doi:10.1111/evo.12193

DISENTANGLING THE EFFECTS OF GEOGRAPHIC AND ECOLOGICAL ISOLATION ON GENETIC DIFFERENTIATION

Gideon S. Bradburd,[1,2] Peter L. Ralph,[3] and Graham M. Coop[1]

[1]*Department of Evolution and Ecology, Center for Population Biology, University of California, Davis, California 95616*

[2]*E-mail: gbradburd@ucdavis.edu*

[3]*Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089*

Received February 11, 2013
Accepted June 5, 2013
Data Archived: Dryad doi:10.5061/dryad.24kp5

Populations can be genetically isolated both by geographic distance and by differences in their ecology or environment that decrease the rate of successful migration. Empirical studies often seek to investigate the relationship between genetic differentiation and some ecological variable(s) while accounting for geographic distance, but common approaches to this problem (such as the partial Mantel test) have a number of drawbacks. In this article, we present a Bayesian method that enables users to quantify the relative contributions of geographic distance and ecological distance to genetic differentiation between sampled populations or individuals. We model the allele frequencies in a set of populations at a set of unlinked loci as spatially correlated Gaussian processes, in which the covariance structure is a decreasing function of both geographic and ecological distance. Parameters of the model are estimated using a Markov chain Monte Carlo algorithm. We call this method Bayesian Estimation of Differentiation in Alleles by Spatial Structure and Local Ecology (BEDASSLE), and have implemented it in a user-friendly format in the statistical platform R. We demonstrate its utility with a simulation study and empirical applications to human and teosinte data sets.

KEY WORDS: **Isolation by distance, isolation by ecology, landscape genetics, partial Mantel test.**

The level ofgenetic differentiation between populations is determined by the homogenizing action of gene flow balanced against differentiating processes such as local adaptation, different adaptive responses to shared environments, and random genetic drift. Geography often limits dispersal, so that the rate of migration is higher between nearby populations and lower between more distant populations. The combination of local genetic drift and distance-limited migration results in local differences in allele frequencies, the magnitude of which increases with geographic distance, resulting in a pattern of isolation by distance (Wright 1943). Extensive theoretical work has described expected patterns of isolation by distance under a variety of models of genetic drift and migration (Charlesworth et al. 2003) in both equilibrium populations in which migration and drift reach a balance, and under nonequilibrium demographic models, such as population expansion or various scenarios of colonization (Slatkin 1993). A range of theoretical approaches have been applied, with authors variously computing probabilities of identity of gene lineages (e.g., Malécot 1975; Rousset 1997) or correlations in allele frequencies (e.g., Slatkin and Maruyama 1975; Weir and Cockerham 1984), or working with the structured coalescent (e.g., Hey 1991; Nordborg and Krone 2002). Although these approaches differ somewhat in detail, their expectations can all be described by a pattern in which allele frequencies are more similar between nearby populations than between distant ones.

In addition to geographic distance, populations can also be isolated by ecological and environmental differences if processes such as dispersal limitations (Wright 1943), biased dispersal (e.g., Edelaar and Bolnick 2012), or selection against migrants due to local adaptation (Wright 1943; Hendry 2004) decrease the rate of

successful migration. Thus, in an environmentally heterogeneous landscape, genome-wide differentiation may increase between populations as either geographic distance or ecological distance increase. The relevant ecological distance may be distance along a single environmental axis, such as difference in average annual rainfall, or distance along a discrete axis describing some landscape or ecological feature not captured by pairwise geographic distance, such as being on serpentine versus nonserpentine soil, or being on different host plants.

Isolation by distance has been observed in many species (Vekemans and Hardy 2004; Meirmans 2012), with a large literature focusing on identifying other ecological and environmental correlates of genomic differentiation. The goals of these empirical studies are generally (1) to determine whether an ecological factor is playing a role in generating the observed pattern of genetic differentiation between populations; and (2) if it is, to determine the strength of that factor relative to that of geographic distance. The vast majority of this work makes use of the partial Mantel test to assess the association between pairwise genetic distance and ecological distance while accounting for geographic distance (Smouse et al. 1986).

A number of valid objections have been raised to the reliability and interpretability of the partial Mantel (e.g., Legendre and Fortin 2010; Guillot and Rousset 2013). First, because the test statistic of the Mantel test is a matrix correlation, it assumes a linear dependence between the distance variables, and will therefore behave poorly if there is a nonlinear relationship (Legendre and Fortin 2010). Second, the Mantel and partial Mantel tests can exhibit high false positive rates when the variables measured are spatially autocorrelated (e.g., when an environmental attribute, such as serpentine soil, is patchily distributed on a landscape), because this structure is not accommodated by the permutation procedure used to assess significance (Guillot and Rousset 2013). Finally, in our view the greatest limitation of the partial Mantel test in its application to landscape genetics may be that it is only able to answer the first question posed earlier—whether an ecological factor plays a role in generating a pattern of genetic differentiation between populations—rather than the first *and* the second—the strength of that factor relative to that of geographic distance. By attempting to control for the effect of geographic distance with matrix regressions, the partial Mantel test makes it hard to simultaneously infer the effect sizes of geography and ecology on genetic differentiation, and because the correlation coefficients are inferred for the matrices of postregression residuals, the inferred effects of both variables are not comparable—they are not in a common currency. We perceive this to be a crucial lacuna in the populations genetics methods toolbox, as studies quantifying the effects of local adaptation (e.g., Rosenblum and Harmon 2011), host-associated differentiation (e.g., Drès and Mallet 2002; Gómez-Díaz et al. 2010), or isolation over ecological distance (e.g., Andrew et al. 2012; Mosca et al. 2012) all require rigorous comparisons to the effect of isolation by geographic distance.

In this article, we present a method that enables users to quantify the relative contributions of geographic distance and ecological distance to genetic differentiation between sampled populations or individuals. To do this, we borrow tools from geostatistics (Diggle et al. 1998) and model the allele frequencies at a set of unlinked loci as spatial Gaussian processes. We use statistical machinery similar to that employed by the Smooth and Continuous AssignmenTs (SCAT) program designed by Wasser et al. (2004) and the BayEnv and BayEnv2 programs designed by Coop et al. (2010) and Günther and Coop (2013). Under this model, the allele frequency of a local population deviates away from a global mean allele frequency specific to that locus, and populations covary, to varying extent, in their deviation from this global mean. We model the strength of the covariance between two populations as a decreasing function of the geographic and ecological distance between them, so that populations that are closer in space or more similar in ecology tend to have more similar allele frequencies. We note that this model is not an explicit population genetics model, but a statistical model—we fit the observed spatial pattern of genetic variation, rather than modeling the processes that generated it. Informally, we can think of this model as representing the simplistic scenario of a set of spatially homogeneous populations at migration–drift equilibrium under isolation by distance.

The parameters of this model are estimated in a Bayesian framework using a Markov chain Monte Carlo algorithm (Metropolis et al. 1953; Hastings 1970). We demonstrate the utility of this method with two previously published data sets. The first is a data set from several subspecies of *Zea mays*, known collectively as teosinte (Fang et al. 2012), in which we examine the contribution of difference in elevation to genetic differentiation between populations. The second is a subset of the Human Genome Diversity Panel (HGDP; Conrad et al. 2006; Li et al. 2008), for which we quantify the effect size of the Himalaya mountain range on genetic differentiation between human populations. We have coded this method—Bayesian Estimation of Differentiation in Alleles by Spatial Structure and Local Ecology (BEDASSLE)—in a user-friendly format in the statistical platform R (R Development Core Team 2013), and have made the code available for download at genescape.org.

Methods

DATA

Our data consist of L unlinked biallelic single nucleotide polymorphisms (SNPs) in K populations; a matrix of pairwise geographic distance between the sampled populations (D); and one or more environmental distance matrices (E). The elements of our

environmental distance matrix may be binary (e.g., same or opposite side of a hypothesized barrier to gene flow) or continuous (e.g., difference in elevation or average annual rainfall between two sampled populations). The matrices D and E can be arbitrary, so long as they are nonnegative definite, a constraint satisfied if they are each matrices of distances with respect to some metric. We summarize the genetic data as a set of allele counts (C) and sample sizes (S). We use $C_{\ell,k}$ to denote the number of observations of one of the two alleles at biallelic locus ℓ in population k out of a total sample size of $S_{\ell,k}$ alleles. The designation of which allele is counted (for convenience, we denote the counted allele as allele '1'), is arbitrary, but must be consistent among populations at the same locus.

LIKELIHOOD FUNCTION

We model the data as follows. The $C_{\ell,k}$ observed '1' alleles in population k at locus ℓ result from randomly sampling a number $S_{\ell,k}$ of alleles from an underlying population in which allele '1' is at frequency $f_{\ell,k}$. These population frequencies $f_{\ell,k}$ are themselves random variables, independent between loci but correlated between populations in a way that depends on pairwise geographic and ecological distance. A flexible way to model these correlations is to assume that the allele frequencies $f_{\ell,k}$ are multivariate normal random variables, inverse logit-transformed to lie between 0 and 1. In other words, we assume that $f_{\ell,k}$ is obtained by adding a deviation $\theta_{\ell,k}$ to the global value μ_ℓ, and transforming:

$$f_{\ell,k} = f(\theta_{\ell,k} + \mu_\ell) = \frac{1}{1 + \exp(-(\theta_{\ell,k} + \mu_\ell))}. \quad (1)$$

Under this notation, μ_ℓ is the transformed mean allele frequency at locus ℓ and $\theta_{\ell,k}$ is the population- and locus-specific deviation from that transformed mean. We can then write the binomial probability of seeing $C_{\ell,k}$ of allele '1' at locus ℓ in population k as

$$P\left(C_{\ell,k}|S_{\ell,k}, f_{\ell,k}\right) = \binom{S_{\ell,k}}{C_{\ell,k}} f_{\ell,k}^{C_{\ell,k}}(1 - f_{\ell,k})^{S_{\ell,k}-C_{\ell,k}}. \quad (2)$$

In doing so, we are assuming that the individuals are outbred, so that the $S_{\ell,k}$ alleles represent independent draws from this population frequency. We will return to relax this assumption later.

To model the covariance of the allele frequencies across populations, we assume that $\theta_{\ell,k}$ are multivariate normally distributed, with mean 0 and a covariance matrix Ω that is a function of the pairwise geographic and ecological distances between the sampled populations. We model the covariance between populations i and j as

$$\Omega_{i,j} = \frac{1}{\alpha_0} \exp\left(-(\alpha_D D_{i,j} + \alpha_E E_{i,j})^{\alpha_2}\right), \quad (3)$$

where $D_{i,j}$ and $E_{i,j}$ are the pairwise geographic and ecological distances between populations i and j, respectively, and α_D and α_E are the effect sizes of geographic distance and ecological distance, respectively. The parameter α_0 controls the variance of population-specific deviate θ (i.e., at $D_{i,j} + E_{i,j} = 0$), and α_2 controls the shape of the decay of the covariance with distance. As alluded to earlier, as many separate ecological distance variables may be included as desired, each with its own α_{E_x} effect size parameter, but here we restrict discussion to a model with one.

With this model, writing $\alpha = (\alpha_0, \alpha_D, \alpha_E, \alpha_2)$, the likelihood of the SNP counts observed at locus ℓ in all sampled populations can now be expressed as

$$P\left(C_\ell, \theta_\ell|S_\ell, \mu_\ell, \alpha\right) = P\left(\theta_\ell \mid \Omega(\alpha)\right) \prod_{k=1}^{K} P\left(C_{\ell,k}|S_{\ell,k}, f(\theta_\ell, \mu_\ell)\right), \quad (4)$$

where we drop subscripts to indicate a vector (e.g., $C_\ell = (C_{\ell 1}, \ldots, C_{\ell K})$), and $P(\theta_\ell|\Omega)$ is the multivariate normal density with mean 0 and covariance matrix Ω.

The joint likelihood of the SNP counts C and the transformed population allele frequencies θ across all L unlinked loci in the sampled populations is just the product across loci:

$$P\left(C, \theta|S, \mu, \alpha\right) = \prod_{\ell=1}^{L} P\left(\theta_\ell|\Omega(\alpha)\right) \prod_{k=1}^{K} P\left(C_{\ell,k}|S_{\ell,k}, f(\theta_\ell, \mu_\ell)\right). \quad (5)$$

POSTERIOR PROBABILITY

We take a Bayesian approach to inference on this problem, and specify priors on each of our parameters. We place exponential priors on α_D and α_E, each with mean 1, and a gamma prior on α_0, with shape and rate parameters both equal to 1. We took the prior on α_2 to be uniform between 0.1 and 2. Finally, we chose a Gaussian prior for each μ_ℓ, with mean 0, variance $1/\beta$, and a gamma distributed hyper-prior on β with shape and rate both equal to 0.001. For a discussion of the rationale for these priors, please see the Appendix.

The full expression for the joint posterior density, including all priors, is therefore given by

$$P(\theta, \mu, \alpha_0, \alpha_D, \alpha_E, \alpha_2, \beta|C, S) \propto \left(\prod_{\ell=1}^{L} P(\theta_{\ell,k}|\Omega) P(\mu_\ell|\beta) \times \prod_{k=1}^{K} P(C_{\ell,k}|S_{\ell,k}, f_{\ell,k})\right) P(\beta)P(\alpha_0)P(\alpha_D)P(\alpha_E)P(\alpha_2), \quad (6)$$

where the various P denote the appropriate marginal densities, and the proportionality is up to the normalization constant given by the right-hand side integrated over all parameters.

MARKOV CHAIN MONTE CARLO

We wish to estimate the posterior distribution of our parameters, particularly α_D and α_E (or at least, their ratio). As the integral of the posterior density given above cannot be solved analytically, we use Markov chain Monte Carlo (MCMC) to sample from the distribution. We wrote a custom MCMC sampler in the statistical platform R (R Development Core Team 2013). The details of our MCMC procedure are given in the Appendix.

MODEL ADEQUACY

Our model is a simplification of the potentially complex relationships present in the data, and there are likely other correlates of differentiation not included in the model. Therefore, it is important to test the model's fit to the data, and to highlight features of the data that the model fails to capture. To do this, we use posterior predictive sampling, using the set of pairwise population F_{ST} values as a summary statistic (Weir and Hill 2002), as we are primarily interested in the fit to the differentiation between pairs of populations. In posterior predictive sampling, draws of parameters are taken from the posterior and used to simulate new data sets, summaries of which can be compared to those observed in the original data sets (Gelman et al. 1996).

Our posterior predictive sampling scheme proceeds as follows. For each replicate of the simulations we

1. Take a set of values of β and all α parameters from their joint posterior (i.e. our MCMC output).
2. Compute a covariance matrix Ω from this set of α and the pairwise geographic and ecological distance matrices from the observed data.
3. Use Ω to generate L multivariate normally distributed θ, and use β to generate a set of normally distributed μ. These θ and μ are transformed using equation (1) into allele frequencies for each population–locus combination, and binomially distributed allele counts are sampled using those frequencies and the per-population sample sizes from the observed data.
4. Calculate F_{ST} between each pair of populations across all loci using the count data. Specifically, we use the F_{ST} estimator defined by the equation given on the top of page 730 in Weir and Hill (2002).

We then use various visualizations of $F_{ST}(i, j)$, for example, plotted against distance between i and j, to compare the patterns in the observed data set to the patterns in the simulated data sets. This functions as a powerful and informative visual summary of the ability of the model to describe the observed data. Because F_{ST} is a good measure of genetic differentiation, users can assess how well the method is able to pick up general trends in the data (e.g., increasing genetic differentiation with ecological or geographic distance) and how well those general trends in the model match the slope of their observed counterparts, and also identify specific pairwise population comparisons that the model is doing a poor job describing. These latter may help reveal other important processes that are generating genetic differentiation between populations, such as unmeasured ecological variables, or heterogeneity in population demography.

ACCOUNTING FOR OVERDISPERSION

A consequence of the form of the covariance given in equation (3) is that all populations have the same variance of allele frequencies about the global mean (and this is $\Omega_{ii} = 1/\alpha_0$). This will be the case in a homogeneous landscape, but is not expected under many scenarios, such as those characterized by local differences in population size, inbreeding rate, historical bottlenecks, or population substructure. In practice, this leads to overdispersion—particular populations deviating more from global means than others. Indeed, in both empirical data sets examined in this article, there are clearly populations with much greater deviation in allele frequencies from the global mean than predicted from their geographical and ecological distances.

To account for this, we will explicitly model the within-population correlations in allelic identity due to varying histories. In so doing, we simultaneously keep outlier populations from having an undue influence on our estimates of α_D and α_E, the effect sizes of the distance variables measured, and highlight those populations that the model is describing poorly. Introducing correlations accounts for overdispersion because a population whose allele frequencies differ more from its predicted frequencies across loci has individuals whose allelic identities are more correlated (and the converse is also true). To see this, observe that, for instance, if one completely selfing population and one outbred population each have a given allele at frequency p, then the variance in sampled allele frequency will be twice as high in the selfing population, because the number of effective independent draws from the pool of alleles is half as large.

To introduce within-population correlations we assume that the allele frequencies from which the allele counts $C_{\ell,k}$ are drawn are not fixed at $f_{\ell,k}$, but rather randomly distributed, with mean given by $f_{\ell,k}$ and variance controlled by another parameter. Specifically, given μ_ℓ and $\theta_{\ell,k}$, we suppose that the allele frequency at locus ℓ in population k is beta-distributed with parameters $\Phi_k f_{\ell,k}$ and $\Phi_k(1 - f_{\ell,k})$, where $f_{\ell,k} = f(\mu_\ell, \theta_{\ell,k})$ as before, and Φ_k is a population-specific parameter, estimated separately in each population, that controls the extent of allelic correlations between draws from individuals in population k. To see why this introduces allelic correlations, consider the following equivalent description of the distribution of $C_{\ell,k}$. We sample the alleles one at a time; if we have drawn n alleles; then the $(n+1)^{\text{st}}$ allele is either: a new draw with probability $\Phi_k/(\Phi_k + n)$ (in which case it is of type '1' with probability $f_{\ell,k}$ and of type '0' with probability

$1 - f_{\ell,k}$); otherwise, it is of the same type as a previously sampled allele, randomly chosen from the n sampled so far. Conceptually, each allele is either a "close relative" of an allele already sampled, or else a "new draw" from the "ancestral population" with allele frequency $f_{\ell,k}$. Smaller values of Φ_k lead to increased allelic correlations, which in turn increase the variance of population allele frequencies.

Conveniently, the random frequency integrates out, so that the likelihood of the count data becomes

$$P(C_{\ell,k}|S_{\ell,k}, f_{\ell,k} = f(\theta_{\ell,k}, \mu_\ell)) = \binom{S_{\ell,k}}{C_{\ell,k}} \frac{B(C_{\ell,k} + \Phi_k f_{\ell,k}, S_{\ell,k} - C_{\ell,k} + \Phi_k(1 - f_{\ell,k}))}{B(\Phi_k f_{\ell,k}, \Phi_k(1 - f_{\ell,k}))}, \quad (7)$$

where $B(x, y)$ is the beta function. This is known as the "beta-binomial" model (Williams 1975), and is used in a population genetics context by Balding and Nichols (1995, 1997); see Balding (2003) for a review.

The parameter Φ_k can be related to one of Wright's F-statistics (Wright 1943). As derived in previous work (Balding and Nichols 1995, 1997), if we define F_k by $\Phi_k = F_k/(1 - F_k)$ ($0 \leq F_k < 1$), then F_k is analogous to the inbreeding coefficient for population k relative to its set of the spatially predicted population frequencies (Cockerham and Weir 1986; Balding 2003), with higher F_k corresponding to higher allelic correlation in population k, as one would expect given increased drift (inbreeding) in that population. However, it is important to note that F_k cannot solely be taken as an estimate of the past strength of drift, because higher F_k would also be expected in populations that simply fit the model less well. We report values of F_k in the output and results, and discuss the interpretation of this parameter further in the discussion.

We have coded this beta-binomial approach as an alternative to the basic model (see Results for a comparison of both approaches on empirical data). To combine estimation of this overdispersion model into our inference framework, we place an inverse exponential prior on Φ_k (i.e., $1/\Phi_k \sim \text{Exp}(5)$). This prior and the beta-binomial probability density function are incorporated into the posterior.

SIMULATION STUDY

We conducted two simulation studies to evaluate the performance of the method. In the first, we simulated data under the inference model, and in the second, we simulated under a spatially explicit coalescent model.

For the data sets simulated under the model, each simulated data set consisted of 30 populations, each with 10 diploid individuals sequenced at 1000 polymorphic bi-allelic loci. Separately for each data set, the geographic locations of the populations were sampled uniformly from the unit square, and geographic distances ($D_{i,j}$) were calculated as the Euclidean distance between them. We also simulated geographically autocorrelated environmental variables, some continuous, some discrete (see Fig. 1A, C). For both discrete and continuous variables, we simulated data sets in which ecological distance had no effect on genetic differentiation between populations; these simulations tested whether our method avoids the false positive issues of the partial Mantel test. We also simulated data sets with an effect of both geographic and ecological distance on genetic distance across a range of relative effect sizes (varying the ratio α_E/α_D) to test our power to detect their relative effects. The study thus consisted of four sections, each composed of 50 data sets: discrete and continuous ecological variables, with or without an effect of ecology.

For each data set, we set $\alpha_0 = 0.5$, and sampled α_D and α_2 from uniform distributions ($U(0.2, 4)$ and $U(0.1, 2)$, respectively); the choice of α_E varied, depending on the specific scenario. These parameters were chosen to give a range of pairwise population F_{ST} spanning an order of magnitude between approximately 0.02 and 0.2, and a realistic allele frequency spectrum. The covariance matrix Ω was calculated using these α and the pairwise geographic and ecological distance matrices (normalized by their standard deviations), and Ω was used to generate the multivariate, normally distributed θ. Values of μ were drawn from a normal distribution with variance $1/(\beta = 0.09)$. [Correction added on September 16, 2013, after first online publication: In the previous sentence, "$\beta = 0.09$" was changed to "$1/(\beta = 0.09)$".] Allele frequencies at each locus were calculated for each population from the θ and μ using equation (1), and SNP counts at each locus in each population were drawn from binomial distributions parameterized by that allele frequency with the requirement that all loci be polymorphic. We simulated under the following ecological scenarios.

1. *Continuous, autocorrelated ecological variable*. For the continuous case, we simulated the values of an ecological variable across populations by sampling from a multivariate normal distribution with mean 0 and covariance between population i and population j equal to $\text{Cov}(E(i), E(j)) = \exp(-D_{i,j}/a_c)$, where a_c determines the scale of the autocorrelation (following Guillot and Rousset 2013). For all simulations, we set $a_c = 0.7$, to represent a reasonably distributed ecological variable on a landscape.
2. *Binary ecological variable*. A binary variable was produced by declaring that the latitudinal equator in the unit square was a barrier to dispersal, so that all populations on the same side of the barrier were separated by an ecological distance of 0, and all population pairs that spanned the equator were separated by an ecological distance of 1.

A. *Zero effect size* For each type of ecological variable, we produced 50 simulated data sets with $\alpha_E = 0$, so that ecological

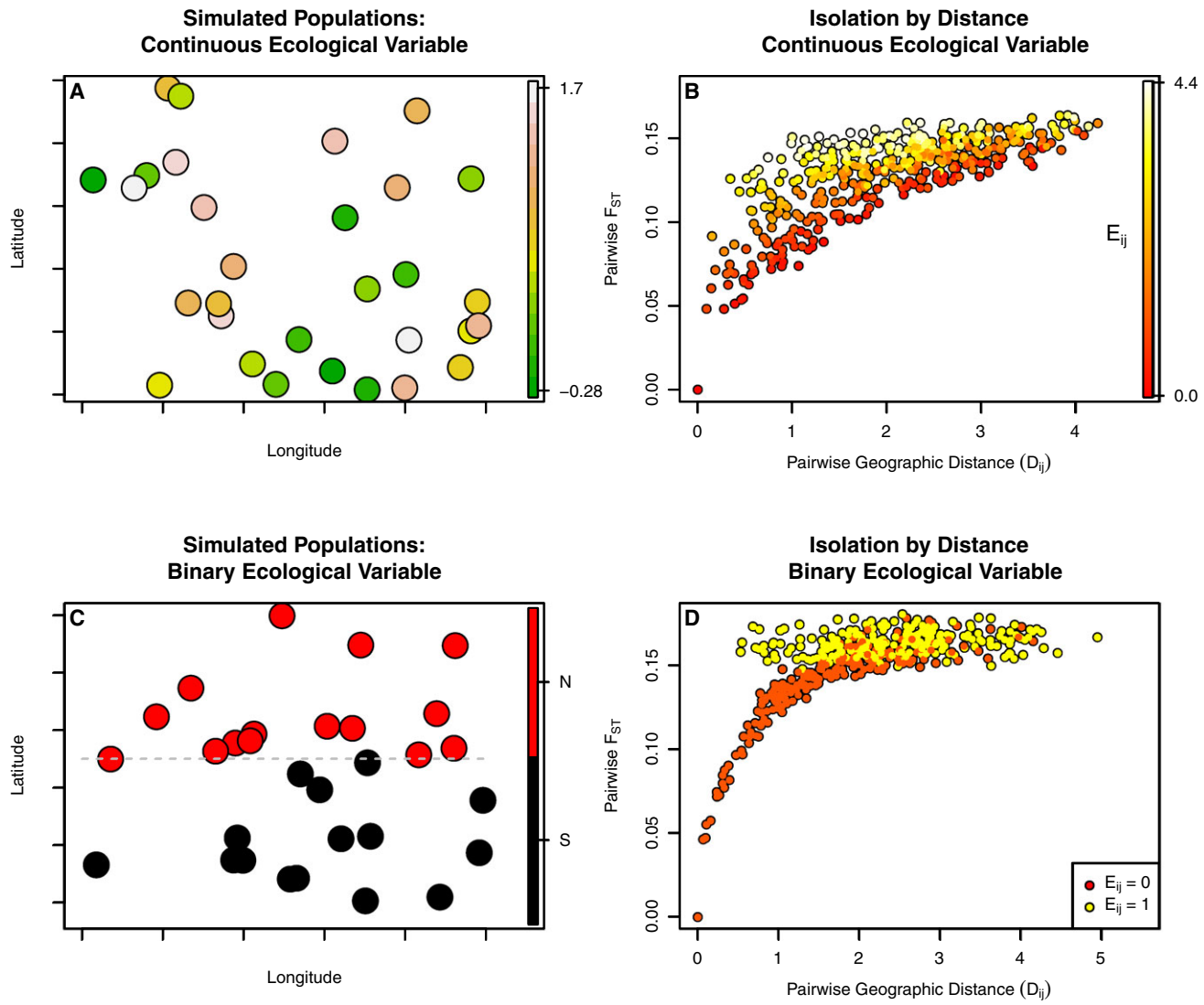

Figure 1. **(A) Populations simulated in the unit square, colored by their value of a continuous ecological variable. (B) Pairwise F_{ST} between simulated populations from (A), colored by difference in their values of the continuous ecological variable. (C) Populations simulated in the unit square, colored by their value of a binary ecological variable. (D) Pairwise F_{ST} between simulated populations from (C), colored by difference in their values of the binary ecological variable.**

distance had no effect on the covariance of θ, and hence on genetic differentiation between populations. For each of these simulated data sets, we performed a partial Mantel test in R using the package *ecodist* (Goslee and Urban 2007) with 1,000,000 permutations.

B. *Varying effect size* We also produced 50 simulated data sets for each type of ecological variable by simulating 10 data sets for each value of α_E/α_D from 0.2 to 1.0 in intervals of 0.2 (see Fig. 1 B, D). (As earlier, values of α_D were drawn from a uniform distribution ($U(0.2, 4)$), so this determines α_E.)

For the data sets simulated using a spatially explicit coalescent process, allelic count data were simulated on a fixed lattice using the program *ms* (Hudson 2002). A total of 49 populations were simulated, evenly spaced in a seven-by-seven grid, of which a subset of 25 populations were sampled to make the final data set; these 25 sampled populations were arranged in a five-by-five grid, as shown in Figure 2. Each population consisted of 10 chromosomes sampled at 1000 polymorphic, unlinked, biallelic loci. Migration occurred between neighboring populations (with no diagonal migration) at a rate of $4Nm_{i,j} = 4$. In all simulations, a longitudinal potential barrier to gene flow was included just to the east of the central line (see Fig. 2). Migration rate between populations that were separated by this barrier was diminished by dividing by some barrier effect size, which varied between simulation sets. For 40 data sets, the barrier effect size was set to 1, so that the barrier had no effect on genetic differentiation across

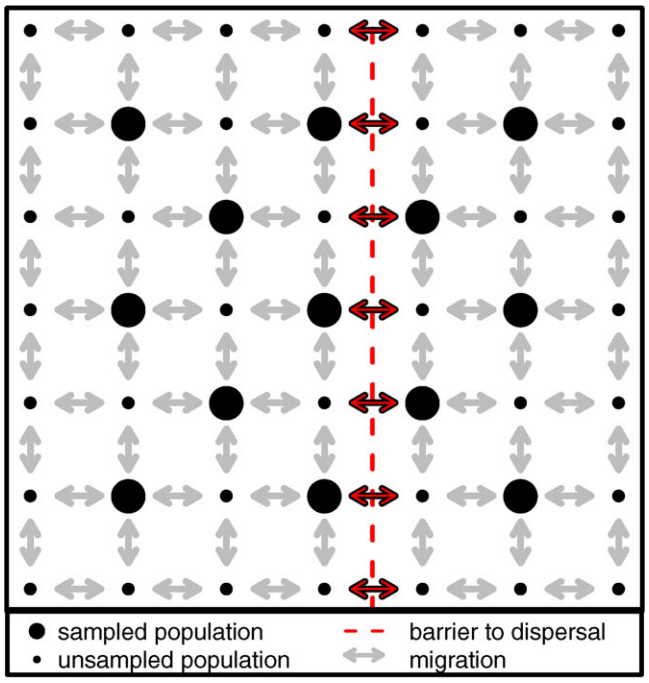

Figure 2. Populations simulated using a spatially explicit coalescent model in the unit square. All simulated populations are indicated with black dots, whereas populations that were sampled for inclusion in each data set are indicated by large black dots. All pairwise migration is indicated with gray arrows. The barrier to dispersal is given by the red dotted line, across which the standard migration rate was divided by a barrier effect size, which we varied.

it. The barrier effect size was set to 5, 10, and 15, for 20 data sets each, for a total of 100 data sets simulated under the spatial coalescent. For all data sets, geographic distance was measured as the pairwise Euclidean distance between populations on the lattice, and ecological distance was defined as 0 between populations on the same side of the barrier, and 1 between populations on opposite sides.

All analyses on the simulated data sets were run for 1,000,000 MCMC iterations, which appeared sufficient in most cases for convergence on the stationary distribution. The chain was sampled every 1000 generations, and all summary statistics from the simulation study were calculated after a burn-in of 20%. The metrics of method performance used on the data sets simulated under the inference model were precision, accuracy, and coverage of the $\alpha_E : \alpha_D$ ratio. We defined *precision* as breadth of the 95% credible set of the marginal posterior distribution; *accuracy* as the absolute value of the difference between the median value of the marginal posterior distributions and the values used to simulate the data in each data set; and *coverage* as the proportion of analyses for which the value used to simulate the data fell within the 95% credible set of the marginal posterior distribution for that parameter. For the data sets simulated under the spatial coalescent process, we wished to assess the ability of the method to accurately recover the relative strength of the barrier to gene flow.

For approximately 30% of all analyses, the MCMC runs displayed obvious difficulty with convergence within the first 1,000,000 generations. The signs of potentially poor single-chain MCMC behavior that we looked for included: acceptance rates that are too low or too high (generally 20–70% acceptance rates are thought to be optimal); parameter trace plots that exhibit high autocorrelation times; acceptance rates that have not plateaued by the end of the analysis; and marginal distributions that are multimodal, or not approximately normal (for a more complete discussion on MCMC diagnosis, please see Gilks et al. 1996, for plots of example MCMC output, see Figs. S5–S7). In some cases, this was because the naive scales of the various tuning parameters of the random-walk proposal mechanisms were inappropriate for the particular data set, and mixing was too slow over the number of generations initially specified (as diagnosed by visualizing the parameter acceptance rates of MCMC generations). This was addressed by re-running analyses on those data sets using different random-walk tuning parameters, or by increasing the number of generations over which the MCMC ran. In the other cases, failure to converge was due to poor performance of the MCMC in regions of parameter space too near the prior boundaries. Specifically, when the chain was randomly started at values of some α parameters too close to 0, it was unable to mix out of that region of parameter space. This problem was addressed by re-running the analyses using different, randomly chosen initial values for the α parameters. In our R package release of the code we provide simple diagnostic tools for the MCMC output, and further guidance for their use.

EMPIRICAL DATA

To demonstrate the utility of this method, we applied it to two empirical data sets: one consisting of populations of teosinte (*Zea mays*), the wild progenitor of maize, and one consisting of human populations from the HGDP panel. Both processed data sets are available for download at genescape.org. See Tables S1 and S2 in the Supporting Information for names and metadata of populations used.

The teosinte data set consisted of 63 populations of between 2 and 30 diploid individuals genotyped at 978 biallelic, variant SNP loci (Fang et al. 2012). Each population was associated with a latitude, longitude, and elevation at the point of sampling (see Fig. S2 and Table S1). Pairwise geographic great-circle distances and ecological distances were calculated for all pairs of populations, where ecological distance was defined as the difference in elevation between populations. Both pairwise distance variables were normalized by their standard deviations.

The human data set was the Eurasian subset of that available from the HGDP (Conrad et al. 2006; Li et al. 2008), consisting of 33 populations of between 6 and 45 individuals genotyped at 1000 biallelic, variant SNP loci (see Fig. S3 and Table S2). Pairwise geographic great-circle distances and ecological distances were calculated for all pairs of populations, where ecological distance was defined as 0 or 1 if the populations were on the same or opposite side of the Himalaya mountain range, respectively. For the purposes of our analysis the western edge of the Himalaya was defined at 75° East.

For comparison, the method was run on each of the two data sets both with and without the beta-binomial overdispersion model. Markov chain Monte Carlo marginal traces were examined visually to assess convergence on a stationary distribution. The chain was thinned by sampling every 1000 generations, and the median and 95% credible sets were reported on the marginal distribution after a burn-in of 20%. The MCMC analysis for the teosinte data set without the overdispersion model was run for 10 million generations; the analysis with the overdispersion model was run for 15 million generations. For the HGDP data set, the numbers of generations were 25 million and 35 million, for the analyses without and with the overdispersion model, respectively.

Results

SIMULATION RESULTS

As described earlier, we conducted two simulation studies. The performance of the method in inference of the parameters of greatest interest is given below.

First we note that, consistent with the results of Guillot and Rousset (2013), the spatial autocorrelation in our ecological variable caused the partial Mantel to have a high false positive rate when $\alpha_E = 0$, which suggests that the partial Mantel test is not well calibrated to assess the significance of ecological distance on patterns of genetic differentiation. At a significance level of $P = 0.05$, the false positive rate for the data sets simulated under the inference model with a binary ecological distance variable was 8%, and for the continuous ecological variable, the false positive error rate was 24%. For the data sets simulated under the spatial coalescent process with a barrier effect size of 1 (meaning that the barrier had no effect on genetic differentiation across it), the false positive error rate was 37.5% (see Fig. S4).

The precision and accuracy results for the data sets simulated under the model with a continuous and discrete ecological variable are visualized in Figure 3A and B, respectively, across the six simulated values of the ratio α_E/α_D. Median precision, accuracy, and coverage are reported in Table 1.

The performance of the method on the data sets simulated using the spatial coalescent model is given in Figure 4, which shows the posterior distributions of $\alpha_E : \alpha_D$ ratio from each analyzed data set over the four barrier effect sizes.

EMPIRICAL RESULTS

Teosinte results

For the *Zea mays* SNP data set analysis, the mean and median of the posterior ratio of the effect size of pairwise difference in elevation to the effect size of pairwise geographic distance (i.e., the $\alpha_E : \alpha_D$ ratio) was 0.153, and the 95% credible set was 0.137–0.171 (see Fig. S10A). The interpretation of this ratio is that 1000 m of elevation difference between two populations has a similar impact on genetic differentiation as around 150 (137–171) km of lateral distance.

Accounting for overdispersion (using the beta-binomial model) we obtain slightly different results, with a mean and median $\alpha_E : \alpha_D$ ratio of 0.205, and a 95% credible set from 0.180 to 0.233 (1000 m difference in elevation $\approx$ 205 km lateral distance, see Fig. S10B). Values of our F statistics F_k estimated across populations ranged from 2×10^{-4} to 0.53, and are shown in Figure S2.

Posterior predictive sampling indicates incorporating overdispersion with the beta-binomial extension results in a better fit to the data (see Fig. 5A, B): the mean Pearson's product moment correlation between the posterior predictive data sets and the observed data without the beta-binomial extension was 0.64, whereas the mean correlation with the beta-binomial model was 0.86 (see Fig. S1A). The ability of the model to predict specific pairwise population F_{ST} is shown Figure S8.

HGDP results

For the human (HGDP) SNP data set analysis, the mean posterior $\alpha_E : \alpha_D$ ratio was 5.13×10^4, the median was 5.00×10^4, and the 95% credible set was 3.09×10^4 to 7.85×10^4 (see Fig. S11A). However, this result seems to be sensitive to outlier populations, as the beta-binomial extension of this method on the same data set yields significantly different results, with a mean $\alpha_E : \alpha_D$ ratio of 1.35×10^4, a median of 1.34×10^4, and a 95% credible set from 1.09×10^4 to 1.65×10^4 (see Fig. S10B). This latter result is broadly consistent with that of Rosenberg (2011), who found an effect size ratio of 9.52×10^3 in a linear regression analysis that treated pairwise population comparisons as independent observations. The interpretation of our result is that being on the opposite side of the Himalaya mountain range has the impact of between approximately 11,000 and 16,000 km of extra pairwise geographic distance on genetic differentiation.

Under our beta-binomial extension values of F_k estimated across populations ranged from 3.2×10^{-4} to 0.06. Population values of F_k are shown on the map in Figure S3.

Posterior predictive sampling again indicates a better fit to the data including overdispersion (see Fig. 5C, D): the mean

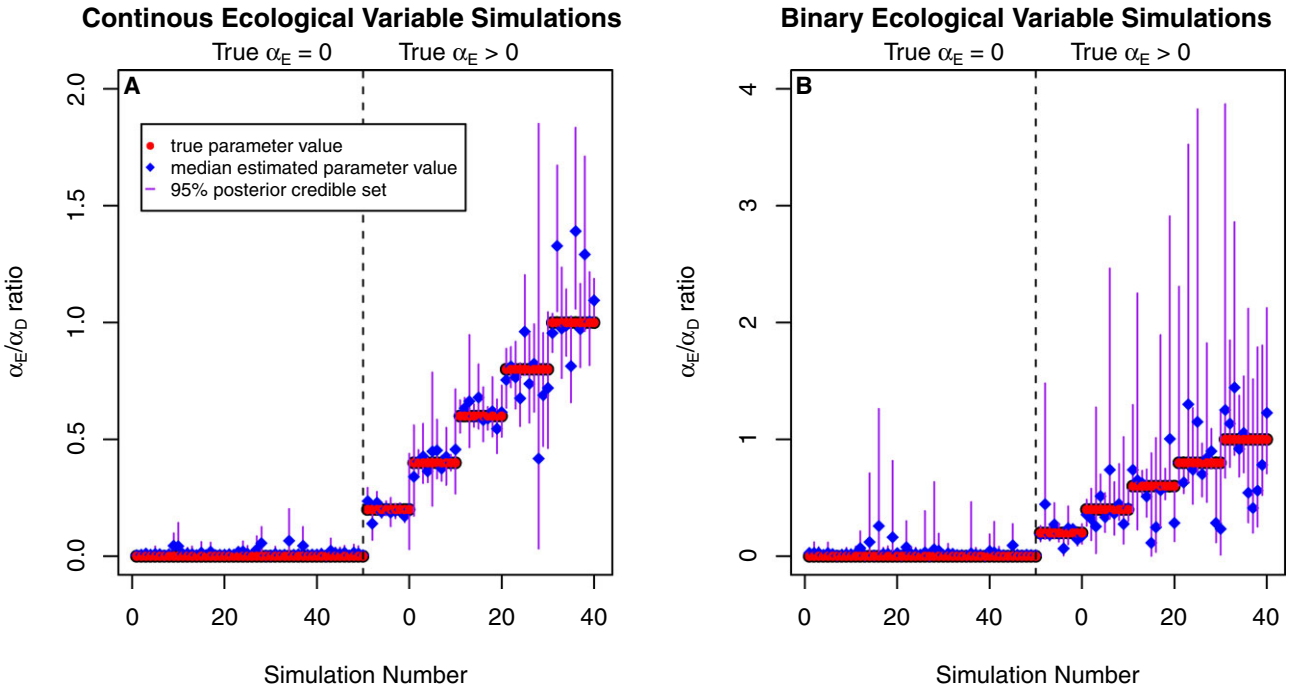

Figure 3. (A) Performance of the method for the 100 data sets simulated with a continuous ecological distance variable. (B) Performance of the method for the 100 data sets simulated with a binary ecological distance variable. In each, the left panel depicts performance on the 50 data sets for which a_E was fixed at 0, and the right panel depicts performance on the 50 data sets for which a_E varied.

Table 1. Simulation studies 1A and 1B were conducted with a continuous ecological variable and $\alpha_E = 0$ and $\alpha_E > 0$, respectively. Simulation studies 2A and 2B were conducted with a binary ecological variable and $\alpha_E = 0$ and $\alpha_E > 0$, respectively. The table reports precision, accuracy, and coverage on the $\alpha_E : \alpha_D$ ratio. *Precision* is breadth of the 95% credible set of the marginal posterior distribution (smaller values indicate better method performance). *Accuracy* is the absolute value of the difference between the median value of the marginal posterior distributions and the values used to simulate the data (smaller values indicate better method performance). *Coverage* is the proportion of analyses for which the value used to simulate the data fell within the 95% credible set of the marginal posterior distribution for that parameter (higher values indicate better method performance). *Coverage* is not reported for the simulations in which the effect size of the ecological distance variable was fixed to zero ($\alpha_E = 0$), as the parameter value used to generate the data is on the prior bound on α_E, and *Coverage* was therefore 0.

	Simulation study 1A	Simulation study 1B	Simulation study 2A	Simulation study 2B
Precision	0.041	0.30	0.15	0.96
Accuracy	0.013	0.0066	0.031	0.033
Coverage	NA	94%	NA	94%

Pearson's product moment correlation between the posterior predictive data sets and the observed data without the beta-binomial extension was 0.88, whereas the mean correlation with the beta-binomial model was 0.91 (see Fig. S1B). The ability of the model to predict specific pairwise population F_{ST} is shown in Figure S9.

Discussion

In this article, we have presented a method that uses raw allelic count data to infer the relative contribution of geographic and ecological distance to genetic differentiation between sampled populations. The method performs quite well: we have shown that it reliably and accurately estimates correct parameter values using simulations, and produces sensible models that give a good fit to observed patterns of differentiation in real data sets. We feel that our method has broad utility to the field of landscape genetics and to studies of local adaptation, and holds a number of advantages over existing methods (although see Wang et al. 2012, for another recent approach). It allows users to simultaneously quantify effect sizes of geographic distance and ecological distance (rather than assessing the significance of a correlation once the effect of geography has been removed, as in the partial Mantel test). Explicitly modeling the covariance in allele frequencies allows users to accommodate nonindependence in the data, and the method's Bayesian framework naturally accommodates uncertainty and provides a means of evaluating model adequacy. The inclusion of overdispersion allows fit to a set of populations with heterogeneous demographic histories. In addition, the basic

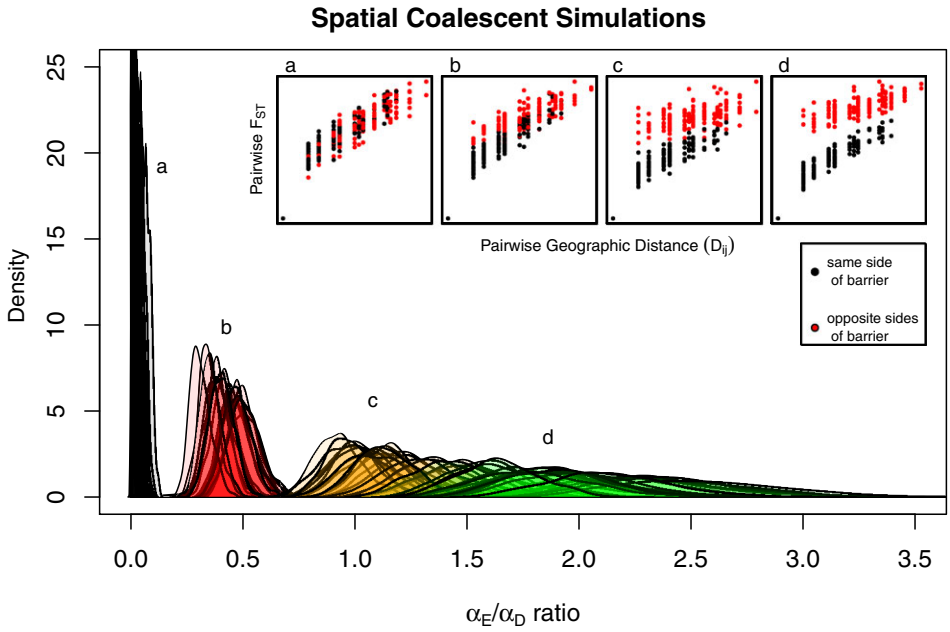

Figure 4. **The marginal distributions on the α_E/α_D ratio from the analyses performed on the data sets simulated using a spatially explicit coalescent process. The migration rate between populations separated by the barrier was divided by a barrier effect size, which varied among simulations. Inset: Pairwise F_{ST}, colored by whether populations were on the same or opposite sides of a barrier to dispersal, plotted against pairwise geographic distance for example data sets for each of the four barrier effect sizes. (A) Barrier effect size of 1 ($n = 40$); (B) Barrier effect size of 5 ($n = 20$); (C) Barrier effect size of 10 ($n = 20$); and (D) Barrier effect size of 15 ($n = 20$).**

model presented here—a parametric model of spatial covariance in allele frequencies—is extremely versatile, allowing for the inclusion of multiple ecological or geographic distance variables, as well as great flexibility in the function used to model the covariance.

SIMULATION STUDY

Our method performed well in both simulation studies (see Figs. 3, 4, and Table 1), and was able to effectively recognize and indicate when an ecological variable contributes significantly to genetic differentiation. This is in contrast to the partial Mantel, which has a high false positive rate in the presence of spatial autocorrelation of environmental variables (see Fig. S4).

For data sets simulated under the inference model, coverage, accuracy, and precision were all satisfactory (see Table 1). The precision of our estimator of α_E was generally lower for our discrete ecological variable, likely due to the strong spatial structure of the discrete ecological variable.

For data sets simulated using the spatial coalescent, there were no true values for the $\alpha_E : \alpha_D$ ratio to compare with those inferred by the method. However, we note that the $\alpha_E : \alpha_D$ ratios estimated across analyzed simulated data sets tracked the barrier effect sizes used to simulate them, and that when the barrier had no effect on migration, the marginal distributions on the $\alpha_E : \alpha_D$ ratio estimated were stacked up against the prior bound at 0 and had very low median values. The width of the 95% credible set of the marginal posteriors grew with the barrier effect size as a result of the flattening of the posterior probability surface as true parameter value increased. Overall, the method performed well on the data sets simulated under a model different from that used for inference (and presumably closer to reality).

An issue we observed in practice is that at some parameter values, different combinations of α are essentially nonidentifiable—the form of the covariance given in equation (3) sometimes allows equally reasonable fits at different values of α_2, or at different combinations of α_0, α_D, and α_E. (In other cases, all four parameters can be well-estimated.) Even when this is the case, the α_E:α_D ratio, which is the real parameter of interest, remains constant across the credible region, even as α_E and α_D change together to compensate for changes in α_2 and α_0. Such 'ridges' in the likelihood surface are readily diagnosed by viewing the trace plots and joint marginals of the α parameters (see Figs. S5, S6).

EMPIRICAL RESULTS

Teosinte

The application of our method to the teosinte SNP data set indicated that difference in elevation has a potentially substantial contribution to genetic differentiation between teosinte populations. Difference in elevation could be correlated with another, as yet unmeasured ecological variable, so we cannot claim to report a causal link, but these results are certainly suggestive, especially in

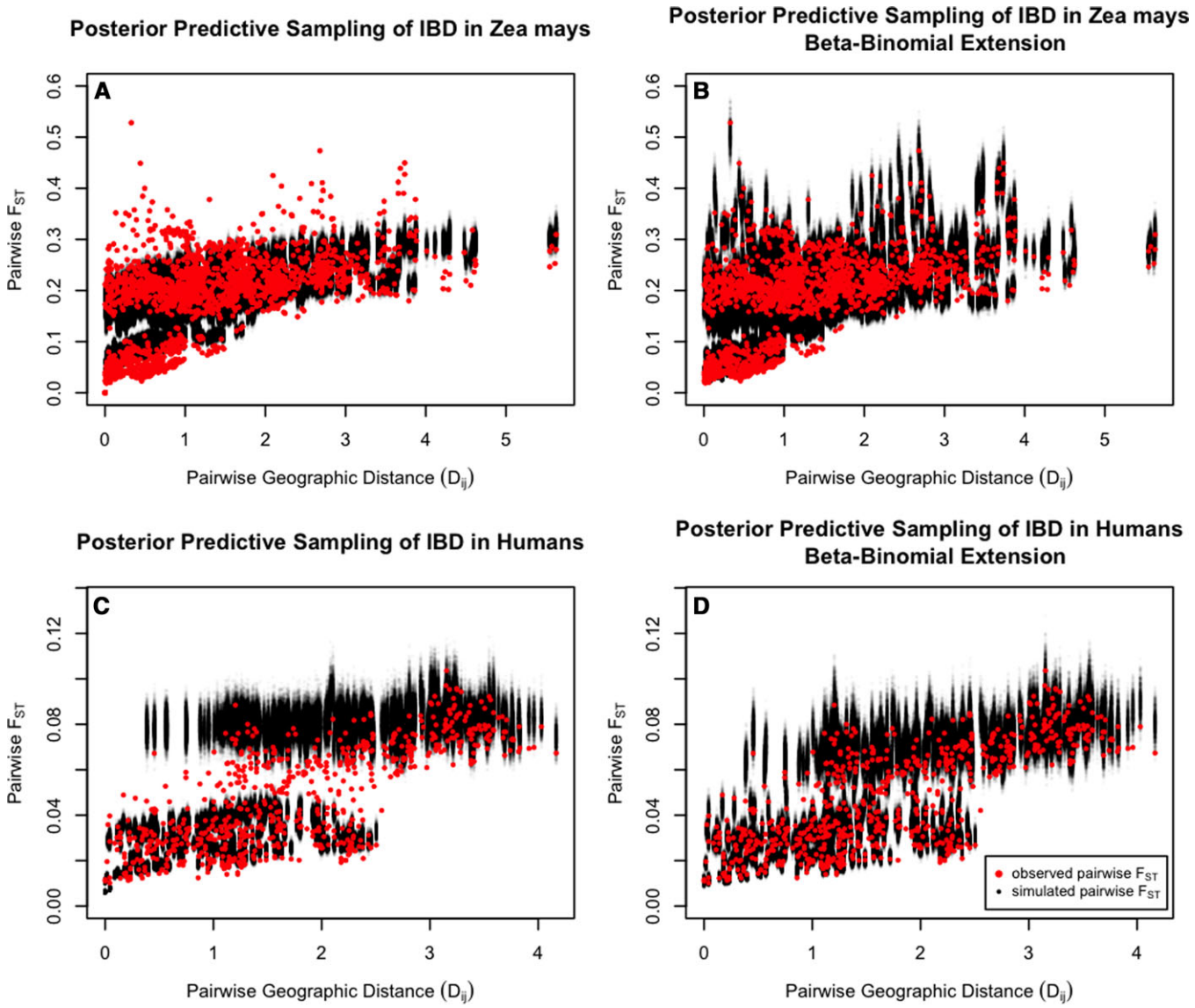

Figure 5. Posterior predictive sampling with 1000 simulated data sets, using pairwise F_{ST} as a summary statistic of the allelic count data for: (A) the teosinte data set, using the standard model; (B) the Teosinte data set, using the overdispersion model; (C) HGDP data set, standard model; and (D) HGDP data set, overdispersion model.

the light of the research on morphological adaptations in teosinte to high altitude (Eagles and Lothrop 1994).

The analysis of the teosinte SNP data with the beta-binomial extension of our method shows a much better model fit, and highlights a number of populations with particularly high F_k values. These populations (highlighted in Fig. S2) all belong to the subspecies *Zea mays mexicana*, which primarily occurs at higher altitudes and is hypothesized to have undergone significant drift due to small effective population sizes or bottlenecks (Fukunaga et al. 2005). In addition, a number of these populations occur in putative hybrid zones between *Zea mays mexicana* and *Zea mays parviglumis*, a separate, co-occuring subspecies (Heerwaarden et al. 2011). Like drift, admixture would have the effect of increasing the variance in observed allele frequencies around the expectation derived from the strict geographic/ecological distance model, and would drive up the inferred F_k parameters for admixed populations.

Human genome diversity panel

In the HGDP data, we find a strong effect of separation by the Himalayas on genetic differentiation, confirming previous results (e.g., Rosenberg et al. 2005). To obtain a good fit to the data it is necessary to model overdispersion (with the beta-binomial extension). This lack of model fit of the basic model can be seen in the posterior predictive sampling in Figure 5C and D, which highlights the importance of assessing model adequacy

during analysis. Under the beta-binomial extension, the α_E/α_D ratio estimates an effect of the Himalayas far greater than the distance simply to circumnavigate around the Himalayas. We think this likely reflects the fact that Eurasian populations are away from migration–selection equilibrium, reflecting past large-scale population expansions (Keinan et al. 2007).

With overdispersion included, the model appears to describe the data reasonably well, suggesting substantial heterogeneity beyond that dictated by geographic distance and separation by the Himalayas between the sampled populations. A number of populations stand out in their F_k values, in particular the Kalash, the Lahu, the Mozabites, the Hazara, and the Uygur (highlighted in Fig. S3). This is consistent with the known history of these populations and previous work on these samples (Rosenberg et al. 2002), which suggests that these populations are unusual for their geographic position (i.e., they depart from expectations of their covariance in allele frequencies with their neighbors). The Hazara and Uygur populations are known to be recently admixed populations between central Asian and East ancestry populations. The Mozabite population has substantial recent admixture from Sub-Saharan African populations (Rosenberg et al. 2002; Rosenberg 2011). The Kalash, who live in northwest Pakistan, are an isolated population with low heterozygosity, suggesting a historically small effective population size. Finally, the Lahu have unusually low heterozygosity compared to the other East Asian populations, suggesting that they too may have had an unusually low effective population size. Thus, our beta-binomial model, in addition to improving the fit to the data, is successfully highlighting populations that are outliers from simple patterns of isolation by distance.

Population-specific variance

As noted earlier, in both empirical data sets analyzed, the beta-binomial extension to the basic model offers substantially better model fit. This could in part reflect ecological variables not included in the analyses in addition to heterogeneity in demographic processes, both of which could shape genetic variation in these populations by pushing population allele frequencies away from their expectations under our simple isolation by distance and ecology model. Our F_k statistic provides a useful way to highlight populations that show the strongest deviations away from our model, and to prevent these deviations from obscuring environmental correlations or causing spurious correlations. Therefore, we recommend that the extended model be used as the default model for analyses.

LIMITATIONS

The flexibility of this statistical model is accompanied by computational expense. Depending on the number of loci and populations in a data set, as well as the number of MCMC generation required to accurately describe the stationary distribution, analyses can take anywhere from hours to days. Speedups could be obtained by parallelization or porting code to C. In addition, as with any method that employs an MCMC algorithm, users should take care to assess MCMC performance to ensure that the chain is mixing well, has been run for a sufficient number of generations, and has converged on a stationary distribution (Gilks et al. 1996). Users are well advised to run multiple independent chains from random initial locations in parameter space, and to compare the output of those analyses to confirm that all are describing the same stationary distributions.

Our model rests on a number of assumptions, principal among which is that population allele frequencies are well represented by a spatially homogeneous process, such as are obtained under mutation–migration equilibrium. That is, we assume that current patterns of gene flow between populations are solely responsible for observed patterns of genetic differentiation. Some examples of biological situations that may violate the assumptions of our model include: two populations that have higher genetic differentiation than expected based on their pairwise geographic distance because they arrived in nearby locations as part of separate waves of colonization; or two populations that have been recently founded on either side of some landscape element that truly does act as a barrier to gene flow, but that do not exhibit strong genetic differentiation yet, because the system is not in equilibrium. In reality, we expect that very few natural populations will conform perfectly to the assumptions of our model; however, we feel that the method will provide valid approximations of the patterns for many systems, and that it will be a useful tool for teasing apart patterns of genetic variation in populations across heterogeneous landscapes.

EXTENSIONS

The flexibility of this method translates well into extendability. Among a number of natural extensions the community might be interested in implementing, we highlight a few here.

One natural extension is to incorporate different definitions of the ecological distance between our populations. Just because two populations have no difference in their ecological variable state does not guarantee that there is not great heterogeneity in the distance between them. For example, a pair of populations separated by the Grand Canyon might have nearly identical elevations, but the cost to migrants between them incurred by elevation may well be significant. One solution to this would be to enter a simple binary barrier variable, or to calculate least-cost paths between populations, and use those distances in lieu of geographic distance. A more elegant solution would be to use "isolation by resistance" distances, obtained by rasterizing landscapes and employing results relating mean passage rates of random walks in a heterogeneous environment to quantities from circuit

theory to calculate the conductance (ease of migration) between nodes on that landscape (McRae and Beier 2007). This method has the advantage of integrating over all possible pathways between populations. Currently, users must specify the resistance of landscape elements a priori, but those resistance parameters could be incorporated into our parametric covariance function, and estimated along with the other parameters of our model in the same MCMC. This approach carries great appeal, as it combines the conceptual rigor of accommodating multiple migration paths with the methodological rigor of statistically estimated, rather than user-specified, parameter values.

Another extension is the further relaxation of the assumption of process homogeneity in decay of allelic covariance over geographic and ecological distance. Specifically, the method currently requires that a single unit of pairwise ecological distance translate into the same extent of pairwise genetic differentiation between all population pairs. This assumption is unlikely to be realistic in most empirical examples, especially if populations are locally adapted. For example, individuals from populations adapted to high elevation may be able to migrate more easily over topography than individuals from populations adapted to low elevations. Such heterogeneity could be accommodated by using different covariance functions for different, prespecified population pairings.

A final extension that could be integrated into this method is a model selection framework, in which models with and without an ecological distance variable, or with different combinations of ecological distance variables, can be rigorously compared. Because our method is implemented in a Bayesian framework, we could select between models by calculating Bayes factors (the ratio of the marginal likelihoods of the data under two competing hypotheses; Dickey 1971; Verdinelli and Wasserman 1995). This approach would seem to offer the best of both worlds: robust parameter inference that accommodates uncertainty in addition to output that could be interpreted as definitive evidence for or against the association of an ecological variable of interest with genetic differentiation between populations.

CONCLUSION

In closing, we present a tool that can be useful in a wide variety of contexts, allowing a description of the landscape as viewed by the movements of genetic material between populations. We urge users to be cautious in their interpretation of results generated with this model. A correlation between genetic differentiation and an ecological distance variable does not guarantee a causal relationship, especially because unmeasured ecological variables may be highly correlated with those included in an analysis. In addition, evidence of a correlation between genetic differentiation and an ecological variable may not be evidence of local adaptation or selection against migrants, as both neutral and selective forces can give rise to an association between genetic divergence and ecological distance.

Finally, we are making this method available online at genescape.org, and we hope that users elaborate on the framework we present here to derive new models that are better able to describe empirical patterns of isolation by distance—both geographic and ecological.

ACKNOWLEDGMENTS

The authors thank Y. Brandvain, M. Weber, L. Mahler, W. Wetzel, B. Moore and the Coop lab for their counsel, J. Ross-Ibarra and T. Günther for their help with empirical data sets, J. Novembre and D. Davison for their code, and J. Wilkins and two anonymous reviewers for their comments on previous drafts. This material is based upon work supported by the National Science Foundation under Grant No. 1262645 (PR and GC), National Science Foundation GRFP No. 1148897 (GB), a National Institutes of Health Ruth L. Kirschstein NRSA fellowship F32GM096686 (PR), and a Sloan Foundation fellowship (GC).

LITERATURE CITED

Andrew, R. L., K. L. Ostevik, D. P. Ebert, and L. H. Rieseberg. 2012. Adaptation with gene flow across the landscape in a dune sunflower. Mol. Ecol. 21:2078–2091.

Balding, D. J. 2003. Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. 63:221–230.

Balding, D. J., and R. A. Nichols. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12.

Balding, D. J., and R. A. Nichols. 1997. Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 108:583–589.

Charlesworth, B., D. Charlesworth, and N. H. Barton. 2003. The effects of genetic and geographic structure on neutral variation. Ann. Rev. Ecol. Evol. Syst. 34:99–125.

Christensen, O. F., and R. Waagepetersen. 2002. Bayesian prediction of spatial count data using generalized linear mixed models. Biometrics 58:280–286.

Cockerham, C. C., and B. S. Weir. 1986. Estimation of inbreeding parameters in stratified populations. Ann. Human Genet. 50:271–281.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, N. a. Rosenberg, and J. K. Pritchard. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 38:1251–1260.

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

Dickey, J. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann. Math. Stat. 42:204–223.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. Model-based geostatistics. J. R. Stat. Soc. Ser. C (Appl. Stat.) 47:299–350.

Mallet, Drès, M. and J. 2002. Host races in plant-feeding insects and their importance in sympatric speciation. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 357:471–492.

Eagles, H. A., and J. E. Lothrop. 1994. Highland maize from central Mexico-Its origin, characteristics, and use in breeding programs. Crop Sci. 34:11–19.

Edelaar, P., and D. I. Bolnick. 2012. Non-random gene flow: an underappreciated force in evolution and ecology. Trends Ecol. Evol. 27:659–665.

Fang, Z., T. Pyhäjärvi, A. L. Weber, R. K. Dawe, J. C. Glaubitz, J. D. Jesus, S. González, C. Ross-ibarra, J. Doebley, and P. L. Morrell. 2012. Megabase-scale inversion polymorphism in the wild ancestor of maize. Genetics 191:883–894.

Fukunaga, K., J. Hill, Y. Vigouroux, Y. Matsuoka, J. Sanchez G., K. Liu, E. S. Buckler, and J. Doebley. 2005. Genetic diversity and population structure of teosinte. Genetics 169:2241–2254.

Gelman, A., X.-l. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6:733–807.

Gilks, W., S. Richardson, and D. Spiegelhalter. 1996. Markov Chain Monte Carlo in practice. Interdisciplinary statistics. Chapman & Hall, Boca Raton, Florida.

Gómez-Díaz, E., P. F. Doherty Jr, D. Duneau, and K. D. McCoy. 2010. Cryptic vector divergence masks vector-specific patterns of infection: an example from the marine cycle of Lyme borreliosis. Evol. Appl. 3:391–401.

Goslee, S. C., and D. L. Urban. 2007. The ecodist package for dissimilarity-based analysis of ecological data. J. Stat. Software 22:1–19.

Guillot, G., and F. Rousset. 2013. Dismantling the Mantel tests. Methods Ecol. Evol. 4:336–344.

Günther, T., and G. Coop. 2013. Robust identification of local adaptation from allele frequencies. Genetics genetics.113.152462. doi:10.1534/genetics.113.152462.

Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Heerwaarden, J. V., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman, J. D. J. S. Gonzalez, and J. Ross-Ibarra. 2011. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. PNAS 108:1088–1092.

Hendry, A. P. 2004. Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. Evol. Ecol. Res. 6:1219–1236.

Hey, J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Popul. Biol. 39:30–48.

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

Keinan, A., J. C. Mullikin, N. Patterson, and D. Reich 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. 39:1251–5.

Legendre, P., and M.-J. Fortin. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol. Ecol. Res. 10:831–844.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Malécot, G. 1975. Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. 8:212–241.

McRae, B. H., and P. Beier. 2007. Circuit theory predicts gene flow in plant and animal populations. PNAS 104:19885–19890.

Meirmans, P. G. 2012. The trouble with isolation by distance. Mol. Ecol. 21:2839–2846.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of State Calculations. J. Chem. Phys. 21:1087–1092.

Møller, J., A. R. Syversveen, and R. P. Waagepetersen. 1998. Log Gaussian Cox Processes. Scand. J. Stat. 25:451–482.

Mosca, E., A. J. Eckert, E. A. D. I. Pierro, D. Rocchini, and N. L. A. Porta. 2012. The geographical and environmental determinants of genetic diversity for four alpine conifers of the European Alps. Mol. Ecol. 21:5530–5545.

Nordborg, M., and S. M. Krone. 2002. Separation of time scales and convergence to the coalescent in structured populations. Pp. 130–164, *in* M. Slatkin and M. Veuille, eds. Modern developments in theoretical populations genetics. Oxford Univ. Press, Oxford, U.K.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org. ISBN 3-900051-07-0.

Rosenberg, N. A. 2011. A population-genetic perspective on the similarities and differences among worldwide human populations. Human Biol. 83:659–684.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. 2002. Genetic structure of human populations. Science 298:2381–2385.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1:660–671.

Rosenblum, E. B., and L. J. Harmon. 2011. "Same same but different": replicated ecological speciation at White Sands. Evolution 65:946–960.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219–1228.

Slatkin, A. M., and T. Maruyama. 1975. The influence of gene flow on genetic distance. Am. Nat. 109:597–601.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264–279.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence extensions of the multiple regression and correlation Mantel test of matrix correspondence. Syst. Zool. 35:627–632.

Vekemans, X., and O. Hardy. 2004. New insights from fine-scale spatial genetic structure analyses in plant populations. Mol. Ecol. 13:921–935.

Verdinelli, I., and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Am. Stat. Assoc. 90:614–618.

Wang, I. J., R. E. Glor, and J. Losos. 2012. Quantifying the roles of ecology and geography in spatial genetic divergence. Ecol. Lett. 16:175–182.

Wasser, S. K., A. M. Shedlock, K. Comstock, E. A. Ostrander, B. Mutayoba, and M. Stephens. 2004. Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. PNAS 101:14847–14852.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

Weir, B. S., and W. G. Hill. 2002. Estimating F-statistics. Ann. Rev. Genet. 36:721–750.

Williams, D. A. 1975. 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31:949–952.

Wright, S. 1943. Isolation by distance. Genetics 28:114–138.

Associate Editor: J. Wilkins

Appendix

PRIORS

We denote a gamma distribution with given shape and rate parameters as Γ(shape, rate), a normal distribution with given mean and variance parameters as N(mean, variance), an exponential distribution with given rate parameter Exp(rate), and a uniform distribution between given upper and lower boundaries as U(lower, upper). The priors specified on the parameters

of this model are: $\alpha_0 \sim \Gamma(0.001, 0.001)$; $\alpha_D \sim \text{Exp}(1)$; $\alpha_E \sim \text{Exp}(1)$; $\alpha_2 \sim U(0.1, 2)$; and $\mu_\ell \sim N(0, 1/\beta)$, with a hyperprior $\beta \sim \Gamma(0.001, 0.001)$. [Correction added on September 16, 2013, after first online publication: In the previous sentence, "(1, 1)" was changed to "(0.001, 0.001)".]

The priors on α_D and α_E were chosen to reflect the assumption that there is some, and potentially very great, effect of isolation by geography and ecology. The priors on α_2, α_0, and β were the same as those used by Wasser et al. (2004), and, in the case of the latter two (on β and α_0), were chosen because they were conjugate to the likelihood, so their parameters could therefore be updated by a Gibbs sampling step.

In early implementations of our method, we experimented with uniform priors on α_D and α_E (U(0,4)), as used by Wasser et al. (2004) (although they did not have a parameter analogous to α_E). We replaced these uniform priors with exponentials to reflect the fact that we have no prior belief that there should be any upper bound to the effects geographic or ecological distance may have on genetic differentiation. In practice, we found that for all simulated and empirical data sets tested, there was sufficient information in the data for the likelihood function to swamp the effect of the priors—whether uniform or exponential—on α_D and α_E.

However, in all analyses, we encourage users to visualize the marginal distributions of each parameter at the end of a run and compare it to its prior. If the marginal distribution looks exactly like the prior, there may be insufficient information in the data to parameterize the model effectively, and the prior may be having an unduly large impact on analysis. If the marginal distribution for a parameter shows that it is "piling up" against its prior's hard bound (e.g., the marginal distribution on α_E has a median of $1e-3$, close to its hard bound at 0), that may suggest that the current form of the prior is not describing the natural distribution of the parameter for that particular data set well (e.g., α_E "wants" to be 0, but the prior is constraining it). In both cases (the marginal posterior and the prior have significant overlap; the prior is exhibiting an edge effect), we suggest that the user experiment with different priors and/or model parameterizations to see what effect they are having on inference.

MARKOV CHAIN MONTE CARLO

Our MCMC scheme proceeds as follows. The chain is initiated at maximum likelihood estimates (MLEs) for θ and μ, and, for $\alpha_0, \alpha_D, \alpha_E$, and α_2, at values drawn randomly from their priors. The multiplicative inverse of the empirical variance of the MLEs of μ is used as the initial value of β. [Correction added on September 16, 2013, after first online publication: In the previous sentence, "The empirical standard deviation" was changed to "The multiplicative inverse of the empirical variance".]

In each generation one of $\{\mu, \beta, \theta, \alpha_0, \alpha_D, \alpha_E, \alpha_2\}$ is selected at random to be updated.

The priors on β and α_0 are conjugate to their marginal posteriors, and each is updated via a Gibbs sampling step. The updated value of β given the current μ is drawn from

$$\beta \mid \mu_1, \ldots, \mu_L \sim \Gamma\left(0.001 + \frac{L}{2}, 0.001 + \frac{1}{2}\sum_{\ell=1}^{L}\mu_\ell^2\right), \qquad (8)$$

and the updated value of α_0 conditional on the current set of θ is drawn from

$$\alpha_0 \mid \theta_1, \ldots, \theta_L \sim \Gamma\left(1 + \frac{Lk}{2}, 1 + \frac{1}{2}\sum_{\ell=1}^{L}\theta_{\ell,k}\chi^{-1}\theta_{\ell,k}^T\right), \qquad (9)$$

where k is the number of populations sampled, L is the number of loci sequenced, and $\chi_{i,j} = \alpha_0\Omega_{i,j} = \exp\left(-(\alpha_D D_{i,j} + \alpha_E E_{i,j})^{\alpha_2}\right)$.

The remaining parameters are updated by a Metropolis–Hastings step; here we describe the proposal mechanisms. The proposed updates to θ do not affect each other, and so are accepted or rejected independently. Following Wasser et al. (2004) (derived from Christensen and Waagepetersen 2002; Møller et al. 1998), the proposal is chosen as $\theta_\ell' = \theta_\ell + R_\ell Z$, where R is a vector of normally distributed random variables with mean 0 and small variance (controlled by the scale of the tuning parameter on θ) and Z is the Cholesky decomposition of Ω (so that $ZZ^{\mathrm{T}} = \Omega$). Under this proposal mechanism, proposed updates to θ_ℓ tend to stay within the region of high posterior probability, so that more updates are accepted and mixing is improved relative to a scheme in which the θ in each population were updated individually.

Updates to α_D, α_E, and α_2 are accomplished via a random-walk sampler (adding a normally distributed random variable with mean 0 and small variance to the current value; Gilks et al. 1996). Updates to elements of μ_ℓ are also accomplished via a random-walk sampler, and again the updates to each locus are accepted or rejected independently.

In the overdispersion model, initial values of Φ_k are drawn from the prior for each population. Updates are proposed one population at a time via a random-walk step, and are accepted or rejected independently.

Well-suited values of tuning parameters (variances in the proposal distributions for $\mu, \theta, \alpha_D, \alpha_E$, and α_2) and the number of generations required to accurately describe the joint posterior will vary from data set to data set, and so may require tuning.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Figure S1. Distribution of Pearson's correlations between each posterior predictive simulated data set and the observed data, highlighting the improved fit of the overdispersion model to describe: (A) the Teosinte data set and (B) the HGDP data set.

Figure S2. Map of teosinte populations sampled, colored by their median estimated population-specific overdispersion parameter, F_k.

Figure S3. Map of human populations included in the analysis, colored by their median estimated population-specific overdispersion parameter, F_k.

Figure S4. Histograms of *P*-values produced by the partial Mantel test (with 1,000,000 permutations) on the 140 data sets for which the true contribution of ecological distance to genetic differentiation was 0.

Figure S5. Trace plots of the α parameters of the covariance matrix Ω.

Figure S6. Joint marginal plots of the α parameters of the covariance matrix Ω, colored by the MCMC generation in which they were sampled.

Figure S7. Acceptance rates for the parameters of the model that are updated with random-walk samplers, plotted over the duration of an individual MCMC run. Dashed green lines indicate the bounds of acceptance rates that indicate optimal mixing: 20%-70%.

Figure S8. Heatmapped matrices showing the performance of the model at all pairwise population comparisons. The posterior predictive p-value was defined as $1\text{-}2\times \mid 0.5\text{-}ecdf(F_{STobs})\mid$, in which ecdf($F_{STobs}$) is the empirical cumulative probability of the observed F_{ST} between two populations from a distribution defined by the posterior predictive sample for that population comparison, representing the p-value of a two-tailed t-test. Higher p-values indicate better model fit. Populations are enumerated on the margins, and may be referenced in SuppMat Table 1. a) The standard model. b) The overdispersion model.

Figure S9. Heatmapped matrices indicating the performance of the model at all pairwise population comparisons. The posterior predictive p-value was defined as $1\text{–}2\times\mid 0.5\text{-}ecdf(F_{STobs})\mid$, in which ecdf($F_{STobs}$) is the empirical cumulative probability of the observed F_{ST} between two populations from a distribution defined by the posterior predictive sample for that population comparison, representing the p-value of a two-tailed t-test. Higher p-values indicate better model fit. Populations are enumerated on the margins, and may be referenced in SuppMat Table 2. a) The standard model. b) The overdispersion model.

Figure S10. Trace plots of the marginal posterior estimates for the α_E/α_D ratio from MCMC analysis of the teosinte dataset. Inset figures give the marginal densities and 95% credible set for the samples after a burn-in of 20% a) The standard model. b) The overdispersion model.

Figure S11. Trace plots of the marginal posterior estimates for the α_E/α_D ratio from MCMC analysis of the HGDP dataset. Inset figures give the marginal densities and 95% credible set for the samples after a burn-in of 20% a) The standard model. b) The overdispersion model.

Table S1. Metadata for populations used in the teosinte data set.

Table S2. Metadata for populations used from the HGDP data set.