resolution (Hijmans et al. 2005) (supplementary table S2, Supplementary Material online).

A total of 2,624 (0.4%) SNPs obtained $|z|$-scores greater than 5 (supplementary fig. S3, Supplementary Material online). The cutoff $|z| > 5$ ($P < 10^{-7}$) corresponds to the standard Bonferroni correction for a nominal value of type I error $\alpha < 0.01$ and *L* of order 10^5. Among loci with $|z|$-scores greater than 5, 28 genome-wide association study (GWAS) SNPs with known disease or trait association were found (Hindorff et al. 2009). These include several SNPs discovered by Hancock et al. (2011). For example, the SNPs rs12913832 and rs28777 have $|z|$-scores greater than 6 and

Table 3. Loblolly Pines.

Annotation	Gene Ontology	$-\text{Log}_{10}(P\ \text{Value})$
Thylakoid lumenal 19 kDa chloroplast	Oxygen-evolving complex; Photosystem II	9.87
Pentatricopeptide repeat protein	Oxidative stress; salt stress	8.44
Conserved hypothetical protein	Ubiquitin-specific protease	8.28
Chalcone synthase	Flavonoid biosynthesis; wound response; oxidative stress	7.80
Heat shock	Temperature stress	7.67
Dirigent protein pdir18	Disease response	6.56
Heat shock transcription factor hsf5	Regulation of transcription; response to stress	6.15
Zinc finger	Transcription; DNA binding; zinc ion binding	5.84
Probable *n*-acetyltransferase hookless 1	Auxin signaling; photomorphogenesis; ethylene response	5.78
Calcium-binding pollen allergen	Polcalcin; calcium ion binding	4.61
Geranylgeranyl diphosphate synthase	Cholesterol biosynthesis; isoprenoid biosynthesis	4.59
Hypothetical protein OsI_04393	Trehalose-6-phosphate phosphatase	4.59
Potassium proton antiporter	Potassium ion transport; solute:hydrogen antiporter	5.54
DNA mismatch repair	DNA repair; regulation of DNA recombination	5.44

Note.—Annotation and gene ontology for some interesting SNPs with z-scores with absolute value greater than 4 for the first two components of 60 climatic variables.

are associated with genes *OCA2* and *SLC45A2* (table 4). Among the SNPs significantly correlated with climatic gradients, several notable examples include genes associated with celiac disease (ICOSLG), height (*LHX3-QSOX2* and *IGF1*), and vitamin D synthesis or activation (*NADSYN1*-encoding nicotinamide adenine dinucleotide synthetase and *DHCR7* the gene encoding 7-dehydrocholesterol reductase, an enzyme catalyzing the production in skin of cholesterol from 7-dehydrocholesterol) (table 4).

We performed a Gene Ontology enrichment analysis on human genes with $|z|$-scores greater than 5 (2,624 SNPs). Using a threshold of 0.001 for the false discovery rate (FDR) q values, we found significant enrichment of gene ontology terms associated with six biological processes linked to cell adhesion and locomotion, neural and organismal development (supplementary tables S3–S4, Supplementary Material online). The FDR q values for the regulation of developmental processes (76 genes) and the regulation of multicellular organismal processes (88 genes) were equal to $q = 0.006$ and $q = 0.003$. For examples of interesting genes with a high level of association with climatic variables, we focus of the 65 SNPs with $|z|$-scores greater than 7. Among the 65 SNPs, *EPHB4* ($|z| = 8.90$) is involved in heart morphogenesis and angiogenesis, *NRG1* ($|z| = 7.15$) is involved with nervous system development and cell proliferation, *RBM19* ($|z| = 7.04$) is involved with positive regulation of embryonic development, *EYA2* ($|z| = 7.09$) is involved with eye development and DNA repair, and *POLA1* ($|z| = 7.63$) is involved with the mitotic cell cycle and cell proliferation (Saccone et al. 2011; Hornbeck et al. 2012; supplementary table S3, Supplementary Material online). Supplementary table S5, Supplementary Material online, describes a list of 508 SNPs with $|z|$-scores greater than 6.

Discussion

Interpretation of LFMM Results and Other Methods

On the basis of a matrix factorization approach, LFMMs provide a unified framework for estimating effects of environmental and demographic factors on genetic variation. Without environmental variables, LFMMs are similar to performing a sparse version of a probabilistic PCA of allele frequencies (Tipping and Bishop 1999; Engelhardt and Stephens 2010). When environmental variables are included, hidden factors capture the part of genetic variation that cannot be explained by the set of measured environmental variables. This fraction of genetic variation could result from the demographic history of the species, unknown environmental pressures or from IBD patterns.

Although a plethora of statistical tests have been proposed for detecting genes evolving under positive selection and local adaptation (Storz 2005; Novembre and Di Rienzo 2009), the development of tests based on correlations with habitat or landscape variables is still recent (Joost et al. 2007; Hancock et al. 2008). Compared with methods based on summary statistics, tests based on environmental association have increased power to detect selection from standing genetic variation and soft sweeps in a species genome (Pritchard et al. 2010; Schoville et al. 2012). However, simple implementation of these tests, for example, linear or logistic regression models, can be misleading in the presence of IBD patterns (Meirmans 2012). Our simulation results provide clear evidence that tests based on LFMMs significantly reduce the rates of FP associations in the presence of IBD.

Rates of FP and FN were also investigated for three regression methods that include corrections for population genetic structure: PMTs, PCRM, and standard linear mixed models. In the case of phylogenetic comparative analyses that infer environmental correlations for correlated DNA sequences, PMTs were reported to be erroneous (Harmon and Glor 2010). In addition, Legendre and Legendre (2012) warn against using partial Mantel correlations. The high error rate may stem from autocorrelation of matrix elements due to underlying phylogenetic structure. We found that PMTs produced an excess of high and low P values under IBD assumptions. Although PMTs were not correctly calibrated, these tests can provide a useful statistic for ranking loci, and they

Table 4. Human Data.

Landscape-Trait Category	Ref. SNP ID	Nearby Gene	Disease or Trait Association	$-\text{Log}_{10}$ (*P* Value)
Pigmentation and tanning	rs32579	*PPARGC1*	Tanning	9.42
	rs12913832	*OCA2/HERC2*	Eye color, eye color traits, hair color, black vs. blond hair color, black vs. red hair color	9.15
	rs11234027	*DHCR7*	Vitamin D levels	7.78
	rs3129882	*HLA-DRA*	Parkinson's disease	6.97
	rs28777	*SLC45A2*	Black vs. blond hair color, black vs. red hair color	6.90
Immune and autoimmune	rs1250550	*ZMIZ1*	Crohn's disease and inflammatory bowel disease (early onset)	8.77
	rs2735839	*KLK3*	Prostate cancer	8.16
	rs9264942	*RPL3P2*	HIV-1 control	8.02
	rs2179367	Intergenic between *SUMO4* and *ZC3H12D*	Dupuytren's disease	7.57
	rs1551398	Intergenic between *TRIB1* and *LOC100130231*	Crohn's disease	7.45
	rs2289700	*CTSH*	Bipolar disorder	6.98
	rs4819388	*ICOSLG*	Celiac disease	6.67
	rs703842	*CYP27B1/METTL1*	Multiple sclerosis	6.59
	rs12593813	*MAP2K5*	Restless legs syndrome	6.40
	rs4664308	*PLA2R1*	Nephropathy (idiopathic membranous)	6.28
Metabolism	rs10908907	Intergenic *MUC7*	Alcoholism (heaviness of drinking)	8.91
	rs1566039	Intergenic between *PAPD7* and *MIR4278*	Sphingolipid levels	6.89
	rs7665090	*MANBA*	Primary biliary cirrhosis	6.48
Cardiovascular	rs869244	*ADRA2A*	Platelet aggregation	7.20
	rs12034383	*CR1*	Erythrocyte sedimentation rate	7.15
	rs3129882	*HLA-DRA*	Systemic sclerosis	6.97
	rs11897119	*MEIS1*	PR interval	6.71
Height	rs7678436	*NCAPG-LCORL*	Height	9.43
Other	rs12479254	*BOK*	Brain structure	9.43

NOTE.—HGDP SNPs with the highest $|z|$-scores among those associated with phenotypic traits in GWAS.

can detect interesting associations after choosing a tail cutoff (Fumagalli et al. 2011). *P* values based on PCRM and standard linear mixed models were correctly calibrated. But we found that the three regression methods had low power to detect true associations under IBD assumptions. Although these approaches might be useful to detect alleles with strong associations to environmental gradients, they can miss several interesting associations. FN rates were high for PMT, PCRM, and GEMMA because the simulated environmental variable was strongly correlated with population structure. We suspect all regression methods—including LFMM—have higher power when environmental gradients are uncorrelated to the main axis of neutral genetic variation.

Both the mixed model approach of the computer program BAYENV and the LFMM approach include a covariance structure in a regression model, but there are important differences between the two approaches. A first improvement is that LFMMs estimate latent factors and regression coefficients simultaneously, whereas BAYENV first estimates a covariance matrix, and then uses it when estimating (random) environmental effects. To apply BAYENV, the authors suggest utilizing selectively neutral SNPs to estimate the covariance matrix. Inclusion of adaptive markers in the "neutral set" is sometimes unavoidable, and in this case, methods based on the empirical covariance matrix may overlook interesting associations. For Loblolly pines expressed sequence data, the distinction between the two approaches may explain the differences we observed between the list of loci obtained from LFMM and the list obtained from BAYENV (Eckert et al. 2010). For the pine data, it was difficult to select neutral SNPs from the background a priori. Another distinction between LFMM and BAYENV approaches is our use of low rank approximations of the covariance matrix. LFMMs actually estimate correlations between environmental predictors and allele frequencies while *K* hidden factors explain residual genetic variation, where *K* is much smaller than the sample size. Though program speed is generally difficult to evaluate for Markov chain Monte Carlo methods, we observed that LFMM was computationally faster than BAYENV when analyzing large data sets.

Number of Latent Factors

A potential weakness of tests based on LFMM is that we need to choose *K*. In the LFMM modeling approach, the choice of low values for *K* is important for optimizing the computational performances of the estimation algorithm. This choice is reminiscent of selecting the number of components in PCA or in Bayesian clustering programs, and it has also an impact on test outcomes. For values of *K* taken too large, the tests are conservative, and the power to reject neutrality declines. Estimates of *K* that minimize the trade-off between the bias

and variance for our statistical estimates could be obtained by using cross-validation procedures. Cross-validation procedures are computationally intensive, so instead we use Tracy–Widom theory to select K (Patterson et al. 2006). We evaluated this choice during our simulation analysis and found that P values were well calibrated. Although the choice of Tracy–Widom estimates is suboptimal, the performances of LFMMs were superior to those of BAYENV in simulations of IBD patterns. In the analysis of human data, we restricted K to be less than 50 (approximately the number of population samples). We suggest that, when there is a reasonable estimate of the number of genetic clusters for a species, this should be used in LFMM tests directly. For example, estimates of K based on independent genetic data sets could be obtained from Bayesian clustering programs like STRUCTURE (Pritchard et al. 2000). Although finer grain population structure could also be evaluated (Lawson et al. 2012), our choice was again motivated by a trade-off between accuracy and run-time. A future development of our LFMM approach will be to develop fast numerical optimization procedures based on variational approximations of the likelihood, which will allow us to implement cross-validation algorithms and increase the power of tests.

Plant and Human Data

For *P. taeda*, the LFMM results confirmed that several ESTs previously discovered with BAYENV had functions linked to climate (Eckert et al. 2010). In addition, the LFMM analysis discovered new interesting candidate SNPs. Those variants include functions associated with wound repair and immunity; photosynthetic activity and carotenoid biosynthesis; cellular respiration and carbohydrate metabolism; and heat, salt, and oxidative stress responses (table 3). Applying LFMMs to the HGDP data, we found that a total of 0.4% of all polymorphisms (2,624 SNPs) exhibited significant associations with temperature gradients ($|z| > 5$). For example, we identified SNPs associated with the gene OCA2 that may be functionally linked to blue or brown eye color and the gene *SLC45A2* that may be associated with skin pigmentation (Hancock et al. 2011). This list also contained SNPs identified from GWAS studies of height and vitamin D synthesis and diseases such as gluten intolerance and Crohn's disease. Another interesting result is that the list of genic SNPs with $|z|$-scores greater than 5 ($|z| > 5$) was enriched for gene ontology terms associated with six biological processes linked to cell adhesion and locomotion, neural and organismal development. Among the highest scores, the genes EPHB4, BOK, and NRG1—with functions related to heart and brain development—were associated with climatic gradients. Although cautious interpretations of the results may be required (Pavlidis et al. 2012), our data analysis confirmed that many allele frequencies correlate with climatic gradients or with some evolutionary pressures associated with these gradients.

Conclusion

With ever increasing amounts of genetic data generated by high-throughput sequencing technologies, population genetic methods have shifted from empirical approaches to models that incorporate hidden factors. Estimates of ancestry and other population parameters are commonly obtained from mixture models (Pritchard et al. 2000; Durand et al. 2009; Alexander and Lange 2011), principal component analyses (Patterson et al. 2006), hidden Markov models (Price et al. 2009), and factor analysis (Engelhardt and Stephens 2010). Our study contributes to the factor analysis methods for population and landscape genomic analysis by implementing new tests of gene-environment association. These new tests use comparisons between closely related populations that have adapted to different environments, and they may help to detect modes of selection that differ from the classic selective sweep paradigm.

Materials and Methods

LFMM Implementation Details

Consider the data matrix, $(G_{i\ell})$, where each entry records the allele frequency in individual i at the genomic locus ℓ, $1 \leq i \leq n$, $1 \leq \ell \leq L$, and n and L represent the total sample size and number of loci, respectively. LFMMs were defined by the following equation:

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell}$$

where μ_ℓ is a locus-specific effect, β_ℓ is a d-dimensional vector of regression coefficients, U_i and V_ℓ are scalar vectors with K dimensions ($1 \leq K \leq n$). The residuals $\epsilon_{i\ell}$ are statistically independent Gaussian variables of mean zero and variance σ^2.

We use Bayesian analysis to estimate the regression coefficients and their standard deviations. We assume Gaussian prior distributions on μ_ℓ and $\beta_{\ell j}$ with means equal to zero and variances σ_μ^2 and $\sigma_{\beta_j}^2$ ($\beta_{\ell j} \sim N(0, \sigma_{\beta_j}^2)$). Prior distributions on U_i and V_ℓ are Gaussian distributions with means equal to zero and constant variance for each component (the components are independent random variables). The variance of V_ℓ is set to $\sigma_V^2 = 1$. The prior distributions on σ_μ^2 and $\sigma_{\beta_j}^2$ are noninformative distributions. The variance of each factor, σ_U^2, follows an inverse-Gamma distribution $\Gamma^{-1}(\eta,\eta)$ where $\eta = 10^2 - 10^3$. This parameterization encourages sparsity in factor estimates and provides a more accurate description of underlying population structure (Engelhardt and Stephens 2010).

To simultaneously estimate scores (U) and loadings (V), environmmental effects (β), and biases (μ), we implemented a Gibbs sampler algorithm for LFMMs (supplementary file S1, Supplementary Material online). The Gibbs sampler was based on computing products of matrices of low dimension—typical values of K were less than 50—and its speed scales with the current size of SNP data sets, around $n \approx 1{,}000$ and $L \approx 500{,}000$. We implemented a stochastic algorithm to compute standard deviations for the environmental effects (supplementary file S1, Supplementary Material online). The $|z|$-scores were computed as the ratios between the centered values of the regression coefficients β_ℓ and their standard deviations, and they were converted into P values according to the standard Gaussian

distribution. The cutoff for $|z|$-scores was obtained after applying a Bonferroni correction, corresponding to a type I error of 0.01. From a preliminary set of experiments using data simulated from the model defined in equation (1), we found that the estimates of fixed effects stabilized quickly, after 1,000 to 10,000 sweeps for $n = 100 - 1{,}000$ individuals and $L = 1{,}000 - 100{,}000$ loci. A 10-fold increase in the number of sweeps, however, was necessary to recover the true values of the latent factors. Additionally, we developed numerical optimization methods to compute maximum a posteriori (MAP) estimates for the LFMM. One of these methods, the alternate least square method uses deterministic steps that are similar to our stochastic Gibbs sampler (Koren et al. 2009). When checking for convergence of the Markov chain Monte Carlo (MCMC) algorithm, we also found that least square estimates of regression coefficients were close to the point estimates computed by the Gibbs sampler method. The computational complexity of a single sweep of the LFMM Gibbs sampler algorithm is of order $O(nLK^3)$. For about 1,000 loci and 1,000 individuals, the LFMM MCMC algorithm was run for approximately 1 minute of a 2.4 GHz Intel Xeon 64 bit processor. For larger data sets with 650 K loci and 1,000 individuals, we used a multithreated version of the algorithm, for which a single run lasted approximately 24 h on a multiprocessor computer system (using 10 threads).

Alternative Regression Approaches

The standard linear regression model (LM) was defined as

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \epsilon_{i\ell}\,. \tag{7}$$

and the GLMs used the binomial family and the canonical link. The PCRM was defined as

$$G_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \tilde{U}_i^T V_\ell + \epsilon_{i\ell}\,, \tag{8}$$

where $(\tilde{U}_i)$ are the first K PCs computed from the matrix G. For each SNP, we applied PMTs to assess the relationship between the matrix of allele frequency distances and the matrix for environmental variable distance (Smouse et al. 1986; Legendre and Legendre 2012). PMTs are nonparametric permutation-based tests for quantifying association between two distance matrices, while controlling for the effect of a third matrix. The allele frequency distance matrices were computed using Nei's distance (Nei 1972), and the environmental distance matrix used the Euclidean distance. The third matrix was the Pearson's correlation matrix computed over all loci. P values were computed from the R package `vegan` using 10,000 permutations (R Development Core Team 2012). We used the computer program `GEMMA` to implement a standard linear mixed model for genome-wide association studies (Zhou and Stephens 2012). The model had the following form

$$X_i = G_{i\ell}\beta_\ell + u_i + \epsilon_{i\ell} \quad i = 1, \ldots, n\,,$$

where u is a multivariate random effect having a Gaussian distribution of covariance matrix $\lambda\tau^{-1}\Lambda$, and each $\epsilon_{.\ell}$ is a residual error vector having a Gaussian distribution of covariance matrix $\tau^{-1}\mathrm{I}_n$. The parameter τ is the variance of the residual errors, and λ is the ratio between the variance components. The matrix Λ is an $n \times n$ relatedness matrix, and I_n is the identity matrix. Finally, we used the generalized linear mixed model implemented in the computer program `BAYENV` with the default settings of the software (Coop et al. 2010).

LFMM Generative Model Simulations

We used equation (1) with $\beta = 0$ to generate data under a null hypothesis of no association with any environmental variables. In these experiments, we set the number of individuals to $n = 100$, and the number of loci to $L = 1{,}000$. We used six values, $K = 1$, 3, 5, 7, 10, and 20, for the rank of the factor matrix, V. For each series of experiments, we generated 10 replicates of this generative model, and we studied the distributions of P values for tests using LFMMs. In these tests, we set the rank of the factor matrix equal to the values we used to generate simulations.

Next we used equation (1) to generate data showing various levels of population structure and association with an environmental variable. The environmental variable was uniformly generated in the range (0, 1). Here, we used three values for the rank of the factor matrix, $K = 2$, 20, and 100, representing low, moderate, and high levels of underlying population genetic structure. For each series of experiments, we generated 20 replicates of the generative model.

To compute point estimates of environmental effects and their $|z|$-scores, Gibbs sampler algorithms were run for 1,000 sweeps after a burn-in period of 100 sweeps. For these particular run length parameters, we checked that similar estimates were obtained for distinct initializations of the algorithm. For each locus, we recorded both the true, B_ℓ, and estimated environmental effects, $\hat{B}_\ell$, and evaluated the absolute error

$$E_\ell = |\beta_\ell - \hat{\beta}_\ell|\,.$$

Spatial Coalescent Simulations

To enable comparisons with other models, we simulated genotypic data from spatial coalescent models with the computer program `ms` (Hudson 2002). Ten data sets were generated according to a linear stepping-stone model with 40 demes, setting the effective migration rate between pairs of adjacent demes to the value $4Nm = 25$. Sampling five individuals in each deme, each data set included a total of $n = 200$ haploid individuals genotyped at $L = 1{,}000$ unlinked SNP loci. We ran the LFMM during 100 sweeps for burn-in, and we used the next 900 sweeps to compute point estimates, variances, and $|z|$-scores. An environmental variable, x, was defined for each population as the geographic identifier of the population in the linear stepping-stone model.

We created an environmental gradient for the artificial variable x using a logistic function, $s(x)$, of x as follows

$$s(x) = \frac{1}{1 + e^{\theta(x-20)}}\,, \quad \theta > 0. \tag{9}$$

For each of the 10 previously generated neutral stepping-stone simulations, we simulated binary alleles at 50 unlinked loci for each deme x with frequency $s(x)$, and with the slope of the gradient $\theta = 0.1 - 0.2$. We then obtained 10 data sets with $L = 1050$ unlinked loci including 50 loci correlated with the environmental gradient, $s(x)$. Using Tracy–Widom tests implemented in SmartPCA, we found that the number of principal components with P values smaller than 0.01 was around $K_{TW} = 7$. Using the Bayesian clustering programs STRUCTURE and TESS, we found that $K = 5$ components could better describe our simulated data. A value $\theta = 0.2$ corresponds to a strong intensity of selection through geographic space, whereas $\theta = 0.1$ corresponds to a weak intensity of selection. We used the value $\theta = 0.2$ when comparing tests based on linear and PC regression models. When comparing LFMMs with BAYENV, we used the value $\theta = 0.1$ to better fit the objectives of both models. As BAYENV returns Bayes factors instead of P values, we considered ranked lists recording the M loci corresponding to the strongest associations (M between 1 and $L = 1{,}050$). For each M, we computed the number of TPs and the number of FPs. Locus ranking was performed on the basis of $|z|$-scores in LFMM and on the basis of Bayes factors in BAYENV. The LFMM tests used values of K equal to $K = 1$, 3, 5, 7, 10, and 20, and we used of the BAYENV algorithm to compute Bayes factors. Experiments were assessed by counting the number of FP and FN associations, and by measuring the AUC averaged over 10 replicates.

Real Data

Loblolly Pine

The Loblolly pine data consisted of 1,730 SNPs selected in ESTs for 682 individuals (Eckert et al. 2010). We considered 5 environmental variables representing the 5 first components of a PCA for 60 climatic variables (Eckert et al. 2010). The first component (PC1) was mainly described by latitude, longitude, temperature, and winter aridity. PC2 was described by longitude, spring-fall aridity, and precipitation (Eckert et al. 2010). For each of the 5 environmental variables, we applied the LFMM algorithm using 100 sweeps for burn-in and 400 additional sweeps to compute $|z|$-scores for all loci. On the basis of a prior analysis of the genotypic data with the program SmartPCA and Tracy–Widom tests, we used $K = 10$ latent factors.

Human Data

Genotypes from the HGDP–CEPH data set were generated on Illumina 650 K arrays (Li et al. 2008), and the data were filtered to remove low quality SNPs included in the original files. We extracted climatic data for each of the 52 population samples using the WorldClim data set at 30 arcsecond (1 km^2) resolution (Hijmans et al. 2005). These data included 11 bioclimatic variables interpolated from global weather station data collected during a 50-year period (averaged of the years 1950–2000). The environmental variables were mainly related to temperature data. These variables included annual mean temperature, mean diurnal range, maximum temperature of warmest month, minimum temperature of coldest month, and so forth (supplementary table S2, Supplementary Material online). We summarized them by using the first axis of a PCA (all 11 climatic variables were given similar loadings). For this environmental proxy, we applied the LFMM algorithm and computed $|z|$-scores for each locus, using 100 sweeps for burn-in and 900 additional sweeps to compute estimates. We used $K = 50$ which was of the same order as the number of population samples and the value returned by the Tracy–Widom tests. We investigated whether the gene ontology terms of environmentally associated SNPs were enriched in specific categories of biological processes. The list of target genes with $|z|$-scores greater than 5 was compared with the background list of 14,042 genes represented in the HGDP–CEPH data set using a hypergeometric distribution. This test was implemented using the GORILLA software tool (Eden et al. 2009), with significance determined by at an FDR corrected q value threshold of 0.01.

Software Availability

Source codes and computer programs for fitting LFMMs are available from the author websites (http://membres-timc.imag.fr/Eric.Frichot/ and http://membres-timc.imag.fr/Olivier.Francois/lfmm.html).

Supplementary Material

Supplementary file S1 and tables S1–S5 and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by la Région Rhône-Alpes grant to E.F. and O.F.; National Science Foundation grant OISE-0965038 to S.D.S.; and Grenoble INP grant to O.F. The authors are grateful to Florian Alberto, Pierre De Villemereuil, and Oscar Gaggiotti for useful comments of the LFMM software and to Daniel Wegmann and Matteo Fumagalli for their careful reading and helpful suggestions on a previous version of the manuscript.

References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19:711–722.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246.

Barrett RD, Hoekstra HE. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet.* 12:767–780.

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 B Biol Sci.* 263:1619–1626.

Berry A, Kreitman M. 1993. Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. *Genetics* 134:869–893.

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.

Chen C, Durand E, Forbes F, François O. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Mol Ecol Notes.* 7:747–756.

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

Darwin C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray.
Durand E, Jay F, Gaggiotti OE, François O. 2009. Spatial inference of admixture proportions and secondary contact zones. *Mol Biol Evol.* 26:1963–1973.
Eckart C, Young G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1:211–218.
Eckert AJ, Bower AD, Gonzàlez-Martìnez SC, Wegrzyn JL, Coop G, Neale DB. 2010. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Mol Ecol.* 19:3789–3805.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.
Endler JA. 1977. Geographic variation, speciation, and clines. Princeton (NJ): Princeton University Press.
Engelhardt BE, Stephens M. 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* 6:e1001117.
Frichot E, Schoville SD, Bouchard G, François O. 2012. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Front Genet.* 3:254.
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.
Haldane JBS. 1948. The theory of a cline. *J Genet.* 48:277–284.
Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* 7:e1001375.
Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* 4:e32.
Harmon LJ, Glor RE. 2010. Poor statistical performance of the Mantel test in phylogenetic comparative analyses. *Evolution* 64: 2173–2178.
Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol.* 25:1965–1978.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A.* 106:9362–9367.
Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M. 2012. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 40(Database issue):D261–D270.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Jay F, Manel S, Alvarez N, Durand EY, Thuiller W, Holderegger R, Taberlet P, François O. 2012. Forecasting changes in population genetic structure of alpine plants in response to global warming. *Mol Ecol.* 21: 2354–2368.
Jolliffe IT. 1986. Principal component analysis. New York: Springer Verlag.
Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, Taberlet P. 2007. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol Ecol.* 16:3955–3969.
Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* 16:980–989.
Koren Y, Bell R, Volinsky C. 2009. Matrix factorization techniques for recommender systems. *Computer* 8:30–37.
Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet.* 8:e1002453.
Legendre P, Legendre L. 2012. Numerical ecology, 3rd English ed. Amsterdam (Netherlands): Elsevier.
Lenormand T. 2002. Gene flow and the limits to natural selection. *Trends Ecol Evol.* 17:183–189.
Li JZ, Absher DM, Tang H, et al. (11 co-authors). 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
Manel S, Joost S, Epperson BK, Holderegger R, Storfer A, Rosenberg MS, Scribner KT, Bonin A, Fortin MJ. 2010. Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Mol Ecol.* 19:3760–3772.
Meirmans PG. 2012. The trouble with isolation by distance. *Mol Ecol.* 21: 2839–2846.
Nei M. 1972. Genetic distance between populations. *Am Nat.* 106: 283–292.
Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet.* 39:197–218.
Novembre J, Di Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet.* 10:745–755.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2:e190.
Pavlidis P, Jensen JD, Stephan W, Stamatakis A. 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Mol Biol Evol.* 29:3237–3248.
Poncet BN, Herrman D, Gugerli F, Taberlet P, Holderegger R, Gielly L, Rioux D, Thuiller W, Aubert S, Manel S. 2010. Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Mol Ecol.* 19: 2896–2907.
Price AL, Patterson NJ, Plenge RM, Weinblatt Michael E, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 38:904–909.
Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* 5:e1000519.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:R208–R215.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
Prugnolle F, Manica A, Charpentier M, Gugan JF, Guernier V, Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 15:1022–1027.
R Development Core Team. 2012. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Saccone SF, Quan J, Mehta G, Bolze R, Thomas P, Deelman E, Tischfield JA, Rice JP. 2011. New tools and methods for direct programmatic access to the dbSNP relational database. *Nucleic Acids Res.* 39: D901–D907.
Salakhutdinov R, Mnih A. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. *ICML* 25:880–887.
Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S. 2012. Adaptive genetic variation on the landscape: methods and cases. *Annu Rev Ecol Syst.* 43:23–43.
Smouse PE, Long JC, Sokal RR. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Biol.* 35: 627–632.
Storz JF. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol.* 14:671–688.
Storz JF, Wheat CW. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* 64: 2489–2509.
Thibert-Plante X, Hendry AP. 2010. When can ecological speciation be detected with neutral loci? *Mol Ecol.* 19:2301–2314.
Tipping ME, Bishop CM. 1999. Probabilistic principal component analysis. *J Roy Stat Soc B.* 61:611–622.
West M. 2003. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Stat.* 7:723–732.

Williams GC. 1966. Adaptation and natural selection, Vol. 1996. Princeton (NJ): Princeton University Press.

Young JH, Chang YC, Kim JD, Chretien JP, Klag MJ, Levine MA, Ruff CB, Wang NY, Chakravarti A. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* 1:e82.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet.* 38:203–208.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 44:821–824.