Enhanced computational methods for quantifying the effect of geographic and environmental isolation on genetic differentiation

Botta, Filippo; Eriksen, Casper; Fontaine, Michael Christophe; Guillot, Gilles

Published in:
Methods in ecology and evolution

DOI:
10.1111/2041-210X.12424

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 1270–1277 doi: 10.1111/2041-210X.12424

Enhanced computational methods for quantifying the effect of geographic and environmental isolation on genetic differentiation

Filippo Botta[1,2], Casper Eriksen[1], Michaël C. Fontaine[3] and Gilles Guillot[1]*

[1]*Applied Mathematics and Computer Science Department, Technical University of Denmark, Copenhagen, Denmark;* [2]*Now at Centre for Macro-Ecology, Copenhagen University, Copenhagen, Denmark; and* [3]*Marine Evolution and Conservation, Groningen Institute for Evolutionary Life Sciences, University of Groningen, The Netherlands*

Summary

1. In a recent paper, Bradburd et al. (*Evolution*, 67, 2013, 3258) proposed a model to quantify the relative effect of geographic and environmental distance on genetic differentiation. Here, we enhance this method in several ways.

2. We modify the covariance model so as to fit better with mainstream geostatistical models and avoid mathematically ill-behaved covariance functions. We extend the model – initially implemented only for co-dominant bi-allelic markers such as single nucleotide polymorphisms – to encompass highly polymorphic markers such as microsatellites. We implement and test a model selection procedure that allows users to assess which model (e.g. with or without an environment effect) is most suited. We code all our MCMC algorithms in a mix of compiled languages which allows us to decrease computing time by at least one order of magnitude. We propose an approximate inference and model selection method allowing us to deal with genomic data sets (several hundred thousands loci).

3. We also illustrate the potential of the method by re-analysing three data sets, namely harbour porpoises in Europe, coyotes in California and herrings in the Baltic Sea.

4. The computer program developed here is freely available as an R package called SUNDER. It takes as input georeferenced allele counts at the individual or population level for co-dominant markers. Program homepage: http://www2.imm.dtu.dk/~gigu/Sunder/.

Key-words: genomic data, geostatistical model, isolation by distance, isolation by environment, Mantel tests, micro-saetellite, SNPs

Background

The magnitude of gene flow between two populations is expected to relate to the geographical distance between them, a phenomenon known since (Wright 1943) as isolation by distance (IBD). Variation in environmental conditions can also restrict gene flow, a process referred to as isolation by environment (IBE, Wang & Summers 2010; Shafer & Wolf 2013; Sexton, Hangartner & Hoffmann 2014). In their recent review, Wang & Bradburd (2014) list four processes that can potentially generate isolation by environment: biased dispersal, natural selection against immigrants, sexual selection against migrants and reduced hybrid fitness. Disentangling the role of geographic distance and environmental heterogeneity in shaping genetic variation is a critical issue in landscape genetics studies. This can help to understand better micro-evolutionary processes towards incipient speciation and to address more practical questions involved in populations management and conservation decisions.

So far, this question has mainly been addressed using partial Mantel tests, which can lead to erroneous conclusions in the presence of autocorrelation (Guillot & Rousset 2013). In a recent paper, Bradburd, Ralph & Coop (2013) proposed an alternative method based on a geostatistical model, which does not suffer from the flaw affecting the partial Mantel test. In the latter approach, the key modelling ingredient is a covariance matrix model that summarises individual or population pairwise genetic variation. It assumes that covariance decays in an exponential fashion as a function of geographic and environmental distances. The main output of the method is an estimate of two parameters that quantify how genetic covariance relates to geographic and environmental dista nces. The method is implemented in the R package BEDASSLE (Bradburd 2013) and has been used for example by Bradburd, Ralph & Coop (2013) to analyse human and teosinte data and by Harvey & Brumfield (2014) to analyse tropical bird data.

*Correspondence author. E-mail: gilles.b.guillot@gmail.com

In their conclusion, Bradburd, Ralph & Coop (2013) made the wish that users would elaborate on the framework they presented. Here, we take up this assignment and propose to enhance their method in several ways: (i) we modify the covariance model to fit better with mainstream geostatistical models, (ii) we extend the model – initially implemented only for co-dominant bi-allelic markers such as single nucleotide polymorphisms (SNPs) – to encompass highly polymorphic markers such as microsatellites, (iii) we implement a model selection procedure that allows users to assess which model (e.g. with or without an environment effect) is most suited, (iv) we code all our algorithms in a mix of C and Fortran language which allows us to decrease computing times significantly, (v) we propose an approximate inference and model selection method allowing to deal with large genomic data sets consisting of millions of loci now more and more frequent in model and non-model species. The next section presents our models and algorithm, they are partly reminiscent of Wasser *et al*. (2004), Guillot & Santos (2009) and Bradburd, Ralph & Coop (2013), but for the sake of clarity, we attempt to give a self-contained description here and we list in Supporting information the detail of similarities and differences between our program called SUNDER and the BEDASSLE program. The remaining part of the paper is devoted to the analysis of data simulated under two different models and to the re-analysis of three previously published data sets.

Methods

STATISTICAL MODELS

Model with binomial/multinomial distribution

Genotype and allele frequency model: We assume that the data at hand are a collection of allele counts over groups of individuals (or possibly a single individual) observed at various geographic locations and we denote by g_{sla} the count of alleles of type a, at locus l observed at geographical location s and by n_{sl} the haploid sample size at geographical location s for locus l ($n_{sl} = 2$ if a single diploid individual is observed at site s and genotyped at locus l). A_l denotes the number of alleles observed for locus l (usually $A_l = 2$ for SNPs, more for microsatellite markers). We denote by f_{sla} the frequency of allele a at locus l in a population located at geographical site s. We assume that the alleles observed at location s form a random sample of the underlying local population which we assume to be at Hardy–Weinberg equilibrium. For co-dominant markers, this translates into the assumption that allele counts are multinomials (in particular binomials for bi-allelic markers), which we denote $(g_{sl1}, ..., g_{slA_l}) \sim Multinom(n_{sl}; f_{sl1}, ..., f_{slA_l})$. To comply with standard statistical genetics models, we assume that the vector $(f_{sl1}, ..., f_{slA_l})$ follows a $Dirichlet(\alpha,...,\alpha)$ distribution, where α is an unknown parameter that controls the variance of allele frequencies. This extends the beta distribution for bi-allelic loci assumed in the BEDASSLE program and makes the global model suitable for the analysis of microsatellite markers. We assume that allele frequencies are independent across loci but autocorrelated in space. To model this, we assume that a vector $(f_{sl1}, ..., f_{slA_l})$ is equal – up to a deterministic transform – to a vector of Gaussian random fields $(y_{sl1}, ..., y_{slA_l})$. The various components for $a = 1,...,A_l$ of this vector are mutually independent but each component y_{sla} is spatially autocorrelated. See Guillot & Santos (2009) for details.

Covariance model: Denoting by h_D the geographical distance between sites s and s' and h_E the environmental distance between sites s and s', we consider that

$$\mathrm{Cov}(y_{sla}, y_{s'la}) = C(h_D, h_E) = \exp[-(h_D/\beta_D + h_E/\beta_E)^\gamma] \qquad \text{eqn 1}$$

In the above, β_D and β_E are unknown parameters that have the dimension of a geographic distance and of an environmental distance, respectively. They quantify the magnitude of the effect of these two variables on genetic covariance. Large values of the β_D (resp. β_E) parameter correspond to a slow decay of the covariance as h_D (resp. h_E) increases, i.e. a small influence of geographical (resp. environmental distance). Two limiting cases are worth noting: $\beta_D = +\infty$ would correspond to a situation of panmixia and $\beta_D = 0$ would correspond to a situation of complete geographical isolation of the various populations. We warn against a hurried interpretation of the β_D parameter: even though β_D has the dimension of a geographical distance, it cannot be interpreted straightforwardly as a demographic parameter such as an average dispersal distance. Strictly speaking, β_D solely describes the rate of decay of the covariance in space which relates not only to the average dispersal distance but also to population density and migration rates (cf e.g. Rousset, 1997,, 2001). Besides, the system may not be in migration–drift equilibrium (e.g. due to recent expansions), which may affect the estimate of β_D. The parameter γ is a dimensional and quantifies the smoothness of spatial variation of the hidden variables y and therefore of the allele frequencies f. Anticipating on the analysis of the harbour porpoise (*Phocoena phocoena*) data that comes below, we invite the reader to take a look at Fig. 2 which illustrates the main pattern captured by eqn 1: the spatial correlation decays with geographical distance, but the decay is specific to the environmental distance between populations.

Covariance model with nugget effect In the model defined by eqn 1, the correlation becomes arbitrarily close to one when both the geographical and the genetic distance become arbitrarily close to zero. In other words, the model defined by eqn 1 implies that nearby populations cannot exhibit any large genetic difference. As noted by Bradburd, Ralph & Coop (2013), this property might conflict with certain data, for example in case of local introduction, secondary contact, barrier to gene flow, where some pairs of geographically close populations can exhibit a high level of genetic differentiation. To handle this, we modify the model of eqn 1 into

$$C(h_D, h_E) = \delta I_0(h_D) + (1 - \delta)\exp[-(h_D/\beta_D + h_E/\beta_E)^\gamma] \qquad \text{eqn 2}$$

The term $I_0(h_D)$ is equal to 1 when $h_D = 0$ and 0 otherwise. It is known as a nugget effect in the geostatistical literature (Cressie & Wikle 2011, pp. 122–123), and it is used to introduce a discontinuity of the covariance function at $h_D = 0$. Including a nugget effect in the covariance function amounts to assuming that the variable considered is the sum of spatially unstructured term (random noise) and a spatially structured term. It is used in geostatistics to account for measurement errors or as an expedient to model variation taking place at a spatial scale smaller than that observable with the data. Here, it is used with an alternative goal in mind: it allows us to model departure from a strict IBD process at equilibrium and to take into account empirical covariance structures with potential large genetic differences between pairs of geographically closely located populations.

Covariance model for several environmental variables: We also extend the covariance structure described by eqn 2 to handle the case where a combination of environmental variables $E_1,...,E_J$ may explain jointly the covariance structure. Denoting a vector of p environmental distances $(h_{E_1}, ..., h_{E_p})$ by $\mathbf{h}_E$ we consider:

$$c(h_D, \mathbf{h}_E) = C(h_D, h_{E_1}, ..., h_{E_p}) = \delta I_0(h_D) + (1-\delta)\times \exp\left[-\left(h_D/\beta_D + \sum_j h_{E_j}/\beta_{E_j}\right)^{\gamma}\right] \quad \text{eqn 3}$$

Covariance model with geographic distance only: The generic model of eqn 2 can also be simplified and used to investigate spatial genetic variation in absence of any obvious environmental factor. By dropping h_E (or setting $\beta_E = +\infty$) in eqn 2, we get:

$$C(h_D) = \delta I_0(h_D) + (1-\delta)\exp[-(h_D/\beta_D)^{\gamma}] \quad \text{eqn 4}$$

Making inference about remaining parameters in eqn 4 allows one to estimate the spatial rate of decay of the genetic covariance. Also, comparing estimates of the β_D parameter across populations observed in different environmental conditions can help to better understand how environmental heterogeneity impacts genetic variation.

*Covariance model with environmental distance only:*Finally, the covariance structure of eqn 2 can be used to investigate spatial genetic variation at a scale where no isolation by distance pattern is expected. By dropping h_D (or setting $\beta_D = +\infty$) in eqn 2, we get :

$$C(h_E) = \delta I_0(h_D) + (1-\delta)\exp[-(h_E/\beta_E)^{\gamma}] \quad \text{eqn 5}$$

Model with Gaussian distribution

Regarding the model outlined above, we provide evidence in subsequent sections that an MCMC algorithm for inference coupled with cross-validation for model selection works well for a number of markers L in the range $L = 100-1000$. However, for next-generation sequencing data consisting of up to a million of SNP loci, an MCMC-based approach becomes impractical. Fortunately, for data sets where sampling units consist of a sufficiently large number of individuals, and SNP loci being mostly bi-allelic, the allele counts (assumed to be binomial in our initial model) can be approximated by a Gaussian distribution. In this case, we identify the set of allele counts g_{sla} to the set of hidden Gaussian variables y_{sla}. Doing so, we skip the intermediate layer of allele frequencies f_{sla} and simply assume that the allele counts g_{sla} are approximately multivariate Gaussian. The covariance matrix is assumed to be derived from the same functional expression as before (eqn 3). Under this approximate model, the parameters have a slightly different meaning as they bear on g rather than on some hypothetical allele frequencies. However, this model still allows us to quantify the relative magnitude of the effect of geographic versus environmental isolation. For the Gaussian approximation to a binomial distribution, a haploid sample size larger than 30 seems to be a minimum. Differences in local sample sizes can be accommodated straightforwardly by working with allele frequencies rather than allele counts.

PARAMETER INFERENCE AND MODEL SELECTION

Restrictions on parameters

We focus here on the model described by eqn 2. The vector of unknown parameters is $\theta = (\alpha, \beta_D, \beta_E, \gamma, \delta)$. Covariance functions enjoy a mathematical property known as positive definiteness which mirrors the fact that a variance is always positive. To satisfy this property, the range of the γ parameter has to be restricted to [0,1] when geographical distances are measured as straight line distances in the plane (Guillot *et al.* 2014). When geographical distances are geodesic on the sphere, the mathematical conditions under which this covariance model is well behaved mathematically are not known beyond the case $\gamma = 1$. The same remark applies when using more than one environmental variables (eqn 3). In this context, we recommend treating γ as a fixed parameter equal to 1 which guarantees positive definiteness. The other parameters do not bring any difficulty: $\alpha \in [0,+\infty)$, $\beta_E \in [0,+\infty)$, $\beta_D \in [0,+\infty)$, $\delta \in [0,1)$.

Prior distribution and inference for model with binomial/multinomial distribution

The data consist of allele counts for alleles $a = 1,...,A_l$, at loci $l = 1,...,L$ over populations $i = 1,...,n$ denoted $g = (g_{ila})$. The vector of unknown parameters is $\theta = (\alpha,\beta_D,\beta_E,\gamma,\delta)$ and we denote by f_{il} the set of underlying allele frequencies. We aim at simulating from the posterior density $p(\theta|g) \propto p(g|\theta)p(\theta)$. This involves the sampling distribution $p(g|\theta)$ that can be expressed as $\int p(g|f,\theta)p(f|\theta)df$ and does not have any analytically tractable expression. We therefore simulate jointly from $p(\theta,f|g) \propto p(g|f,\theta)p(f|\theta)p(\theta)$ which involves only tractable probability distributions. We place independent uniform priors on each component of θ, to do so we choose upper bounds for the α, β_D and β_E that are large enough to make the choice un-consequential. We perform Metropolis-within-Gibbs simulation alternating updates of f and updates of θ. In the updates of f, there is no obvious appealing proposal distribution on the frequencies f themselves, so we follow the suggestion of Wasser *et al.* (2004). It consists in adding increments on the independent Gaussian variables x defined in the transform $y = Lx$ where L is the lower triangular matrix in the Cholesky factorisation of the covariance matrix Σ. In the updates of θ we perform Metropolis-within-Gibbs updates with component-wise moves. The steps we take to do so follow Wasser *et al.* (2004) and Bradburd, Ralph & Coop (2013) to a large extent. See Supporting information for illustration of the behaviour of our MCMC algorithm.

Inference for model with Gaussian distribution

Assuming that allele counts are $MVN(\mu,\Sigma_0)$, we estimate θ by maximising the Gaussian likelihood $p(g|\theta)$. We do this with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

Model selection

Here, we are concerned with the selection of the best submodels among $M_{G+E}:\{\beta_D<+\infty,\beta_E<+\infty\}$, $M_G:\{\beta_D<+\infty,\beta_E=+\infty\}$, $M_E:\{\beta_D=+\infty,\beta_E<+\infty\}$, defined by their covariance structure defined, respectively, by eqns 2, 4 and 5. An approach based on maximising the likelihood on the whole data set is obviously incorrect as models M_G and M_E are embedded in M_{G+E}. The latter model would therefore necessarily achieve the highest likelihood. To avoid this issue, we base our method on cross-validation (CV) as follows: we split the data set into a training set (a random subset of locations × loci) and a validation set (the remaining data points). The reason for defining the training set as a combination of geographical locations and loci [in contrast with (i) a subset of loci at all locations or (ii) a subset of locations at all loci] is related to the structure of the model. With a training set as in (i), because we assume independence across allele frequencies, it would be impossible to predict allele frequencies at loci of the validation set. A training set as in (ii), although easy to implement in practice, would amount to downgrade greatly the density of the spatial sampling and would bring results that would not reflect the actual data set but that of data set characterised by a lower spatial sampling density. Our strategy in defining the training set is an attempt to find a trade-off between degrading the spatial and the genetic sampling in the training data set.

We use the training set to make inference on the parameters and hidden variables y under the three submodels. This provides us with an estimate of the y variables (and therefore the allele frequencies after a deterministic transform) for all combinations (s,l) of the validation set. This estimate is the posterior mean obtained by MCMC for the model with binomial marginal and the maximum likelihood for the Gaussian approximation.

Then, we plug these estimates in the likelihood function to evaluate the probability of the validation set. These two steps (inference and evaluation of the likelihood for the validation set) are performed for each of the three competing submodels. The model selected is the one achieving the highest probability. The efficiency of this approach is illustrated in the next section, see also Supporting information for further illustration of our cross-validation procedure.

Summary of main program outputs and computing times

The SUNDER program performs parameter inference (by MCMC simulation or likelihood optimisation) and optionally cross-validation on any of the submodels listed in section Model selection. This provides the users with a point estimate of θ (the posterior mean) under each submodel but also a score quantifying which submodel explains the data best. Bayesian inference and model selection on a data set with $n = 100$ sampling sites and $L = 1000$ loci takes typically an hour on a standard PC. The same task is performed in a few seconds under the Gaussian approximation.

Analysis of simulated data

GEOSTATISTICAL SIMULATIONS/BAYESIAN INFERENCE

Our first investigation consists in analysing data simulated according to the exact model with multinomial distribution (described in section Model with binomial/multinomial distribution and referred hereafter to as 'geostatistical model', an approach taken, for example by Novembre & Stephens 2008). We considered three types of structures for the covariance matrix: with effect of both geographic and environmental distances (G+E), effect of geographic distance only (G) and effect of environmental distance only (E). These covariances correspond to eqns 2, 4 and 5. We generated 100 data sets for each of the three models with populations located at 50 geographical sites consisting each of 10 diploid individuals genotyped at 100 SNP loci and then at 1000 SNP loci. Also, because two of the real data sets reanalysed below contain a small number of microsatellite loci, we also investigate simulations at 15 loci with 10–20 alleles per locus, and similar to simulations above in all other respects.

The locations of the geographical sites were sampled uniformly in a $[0,1]\times[0,1]$ square, and the environmental variable was sampled independently from a uniform discrete distribution with three states that mimics, for example the spatial patchy distribution of three habitats. We also considered the case where the environmental variable is continuous and spatially autocorrelated. In this case, it was simulated as a centred and standardised Gaussian variable with an exponential covariance function with parameter scale equal to 0·3. All simulations of genotypes were carried out with the same set of parameters for the covariance matrix, namely $\alpha = \beta_G = \beta_E = \gamma = 1$ and $\delta = 0{\cdot}01$. For these data, we performed Bayesian inference and model selection under the model with multinomial (or binomial) likelihood.

COALESCENT SIMULATIONS

We also simulated data under an isolation by distance model using coalescent simulation with the IBDSIM program (Leblois, Estoup & Rousset 2009). To produce data under conditions that mimic a purely geographic model (referred to as G model above), we produced simulations on a 30×30 grid with 20 diploid individuals per grid node, we took as dispersal distribution a truncated Pareto distribution (probability of moving k steps $\propto M/k^n$ with $M = 0{\cdot}82$, $n = 4{\cdot}11$ and an upper bound equal to 48) and set the migration rate equal to 0·03. To produce data under a G+E model, we simulated two independent data sets by two independent IBDSIM runs at 25 geographical sites each, both with the same parameters as the G model described above. Then, we merged the two sub-data sets together on a square so as to mimic the coexistence of two subpopulations genetically isolated by an impermeable barrier. To generate data under an E model, we did the same as in the G+E case, except that we set the migrations rate equal to 0·999. Here, we generated genotypes at 1000 independent loci. In a last step, we also simulated data as in the G+E and E cases but picked 4% of the individuals in each population and swap them to mimic F0 migrants. In this case, genotypes were simulated at 100 SNP loci. In all cases, we subsampled 50 of the initial 900 populations to produce a data set at 50 irregularly spaced sampling sites. For these data, we carried out Bayesian inference and model selection under the model with binomial likelihood.

Results on model selection based on the Bayesian model with binomial/multinomial distribution are summarised in Table 1. In the conditions studied, our algorithm is able to retrieve the true model except in a small fraction of cases where the algorithm is too permissive: the true model is G or E and the algorithm selects G+E. The accuracy in model selection increases with the number of loci used, with only a handful of model selection error out of 300 simulated data sets for $L = 1000$ loci.

ASSESSING THE VALUE OF THE GAUSSIAN APPROXIMATION

To assess the value of the Gaussian approximation model, we simulated data under the model with binomial likelihood described in section Model with binomial/multinomial distribution but made inference under the approximate model and algorithm described in sections Model with Gaussian distribution and Inference for model with Gaussian distribution. We considered various numbers of geographical locations n ranging from 50 to 500 and a number of loci L ranging from 100 to 100 000. In all cases, the local haploid sampling size was equal to 2 (a single diploid individual). The environmental variable was continuous and spatially autocorrelated. The results are summarised on Fig. 1, where it is clear that the Gaussian approximation performs well as soon as the number of loci is large.

Table 1. Results of model selection on simulated data. In each subtable, a value of 100% on the diagonal indicates a perfect result

True model\ Selected model	G+E	G	E
Geostatistical simulations, discrete environmental variable			
	Bi-allelic loci L=100		
G+E	100	0	0
G	14	86	0
E	0	0	100
	Bi-allelic loci L=1000		
G+E	100	0	0
G	0	100	0
E	0	0	100
Geostatistical simulations, continuous environmental variable			
	Highly polymorphic loci L=15		
G+E	99	1	0
G	32	68	0
E	29	0	71
	Bi-allelic loci L=100		
G+E	100	0	0
G	16	84	0
E	7	0	93
	Bi-allelic loci L=1000		
G+E	100	0	0
G	1	99	0
E	0	0	100
IBDSIM simulations, discrete environmental variable			
	Bi-allelic loci L=1000		
G+E	100	0	0
G	7	93	0
E	0	0	100
	Bi-allelic loci L=100		
G+E with F0 migrants	95	5	0
G	41	55	4
E with F0 migrants	9	0	91

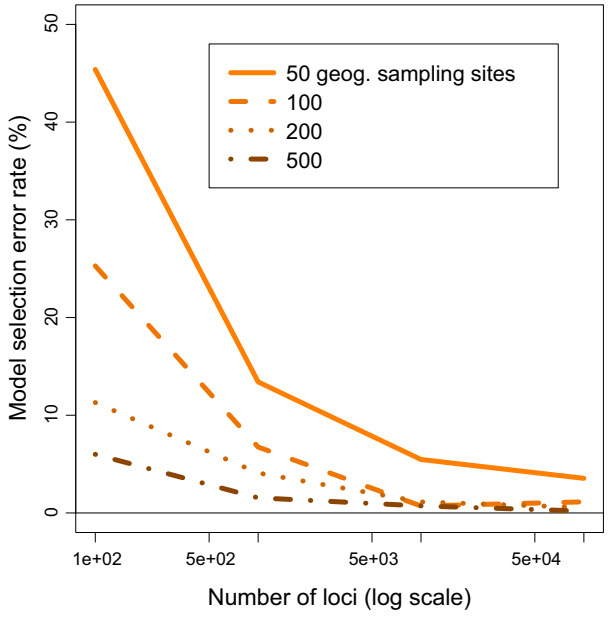

Fig. 1. Results of model selection. Data from geostatistical simulations with allele counts sampled from a binomial distribution. Inference carried out under the approximate Gaussian model.

Analysis of real data

HARBOUR PORPOISE DATA

We re-analyse here a data set consisting of genotypes at 10 microsatellite loci for 752 harbour porpoises (*P. phocoena*) sampled across the North Atlantic continental shelf area in Europe and the Black sea that was initially studied by Fontaine *et al.* (2007) . Among other findings, this initial study conjectured the existence of a sharp genetic discontinuity between the Northern Atlantic samples and the remaining Atlantic samples off the Iberian coast. This is graphically illustrated by Fig. 2, which displays variation of pairwise genetic correlation as a function of pairwise geographic distances. Figure 2 clearly shows that variation in pairwise correlations is not simply explained by geographical distances and suggests that there is a genetic discontinuity among clusters, in particular between Iberia and North Atlantic clusters, which should be therefore roughly located over the Bay of Biscay. Fontaine *et al.* (2007) linked this genetic discontinuity to sharp variation of environmental conditions in the Bay of Biscay.

Here, we re-analyse this data set to investigate further the existence of an IBE process. However, because the Black Sea populations underwent a specific recent history and are geographically isolated by obvious landscape features, we do not include the Black Sea samples in our analysis (Fontaine *et al.* 2012). Also, in their study, Fontaine *et al.* (2007) measured geographical distances as distances along shortest marine path. Because this could bring up mathematical difficulty in the covariance model (Guillot *et al.* 2014), we use straight line distances with planar coordinates. Lastly and although our

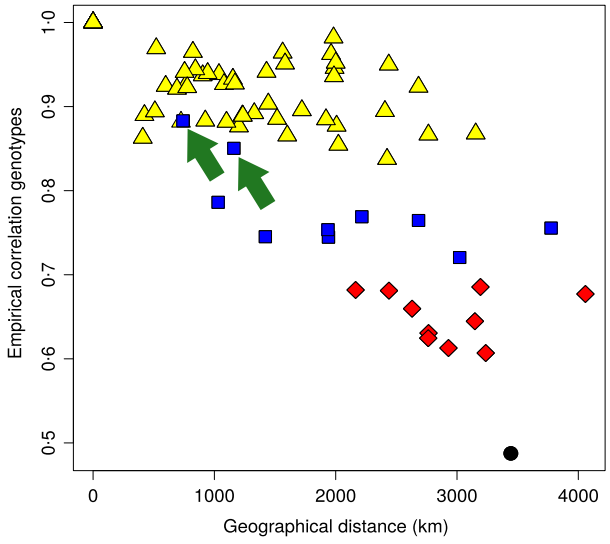

Fig. 2. Pairwise genetic correlation among sampling units as a function of straight line distance. The colour refers to the genetic cluster memberships of the sampling units involved in each pair (estimate from Fontaine *et al.* 2007). Yellow triangles: pairs of population belonging to the same clusters, blue squares: Iberia/North Atlantic, red dots: Black Sea/North Atlantic, black dot: Black Sea/Iberia. Green arrows point towards the pairs of sites Iberia–Ireland and Iberia–Gascony. See Supporting information for details.

method can handle any sample size (including individual genotype data), for consistency with fig. 6 in Fontaine *et al.* (2007), who pooled some populations together to avoid small sampling size (see Supporting information for detail). In the first place, we used as environmental variable a dummy variable taking values 0/1 and encoding the membership to the genetic clusters inferred by Fontaine *et al.* (2007) (North Atlantic/Iberia). We used this dummy variable as a surrogate for a combination of unobserved real environmental variables and performed model selection among the models G, E and G+E. We launched ten MCMC runs of 10^7 iterations. There was no consensus between these runs but out of these ten runs, the model selected most often was *G+E*. This adds weight to the initial conjecture of Fontaine *et al.* (2007) about the existence of a genetic discontinuity between Iberia and North Atlantic. We also carried out similar MCMC runs after removing data from sites from the Bay of Biscay and the Celtic Sea. These sites display a significant amount of admixture Fontaine *et al.* (2010) and Fontaine *et al.* (2014), which may act as a confounder in our analysis (these pairs of populations can be identified on Fig. 2, see also Supporting information). In this second analysis, there is still no consensus across the ten runs, but the model that is now selected most often is E.

North of the Bay of Biscay, Fontaine *et al.* (2007) also observed variation in the IBD strength, which could result from spatial variations in effective population density and/or dispersal pattern. We replicated this analysis using SUNDER to show its capability to address such question. To do so, we estimated parameters under the 'G' model for North Atlantic subareas 2A, 2B, 3A,3B, 3C defined by their latitudes (see Fig. IX, Supporting information). The results of inference for β_D are as follows: 2A: $\widehat{\beta}_D$ = 10392km , 2B: $\widehat{\beta}_D$ = 37386km, 3A: $\widehat{\beta}_D$ = 12939km, 3B: $\widehat{\beta}_D$ = 15627km, 3C: $\widehat{\beta}_D$ = 32880km. In words, genetic similarity decays with geographical distance at a faster pace in the south than in the north. This is consistent with the findings of Fontaine *et al.* (2007) based on the moment based method of Rousset (1997, 2000).

COYOTE DATA

We considered data previously analysed by Sacks *et al.* (2008) consisting of genotypes at 14 autosomal microsatellite loci of 1828 coyotes (*Canis latrans*) sampled in California (USA) in a region including two distinct ecoregions: the California Floristic Province (CFP) and the Desert–Prairie ecoregion (DPE). The CFP ecoregion displays a heterogeneous landscape while the DPE ecoregion displays a homogeneous landscape. Sacks *et al.* (2008) found evidence that coyotes sampled from the CFP exhibit genetic structure concordant with habitat subdivisions, while coyotes from widely dispersed sampling sites within the homogeneous DPE exhibit little or no structure. We analysed this data set using the model with multinomial distribution. To estimate an ecoregion-specific scale parameter for our covariance model, we only considered the geographical distance in our analyses (hence using a G model) and performed runs independently for the two ecoregions. Doing so, we obtained an estimate $\widehat{\beta}_D$ of approximately 780 kms for the CFP ecoregion and 5638 kms for the DPE ecoregion. These result confirm the findings of Sacks *et al.* (2008) and allow to further quantify the magnitude of habitat heterogeneity effect on coyote populations. Indeed, the decorrelation distance is reduced by a factor of approximately 7 when comparing the subdivided CFP region to the homogeneous DPE region.

HERRING DATA

Lastly, we re-analysed a data set consisting of allele counts at 440 817 SNP loci for 400 herrings sampled at eight locations in the Baltic sea and the North Sea analysed by Lamichhaney *et al.* (2012). In these data, the haploid sampling size is equal to 100 for each sampling site and the Gaussian approximation of the Binomial distribution was used. For this spatial sampling, straight line distances are not appealing as they amount to disregard the Scandinavian peninsula land mass between the Baltic Sea and the North Sea. We use distances measured as straight lines along the coast line which amounts to assuming a linear habitat. Following Lamichhaney et al.(2012), we consider salinity as a potential driver of genetic differentiation and perform again model selection with models G, E and G+E.

The model that provided the best fit among E,G and E+G in the cross-validation procedure was E, which corresponds to an absence of a significant isolation by distance pattern and an effect of salinity on genetic differentiation.

Discussion

We have modified and extended the model proposed by Bradburd, Ralph & Coop (2013) in order to make it fit better with traditional geostatistical models and avoid issues related to positive definiteness. We have proposed a statistical model selection method that allows users to go beyond posterior distributions and provides them with a decision criterion as to what model describes best the data. We have also implemented the MCMC inference corresponding to this updated model (and various submodels) in a mix of C and Fortran code. This code is wrapped in an R package available from the Comprehensive R Archive Network called SUNDER. Implementing the main MCMC loop in Fortran allows us to decrease computing times typically by a factor 20 on a data set consisting of about 100 loci and 50 populations. The model selection procedure proves to work well in the conditions investigated. The numerical values reported have to be taken with a grain of salt as they correspond to some best case scenarios where the model assumed in inference complies well with the data-generating process. For data sets consisting of 1000 loci, one could have been worried about MCMC convergence issues. The results about model selection show that there is no major MCMC convergence issue here (see Supporting Information for examples of MCMC runs) and suggest that the algorithm is still well behaved for even larger data sets. We stress also that the results reported here are based on a single MCMC run, in particular we did not experience any of the MCMC run failures reported by Bradburd, Ralph & Coop (2013). In a large majority of cases, erroneous model selection results consist of a preference

for G+E when the true model is either G or E. This preference for the most complex model has been observed in a populations genetics context (see for an example Alexander & Lange 2011) and likely results from the absence of penalisation for the number of model parameters in the cross-validation strategy. In our results, this issue affects more markedly simulations performed under the G model than under the E model. In the covariance models considered here, geography and environment play formally a completely similar role (cf the symmetry in h_G and h_E in eqns 1–5). Therefore, under the family of inference models considered here, there is intrinsically no greater algorithmic difficulty to estimate an IBD effect than an IBE effect. The asymmetry in G and E observed in Table 1 has to result from the specific simulations conditions studied here. In all the geostatistical simulations involving the α_G or the α_E parameter, their values were set equal to one. However, to avoid redundancy of simulations under the G and E scenarios, in our simulations, the distribution of values of the geographic and environmental distances was not the same, the former spreading typically across a broader range than the latter. Turning the results of our simulations into a rule to assess the likelihood to detect a spurious effect would be certainly useful but is practically out of reach as this rule would have to depend on the effect size which is precisely one of the quantity that our model attempts to estimate.

The nugget coefficient δ (eqns 2–5) controls how much an allele frequency at a given location will depart from those at neighbouring locations. In our approach, this coefficient is shared across all populations. This contrasts with the model implemented in BEDASSLE. In the latter approach, there is an over-dispersion model where a parameter accounting for departure from the binomial distribution (and which can be related to a population inbreeding coefficient) plays a role similar to that of our nugget effect. In BEDASSLE, this parameter is population-specific and estimated for each population. The latter approach allows therefore more flexibility in the way population-specific events in population histories (e.g. unequal population sizes, bottlenecks) can be encompassed and understood. In addition to a slightly more flexible modelling framework on this aspect, the BEDASSLE program provides users with a model-fit diagnostic tool based on comparing data to posterior predictive simulations. Such plots can be informative but cannot be implemented in our framework since our Gaussian approximation is based on a pure likelihood approach. Also we believe that the interest of the present covariance-based modelling approach resides in the model selection strategy it offers rather in its ability to fit data. We re-analysed three previously published data sets and could confirm earlier findings on the basis of objective criteria and support conclusions with quantitative facts. On coyote data, we confirm findings of Sacks *et al.* (2008) on specialisation of coyotes by ecotypes. On the herring data, we confirm findings of Lamichhaney *et al.* (2012) about the role of salinity. In the model selection procedure, the combinations (s,l) that define the training and validation sets are randomly chosen. This implies that even under perfect MCMC convergence, the outcome of the model selection procedure (that is based on an MCMC run on a training set and an evaluation of the likelihood at the validation set) remains random and therefore subject to variation from one run to another. This does not appear to be an issue in our analysis of simulated data (cf good results obtained in terms of model selection accuracy in Table 1). In the analysis of the porpoise data, the results from SUNDER clearly support previous findings; however, we faced inconsistencies across MCMC + CV runs several times. This is likely due to the small number of loci and therefore rather inherent to the lack of information in the data than to a genuine weakness of the method.

Guillot & Rousset (2013) showed that the Mantel test and its widely used alternative, the partial Mantel test was flawed when one interprets the p-values as a measure of the significance of the correlation between two spatially autocorrelated variables. This result was also confirmed by Bradburd, Ralph & Coop (2013). This is because the permutation procedure is incorrect in the presence of spatial autocorrelation, and the *P*-values returned are not well calibrated. This often leads to the detection of spurious correlation. The results we report in section Analysis of simulated data suggest that the present method is not prone to this issue.

For example, in the herring data set, there is a clear correlation between the location along the Scandinavian peninsula coastline and the salinity, with increasing salinity from the North of the Baltic Sea (3%) to the North Sea (35%). This means that one of these two variables could act as a confounding factor when analysing the effect of the other one on genetic differentiation. Here, we are able to analyse jointly the effect of these two variables and can conclude on the presence of an effect of salinity only. A similar situation was encountered in the geostatistical simulation of a continuous environmental variable (see Table 1), and this does not affect the accuracy of the method. These results add weight to the idea that our method is a useful alternative to the partial Mantel test.

Acknowledgements

We are grateful to Ben Sacks and Sangeet Lamichhaney for making the coyote and herring data available and to the Associate Editor and two referees for helpful comments.

Funding

This work was supported by the Danish e-Infrastructure Cooperation (DeIC) and the US National Institute for Mathematical and Biological Synthesis working group on Computational Landscape Genomics.

Data accessibility

Data deposited in the Dryad repository: http://datadryad.org/resource/doi:10.5061/dryad.r2rn0

References

Alexander, D. & Lange, K. (2011) Enhancements to the ADMIXTURE algorithm for individual 432 ancestry estimation. *BMC Bioinformatics*, **12**, 246.

Bradburd, G. (2013). *R Package 'BEDASSLE'. Comprehensive R Archive Network*.
Bradburd, G., Ralph, P. & Coop, G. (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, **67**, 3258–3273.
Cressie, N. & Wikle, C. (2011) *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.
Fontaine, M., Baird, S., Piry, S., Ray, N., Tolley, K., Duke, S. et al. (2007) Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters. *BMC Biology*, **5**, 30.
Fontaine, M. C., Tolley, K. A., Michaux, J., Birkun, A., Ferreira, M., Jauniaux, T., Llavona, A. et al. (2010) Genetic and historic evidence for climate-driven population fragmentation in a top cetacean predator: the harbour porpoises in European water. Proceedings of the Royal Society of London. : *BMC Biology*.**277**, 2829–2837.
Fontaine, M., Snirc, A., Frantzis, A., Koutrakis, E., Öztürk, B., Öztürk, A. A. & Austerlitz, F. (2012) History of expansion and anthropogenic collapse in a top marine predator of the Black Sea estimated from genetic data. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E2569–E2576.
Fontaine, M., Roland, K., Calves, I., Austerlitz, F., Palstra, F., Tolley, K. et al. (2014). Postglacial climate changes and rise of three ecotypes of 444 harbor porpoises, Phocoena phocoena, in western Palearctic waters. *Molecular Ecology*, **23**, 3306–3321.
Guillot, G. & Santos, F. (2009) A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies. *Molecular Ecology Resources*, **9**, 1112–1120.
Guillot, G. & Rousset, F. (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution*, **454**, 336–344.
Guillot, G., Schilling, R., Porcu, E. & Bevilacqua, M. (2014) Validity of covariance models for the analysis of geographical variation. *Methods in Ecology and Evolution*, **5**, 329–335.
Harvey, M. & Brumfeld, A. (2014) Genomic variation in a widespread Neotropical bird (*Xenops minutus*) reveals divergence, population expansion, and gene flow. arXiv preprint arXiv:1405.6571.
Lamichhaney, S., Barrio, A., Rafati, N., Sundström, G., Rubin, C., Gilbert, E. et al. (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19345–19350.
Leblois, R., Estoup, A. & Rousset, F. (2009) IBDsim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*, **9**, 107–109.
Novembre, J. & Stephens, M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.
Rousset, F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.
Rousset, F. (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology*, **13**, 58–62.
Rousset, F. (2001) Inferences from spatial population genetics. *Handbook of Statistical Genetics* (eds D.J. Balding, M. Bishop & C. Cannings), pp. 239–269. John Wiley & Sons, Chichester.
Sacks, B., Bannasch, D. L., Chomel, B. B. & Ernst, H. (2008) Coyotes demonstrate how habitat specialization by individuals of a generalist species can diversify populations in a heterogeneous ecoregion. *Molecular Biology and Evolution*, **25**, 1354–1395.
Sexton, J., Hangartner, S. & Hoffmann, A. (2014) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution,* **68,** 1–15.
Shafer, A. & Wolf, J. (2013) Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecology Letters*, **16**, 940–950.
Wang, I. J. & Bradburd, G. S. (2014) Isolation by environment. *Molecular Ecology*, **23**, 5649–5662.
Wang, I. J. & Summers, K. (2010). Genetic structure is correlated with phenotypic divergence rather than geographic isolation in the highly polymorphic strawberry poison-dart frog. *Molecular Ecology*, **19**, 447–458.
Wasser, S., Shedlock, A., Comstock, K., Ostrander, E., Mutayoba, B. & Stephens, M. (2004) Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14847–14852.
Wright, S. (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Received 10 May 2015; accepted 9 June 2015
Handling Editor: Douglas Yu

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Analysis of geostatistical simulations; porpoise data; and similarities and differences between BEDASSLE and SUNDER programmes.

Fig. S1. MCMC inference with 10^5 iterations.

Fig. S2. MCMC inference with 105 iterations.

Fig. S3. Trace of three independent MCMC runs for cross validation, with 10^5 iterations.

Fig. S4. Trace of some allele frequencies f_{sla} for the MCMC run displayed in Fig. S3.

Fig. S5. Trace of three independent MCMC runs for cross validation, with 10^5 iterations.

Fig. S6. Trace of some allele frequencies f_{sla} for the MCMC run displayed in Fig. S5.

Fig. S7. Trace of three independent MCMC runs for cross validation, with 10^5 iterations.

Fig. S8. Trace of some allele frequencies f_{sla} for the MCMC run displayed in Fig. S7.

Fig. S9. Location of sampling locations for the harbour porpoise data, reprinted from Fontaine et al. (2007) with permission.

Fig. S10. Geographic locations of the porpoise samples.

Fig. S11. Geographic locations of the populations after pooling.

Fig. S12. Pairwise genetic differentiation against genetic correlation among sampling units.

Table S1. Similarities and differences between the BEDASSLE and the SUNDER programs.