Available online at www.sciencedirect.com
ScienceDirect
Theoretical Population Biology 71 (2007) 332–341
Theoretical Population Biology
www.elsevier.com/locate/tpb

The Genetical Bandwidth Mapping: A spatial and graphical representation of population genetic structure based on the Wombling method

A. Cercueil[a], O. François[a], S. Manel[b,*]

[a]TIMC, Institut National Polytechnique de Grenoble, UJF-CNRS UMR 5525, TIMB Faculté de Médecine, F38706 La Tronche, France
[b]Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France

Received 1 June 2006
Available online 4 February 2007

Abstract

Characterizing the spatial variation of allele frequencies in a population has a wide range of applications in population genetics. This article introduces a new nonparametric method, which provides a two-dimensional representation of a structural parameter called the genetical bandwidth, which describes genetic structure around arbitrary spatial locations in a study area. This parameter corresponds to the shortest distance to areas of significant allele variation, and its computation is based on the Womble's systemic function. A simulation study and application to data sets taken from the literature give evidence that the method is particularly demonstrative when the fine-scale structure is stronger than the large-scale structure, and that it is generally able to locate genetic boundaries or clines precisely.

Keywords: Spatial genetic structure; Wombling; Systemic function

1. Introduction

Wombling methods aim at detecting regions of abrupt changes in maps of biological variables. They were introduced by Womble (1951), and they were refined afterwards by Barbujani et al. (1989) and Bocquet-Appel and Bacro (1994). In the original approach, Womble assumed the knowledge of surfaces derived from the variables of interest (e.g. allele frequency surfaces) and computed the gradient of these surfaces. The norms of the gradients were then averaged to form a new surface called the systemic map (or the systemic function). Zones of rapid changes could therefore be identified as regions given high values by the systemic function. These regions were called boundaries.

However, the surfaces considered in the original approach are rarely available in real data analysis. Instead of surfaces, the variables of interest are usually measured at scarce geographical locations. Implementing Wombling methods therefore requires the preliminary inference of these surfaces from biological data collected from either regular or irregular geographical sampling designs. Analysis of regular experimental designs can be addressed from a technique called lattice Wombling. In this approach, a lattice tessellates the space into rectangular regions termed pixels, and the variables of interest are assigned to the center of the pixels. The rates of change can then be computed either as the first derivatives among adjacent pixels or the second partial derivatives (Laplacian) (Jacquez et al., 2000; Fagan et al., 2003; Fortin and Dale, 2005) using kernel methods that operate in windows of m pixels, where m is a fixed parameter. Regarding irregular experimental designs, the data can be processed by a different technique called triangulation Wombling, which is based on triangular kernels (Fortin, 1994). Once the systemic function has been estimated using one of these methods, the next step is to discriminate between true boundaries and spurious ones due to the inherent variability of experimental designs. To achieve this,

*Corresponding author. Fax: +33 (0) 4 76 51 42 79.
E-mail address: stephanie.manel@ujf-grenoble.fr (S. Manel).

0040-5809/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2007.01.007

Barbujani et al. (1989) assessed the significance of the boundaries by using a randomization procedure. Their procedure detects whether the estimated values are higher than the values expected under the absence of structure. This approach was successfully applied to the analysis of allele frequency data, and it was able to detect zones of abrupt change within Human and *Drosophila*.

Nevertheless the need for statistical reconstruction from scarce observations poses a difficult problem to the implementation of Wombling methods in computer programs. Since lattice Wombling assigns observations to a regular lattice, it may be subject to bias. Triangulation Wombling estimates the systemic function from three data only, and it is then prone to statistical error. This statistical issue is closely related to the choice of the spatial scales at which the estimation procedures (kernels) are implemented. For instance, the rates of change are strongly dependent on the pixel size (Jacquez et al., 2000; Fagan et al., 2003). The issue was recently addressed by a method called hierarchical Wombling (Csillag and Kabos, 2002) that computed maps at several scales by varying the pixel size. But, in general, Wombling does not result in a unique map because several pixel sizes may be used, and it may be problematic to decide which pixel sizes are the most appropriate to interpret the data.

This study addresses the above scaling problem from a different perspective. The approach introduced here, named the Genetical Bandwidth Mapping (GBM), is a nonparametric technique that deals with allele frequency surfaces. The GBM estimates the systemic function at several scales in order to provide local characterization of the genetic structure around any arbitrary spatial location in a study area. The quantities displayed by the GBM are not the systemic values themselves, but new local quantities called the genetical bandwidths. The GBM overcomes the issue of choosing among of multiple maps by computing these local quantities, which may in turn be interpreted as the shortest distances to areas of significant variation in allele frequencies (Section 2). This study also evaluates the capabilities of the GBM to detect and locate genetic structures such as boundaries or clines using multilocus genotypes and geographical coordinates (Sections 3 and 4).

2. Theory

This section presents a formal description of bandwidths and describes the statistical principles underlying the GBM. The mathematical details are deferred to Appendix. We consider a biological population living in a two-dimensional habitat, and we assume that n individuals have been sampled from this population according to a uniform experimental design (regular or random). In what follows, each location (x_i, y_i) is associated with a genetic observation g_i called the multilocus genotype that indicates the presence of specific alleles at multiple DNA loci.

Typically the coordinates (x_i, y_i) represent either the location of an individual labelled i at its instant of observation, or the location of its habitat. The GBM computes a critical parameter (the genetical bandwidth) at each site of a grid that covers the study area. In the sequel, a grid site is denoted by (x, y), and it may differ from the sampled location (x_i, y_i).

2.1. Definition of bandwidths

In the GBM, the gradients of allele frequencies are computed at each grid site (x, y). In this grid, the neighborhood of a site is not defined precisely. Instead each observation receives a weight that depends on the distance to the current grid site. More specifically, the individual i located at (x_i, y_i) is given the Gaussian weight $w_i(h) = \exp(-d_i^2/2h^2)$, where d_i^2 represents the squared Euclidean distance between (x_i, y_i) and (x, y). This approach is standard in density estimation (Silvermann, 1986). The parameter h is crucial to our method. It is called the bandwidth (or window size), and it controls the exponential decay of the Gaussian weights.

2.2. The systemic function

In order to test for the absence of local genetic structure, the GBM computes an estimate of the Womble's systemic function at each grid site. The systemic function $S(x, y)$ can be defined by the following formula:

$$S(x,y) = \sum_j \left\| \nabla f_j(x,y) \right\|,$$

where $\nabla f_j(x, y)$ is the gradient of the allele frequency f_j with respect to x and y, and the sum runs over all possible alleles j at all loci. The systemic function is representative of the total slope of allele frequency surfaces and it integrates the dependencies between the genetic and the spatial data. The GBM deals with the fact that the derivatives at (x, y) are unknown parameters by estimating them from the presence/absence of alleles at the sample locations. The estimation technique used here relies on local polynomials based on the Gaussian weights $w_i(h)$ (see Fan and Gijbels, 1996 and Appendix). This approach differs from standard techniques significantly because the derivatives are traditionally estimated from difference equations (e.g. Barbujani et al., 1989). Depending on the value given to the bandwidth h, the GBM builds an estimate $S(x, y, h)$ of $S(x, y)$ at each (x, y). In nonparametric statistics, the choice of h is usually motivated by the minimization of a statistical error. Optimal choices nevertheless generate difficult mathematical and practical problems. The GBM avoids this particular issue by computing h in relation to critical regions of tests for the absence of structure.

2.3. Testing for spatial genetic structure

Testing for genetic structure is at the heart of the GBM. The tests for absence of structure are repeated at each site (x, y) of the grid. Their objective is to determine whether

the estimated systemic value $S(x, y, h)$ reflects significant local structure or not. In order to perform the tests, the values $S(x, y, h)$ are compared to the probability distribution of the systemic function $S(x, y)$ obtained under the null hypothesis of absence of structure. Such a probability distribution can be computed by using a permutation procedure. The algorithm actually proceeds with resampling from all possible genotypes, and it reassigns the sampled genotypes to the individual locations. For each (x, y), replicates of systemic values are computed as described in Section 2.2. A P-value is then computed by forming the ratio of the number of replicates greater than the estimated value $S(x, y, h)$ to their total number. For example, Fig. 1 represents individuals genotyped at a single diallelic haploid locus geographically sampled for a structured population. The figure suggests that the test should be significant for large bandwidths ($h > h_1$) but nonsignificant for bandwidths lower than the distance to the closest discontinuity in allele frequencies (e.g. $h < h_2$).

2.4. Genetical bandwidths

To compute the genetical bandwidths, the type I error of the permutation test must be set to a fixed value α (typically $\alpha = 0.05$). The tests are repeated at each grid site for several values of h. The genetical bandwidths are then defined as the largest values of h for which the tests are nonsignificant, i.e. the P-values are greater than α. To compute these quantities, large bandwidths corresponding to the diameter of the study area are first tested. The bandwidths are then decreased by one unit unless the test becomes nonsignificant.

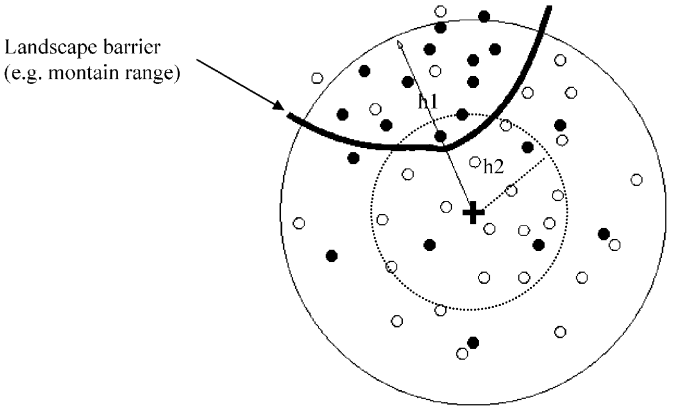

Fig. 1. The permutation test illustrated. The figure represents individuals (dots) genotyped at a two-allele locus (A = black dots, a = white dots), and includes geographical structure. The thick black line indicates the presence of a physical barrier (e.g. a mountain range), and the cross corresponds to a site (x, y) in the grid. When the permutation test is applied to individuals at distance h_1 black and white dots may be mixed up with high probability. The null hypothesis of absence of structure may then be rejected. In contrast, the test may be nonsignificant when the permutation test is applied to individuals at distance h_2.

2.5. Interpretation of the GBM output

The GBM output is stored as a two-dimensional matrix, which can be interpreted as a map. Such a matrix contains the genetical bandwidths computed at each point of the grid covering the study area. These parameters correspond to the shortest distances to the zones of significant variation in allele frequencies. Graphical representations of the GBM therefore provide bases for interpretation of the genetic structure of the population. In this paragraph, we give short guidelines to the interpretation of such outputs by analyzing two typical responses. For sake of clarity, we discuss one-dimensional organization, i.e. populations which consist of continuously distributed individuals along a line. The one-dimensional hypothesis is actually more amenable to analysis, and this enables reasonable guesses of typical shapes of response in two dimensions (cross-sections in two-dimensional maps). In one dimension, two types of response can be expected. We call these responses the *W-shaped* and the *V-shaped* curves (Fig. 2). The W-shaped response may be the signal of a

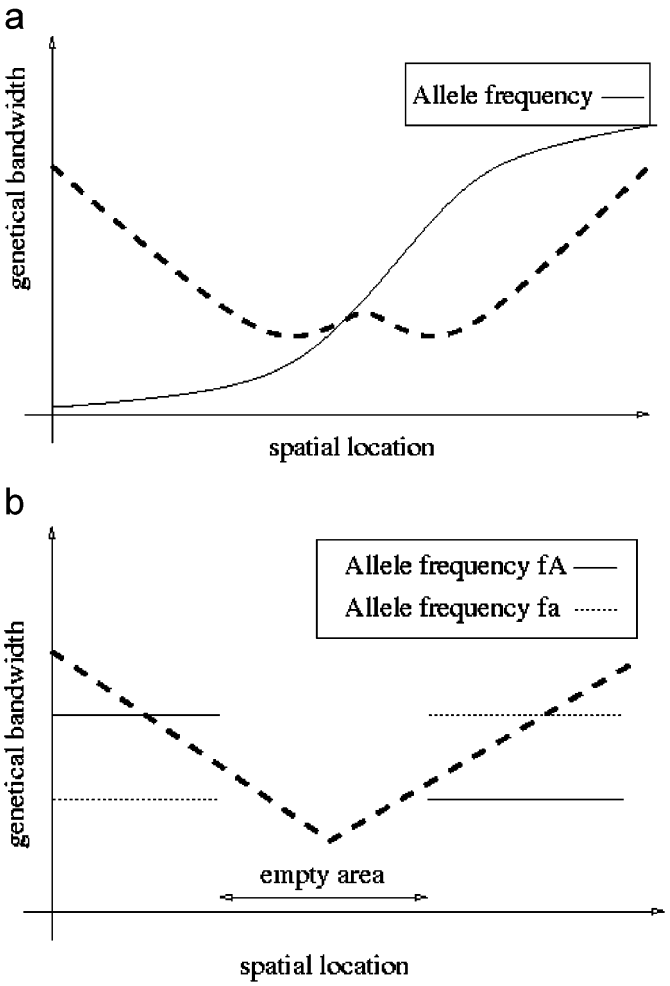

Fig. 2. (a) W-shaped response corresponding to a cline in allele frequency at a locus with two alleles. (b) V-shaped response corresponding to physical barrier to gene flow.

cline in allele frequency, while the V-shaped response is expected when differentiated populations are isolated by a nonhabitat or unsampled area. The W-shaped response is illustrated in Fig. 2a, where the spatial distribution of individuals genotyped at a haploid biallelic locus (a/A) is displayed. Within areas of low frequency of allele A, the genetical bandwidth decreases as the distance to the zone of sharp variation decreases. The same phenomenon occurs within areas of high frequency. Within the transition zone, the genetical bandwidths may undergo erratical variation due to variability in geographical sampling, and the fact that the allele frequencies are close to 50%. Fig. 2b gives an illustration of the V-shaped response. Two populations are separated by an empty area, e.g. a physical barrier to gene flow. In this case, the genetical bandwidths vary linearly with the distance to the farthest population.

2.6. Software

A documented computer software called GENBMAP implements the GBM in C++ and provides a graphical interface for visualizing its outputs. The computer program has been designed for running under the win32 operating system, and it is available from the authors' web pages (http://www-timc.imag.fr/Olivier.Francois or http://www-leca.ujf-grenoble.fr/logiciels.htm).

3. Simulation study

This section illustrates the behavior of the GBM when confronted to typical spatial structures obtained from numerical simulations. Two cases were simulated: (1) clines along a specific direction (longitude) and (2) barriers to gene flow in the two-island model. The choice of such simple scenarios was motivated by the fact that these scenarios clearly enable the measure of the mutual influences of both spatial and genetic effects on the GBM output. Analysis of the behavior of GBM on more complex scenarios is deferred to the next section, where real data sets are considered.

3.1. Experimental design and simulation tools

Random generation of spatial locations and genetic data were performed using the statistical software R 2.3.1 (R Development Core Team, 2006). Simulations were replicated more than 10 times in each case. Sample sizes were increased from 200 to 1000 and 5000 individuals. Maps were computed using 300×300 regular grids (90,000 sites). The type I errors in permutation tests were equal to $\alpha = 0.05$ and $\alpha = 0.1$. Refining the mesh grid had a direct impact on the running time, which increased linearly with the number of sites in the grid. The number of permutations was fixed to 200, and then increased to 500 and 1000 without major influence on the output of the tests. With 200 permutations, the computing time for one single map approximated half an hour on a 2 GHz processor laptop computer. Additional genetic data were generated from the two- and three-island models using EASYPOP (Balloux, 2001) with similar results (not reported). The GBM is a purely descriptive method, and its statistical behavior over many simulated replicates is particularly difficult to summarize. The graphical results presented in this section were chosen as representative of the majority of the simulated cases.

3.2. Simulation of clines

Simulated data included 12 unlinked diallelic loci (alleles were coded 0 and 1). At each locus, the frequencies of 1's varied along the horizontal direction (x coordinate), and their dependence on x was logistic $f(x) = 1/(1+\exp(-(x-a)/b))$ where a and b were specific constants set to $a = 2500$, 3000, 3500 and $b = 200$, 500 (each of the six combinations was produced twice, see Fig. 3a). Spatial coordinates were obtained as mixtures of two independent isotropic Gaussian distributions centered at $x_1 = 1000$ and $x_2 = 3000$. The y coordinate was set to $y = 1800$ and the standard deviation was SD = 1000. These simulations were relevant to clinal variation of allele frequencies along one specific direction (e.g. longitude).

Fig. 3b displays a central horizontal cross-section ($y = 900$) of the GBM. The cross-section exhibits a W-shaped response locating the beginning of the sharp variation around $x = 1500$ and its end around $x = 4000$. The central area of the map corresponds to a wide region where allele frequencies varied significantly. In these experiments, increasing the type I error produced wider minimal areas with weak incidence on the result interpretation. A sensitivity analysis was actually performed for this parameter (not reported). Although the minima appeared to be wider, their shape and their location were unchanged when α was increased to 10% and 15%. As for many statistical tests, fixing the type I error to $\alpha = 5\%$ may be considered as a generally reasonable strategy.

3.3. Simulation of two-island models and sensitivity analysis

A spatial variant of Wright's island model was used in order to assess the sensitivity of the GBM to the simultaneous variation of genetic differentiation and the density of the spatial data. Individuals were sampled from two subpopulations of equal effective sizes N_e. Alleles were simulated according to the infinite allele model with constant mutation rate $\theta = 4\mu N_e = 1$ at each locus (μ is the mutation rate per generation). Migration rates $M = 4mN_e$ were varied from 1 to 10 loci (m is the migration rate per generation). F-statistics (F_{ST}) were computed using Weir and Cockerham estimates (Weir and Cockerham, 1984). In order to assess effects due to the number of loci and the sample size, we used $L = 5$, 20, and 50 loci and $n = 20$, 50, and 100 individuals. Spatial coordinates were simulated as the geographical mixture of two independent Gaussian distributions. Each subpopulation

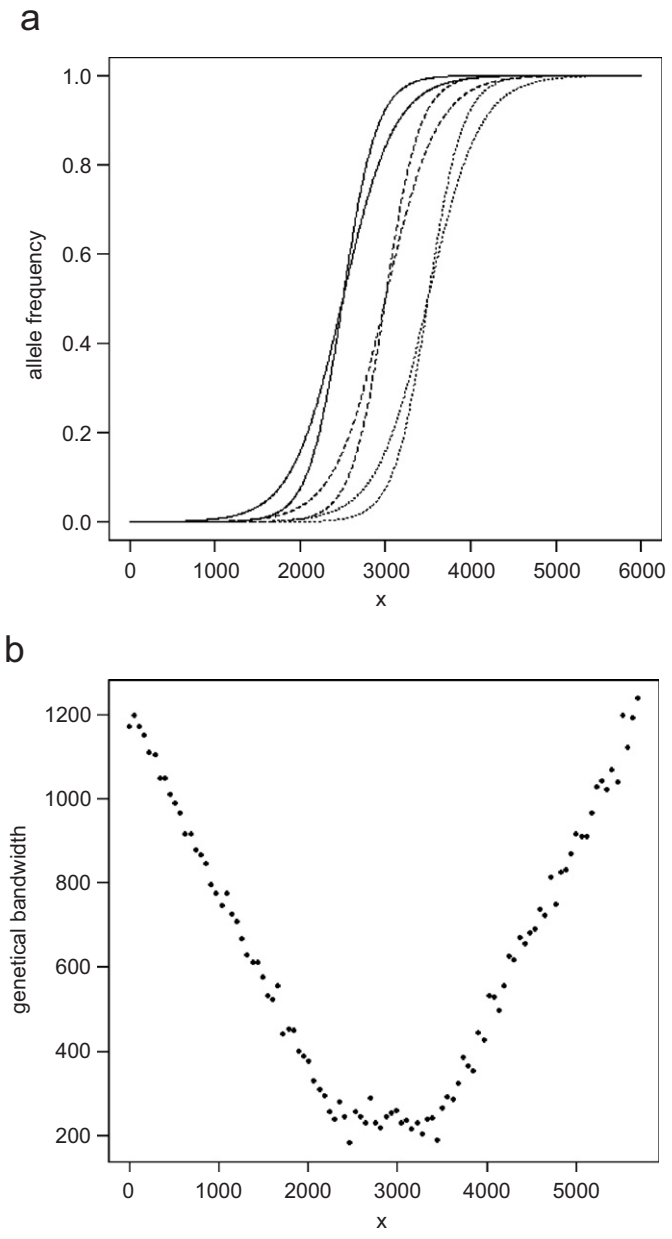

Fig. 3. (a) Allele frequencies used in the simulation of clinal variation. (b) One-dimensional cross-section of the GBM corresponding to the response of a cline in simulated data. Genetical bandwidths ranged from 200 (min) to 1200 (max).

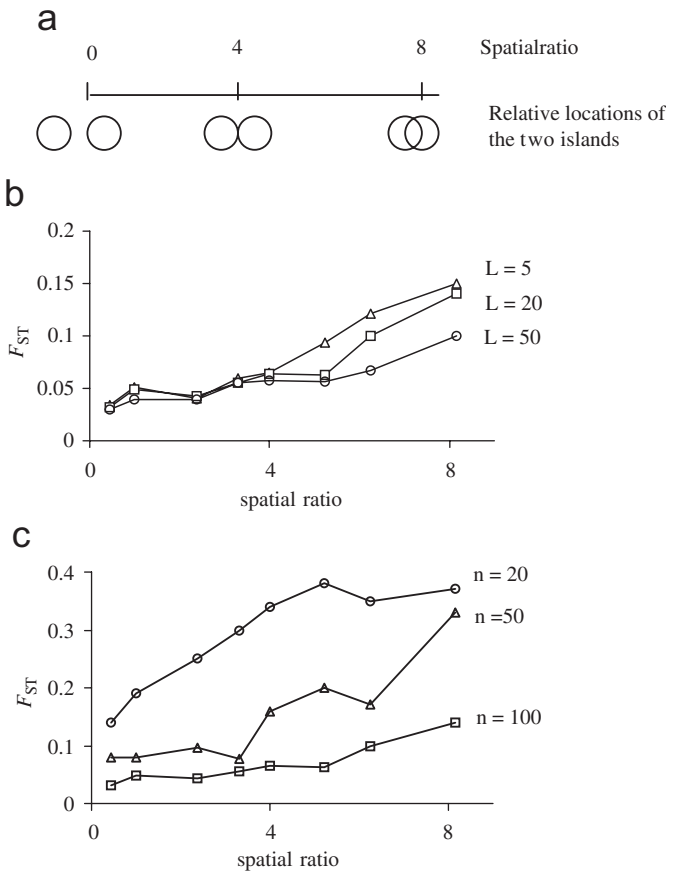

Fig. 4. Simulations of the two-island model with limited gene flow. (a) Relative locations of the two islands along the r-axis. (b) Critical F_{ST} values below which population subdivision cannot be detected (as a function of r). The sample size is set to $n = 100$ individuals. The numbers of loci are set to $L = 5$ (triangles), $L = 20$ (squares), and $L = 50$ (circles). (c) Critical F_{ST} values below which population subdivision cannot be detected (as a function of r). The number of loci is set to $L = 20$. The sample sizes are $n = 20$ (circles), $n = 50$ (triangles), and $n = 100$ (squares).

(or island) density had its own spatial range, and the two islands could intersect. The ratio r of the within-group variance to the between-group variance was used to measure the degree at which the two islands spatially overlapped. This classical discriminant analysis parameter was interpreted as a measure of spatial differentiation. For fixed within-group variances, bringing islands closer had the effect of increasing r while moving islands away from each other had the reverse effect (see Fig. 4a). For r between 0 and 4, the two islands did not generally share an intersecting area and they even remained far apart for the lowest values. For r between 4 and 8, the regions intersected significantly.

Simulations of this model with various values of r and various levels of F_{ST}'s were used to investigate which parameter values were critical to the detection of population structure (Fig. 4b and c). At the critical values, the GBM provided flat responses, and the minima were hardly detectable. We considered that the map was flat when the ratio of the maximum of the GBM to its minimum value minus one was less than 5%. These ratios were obtained from 10 simulated replicates. Fig. 4b represented critical F_{ST} values as a function of r. For r in the range (0,4), structure was detected for F_{ST}'s less than a value around 0.04. In this range, the method was relatively insensitive to the number of loci. For higher levels of spatial mixture (r around 6), the values below which population subdivision remained undetected increased, but the performances clearly rose with the number of loci. With 20 loci, population structure was detected for F_{ST} values less than 0.05–0.06. Fig. 4c showed that the GBM was sensitive to

geographical sampling, and the method was frequently confounded when the sample sizes were small. For $n = 20$, population structure was correctly identified for large F_{ST} (>0.15). For $n = 100$, the GBM was remarkably efficient at detecting structure even when the two subpopulations intersected strongly.

4. Application to real data

This section summarizes the analysis of two well-studied data sets. The first one can be considered as a case of evidence of clinal selection at the Adh F/S locus in *Drosophila melanogaster* (Berry and Kreitman, 1993). The second one is the Human Genome Diversity Panel from the Centre d'Etude du Polymorphisme Human (HGDP-CEPH), which contains the genotypes of 1056 individuals at 377 autosomal microsatellite loci. This data set was used by Rosenberg et al. (2002) to infer the genetic structure of the modern human population. Here these data sets may be viewed as completing the simulation study described in Section 3. Although simulating cline and cluster models was useful for evaluating the sensitivity of the GBM to several parameters, the simulations did not reflect the complexity of realistic situations. The two data sets studied in this section provide insights on how the method works on more complex scenarios.

4.1. Adh locus

Clinal selection at the *Adh F/S* locus in *D. melanogaster* was studied by Berry and Kreitman (1993) using 113 haplotypes from 44 polymorphic markers in 1533 individuals from 25 population sites from the East Cost of North America. The original data contained latitudinal information for the 25 population samples. To create individual locations, the spatial coordinates were simulated by adding a small amount of variability to each site coordinates. These sites were pertubated by adding SD = 0.5°N to the latitudes, and longitudes were simulated within an artificial range of $(-1, +1)$°E. In addition, we used a subsample of 1303 individuals so that the 14 most represented haplotypes were selected (85% of the full data set at the end of Berry and Kreitman's article). Clearly the above described procedures did not favor bias toward the appearance of clines.

The two-dimensional plot displays a band pattern, which reflected the absence of sensitivity to longitude resampling. Most latitudinal cross-sections exhibited similar band patterns (not shown). The curve in Fig. 5 corresponds to a single section of the map. This curve exhibits a W-shaped response locating the beginning of the zone of sharp variation around the latitude 32.4°N and the end of this zone around the latitude 41.1°N. These results gave evidence that the cline was correctly retrieved and that it separated two homogeneous zones in the South and the North of the study area.

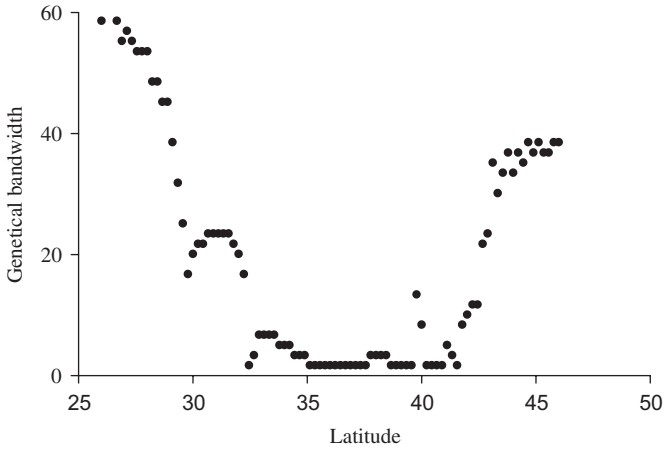

Fig. 5. GBM response to the latitudinal cline in frequencies of the *Adh* locus in the North American *D. melanogaster* populations (latitudinal section).

4.2. CEPH human genome data set

The genetic structure of modern human populations was recently investigated without the use of predefined "groups" by Rosenberg et al. (2002) and Cann et al. (2002). Inference of genetic ancestry was performed by applying a model-based clustering algorithm implemented in the computer program STRUCTURE (Pritchard et al., 2000) that computes individual cluster membership coefficients. Rosenberg et al. (2002) identified six main genetic clusters, five of which corresponded to major geographic regions. The secondary clusters often matched with one of the 52 sample populations.

Here, the GBM was applied to the Eurasia/Asia subset of the original data set. These data contained 451 individuals originating from 25 populations. The study area ranged from south-western Pakistan (latitude 24°N, longitude 66°E) to north-eastern Russia (latitude 64°N, longitude 130°E). Geographical coordinates and spatial ranges of population samples were available from the CEPH web site (http://www.cephb.fr/HGDP-CEPH-Panel/). Because the individual coordinates were not known exactly, Gaussian data simulated within the range specified from the CEPH web site were used instead. This was done by adding small amounts of variability to the geographical coordinates. Several densities of spatial coordinates were investigated. The map was computed using a regular grid of resolution 100×100, 200 permutations, and the type I error set was to $\alpha = 0.05$. The result is displayed in Fig. 6.

The outputs were consistent with recent results regarding the inference of the genetic structure of human populations obtained with Bayesian methods (Rosenberg et al., 2002; Bamshad et al., 2003; Ramachandran et al., 2004; François et al., 2006). Evidence for the separation Eurasia/East Asia was provided by the transversal separation (T). The (L) and (U) signals may be interpreted as genetic barriers to gene flow (a double barrier). The upper signal (U) may be interpreted as the separation between Xibo and Uygur of

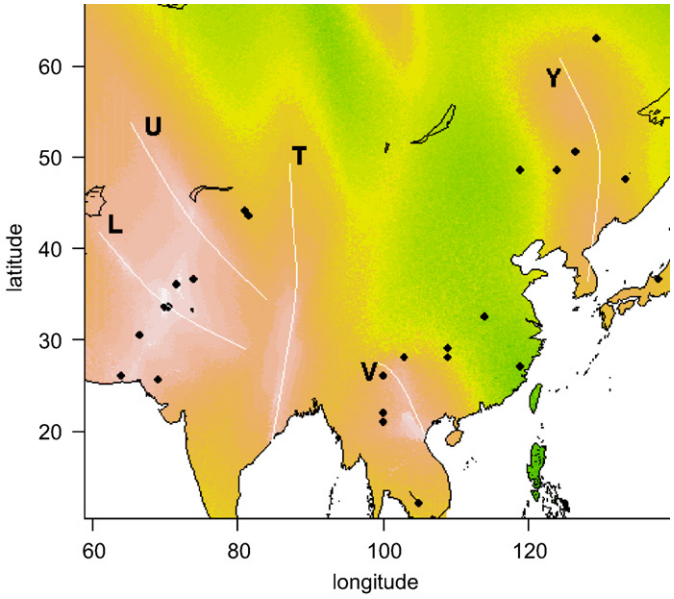

Fig. 6. Map produced by the GBM method for the Eurasian and Asian modern human populations (HGDP-CEPH data set). U, T, and V are signals of the genetic differentiation between Eurasian and Asian populations. L may be explained by the genetic isolate Kalash, and Y is consistent with the separation of populations with languages of altaic origin.

north-western China and the populations from southern Pakistan who speak Indo-European languages. The (V) line appears to be the continuation of (U). Grouped as a single signal (UV) might also indicate the separation between Pakistani and East-Asian samples. The lower separation (L) matched with the presence of the Kalash population, which was clearly identified as a genetic isolate by Rosenberg et al. (2002), and Fig. 6 indicates that the Pakistan area featured higher genetic heterogeneity than the Asia area (lower values of the GBM). As concerned the Asia samples, an heterogeneous area was observed in north-eastern China (Y). The shape of this structure may also indicate a separation between Yakut–Japanese samples, which are populations with altaic language and the rest of the East-Asian samples. The principal features reported in this paragraph were common to all computed maps regardless the spatial densities from which individual coordinates were resampled (not reported).

5. Discussion

The GBM was introduced as a new visual tool for investigating spatial variation of allele frequencies. The information was displayed through a two-dimensional graphical representation of a local structural parameter. This parameter could be interpreted as the shortest distance to areas of significant changes as well as the radius of the largest zone for which the genetic structure can be thought of as being spatially homogeneous.

The definition of genetic bandwidths fit in the framework of Wombling methods because the systemic map provided a natural measure of spatial homogeneity. GBM and Wombling were nevertheless fundamentally distinct. The main difference resided in the fact that Wombling estimated the systemic function by using a fixed local parameter (e.g. a window size). Due to the high sensitivity to this parameter, systemic maps might hardly be estimated unambiguously, because each value of the local parameter might lead to a new map. Deciding which value minimizes the statistical error is a generally difficult issue. In contrast, the GBM avoided these issues by adopting a reverse perspective, focusing on bandwidths rather than on the systemic map itself. The bandwidths were estimated on the basis of local homogeneity tests using all values of the local parameter. The GBM therefore produced a unique map.

The GBM proved successful at identifying genetic discontinuities and sharp clinal variation. Genetic discontinuities or boundaries induced V-shaped responses, while clinal variation was likely to yield W-shaped responses. Application to real data provided additional evidence that the GBM was able to identify clinal variation in allele frequencies (Drosophila *Adh* locus). In addition, the maps obtained from the HGDP data set were in accordance with our current knowledge of the genetic structure of Eurasian and Asian populations.

Variation of allele frequencies in human populations may occur at multiple scales. For example, the genetic discontinuities existing at the transition between Eurasia and East Asia may result from large-scale variation, while the Kalash cluster emerges within Pakistan at a finer scale. From the very definition of genetical bandwidths, the GBM operates at local scales primarly and may then be appropriate for detecting recent differentiation followed by restricted range expansion. Genetic discontinuities within a high-density area may actually result in relatively low values of the GBM (e.g. within Pakistan). At the same time, these areas may form several geographical clusters separated by larger distances and potentially more ancient. In this situation, large-scale variation can still be detected by the GBM, but it may be given higher relative values than the primary responses leading to secondary structures (see Fig. 7).

The GBM seemed particularly relevant to the study of populations in which the fine-scale population structure is stronger than in the long range (Kayser et al., 2005; Marjanovic et al., 2005; Klopfstein et al., 2006). The color coding representation used by the graphical interface (GENBMAP) may however be partly responsible for the fact that large-scale structure may confound the program when local structure is present because the program uses few colors, and users of GENBMAP may be aware that the secondary minima (or ridges) may reflect the imprints of large-scale structures more significantly than the primary mimima do. Nevertheless secondary structures associated with large-scale variation in the human data could be unambiguously identified from Fig. 6 showing that the program can also perform well when the large-scale structure is stronger. A piece of advice to give to users of

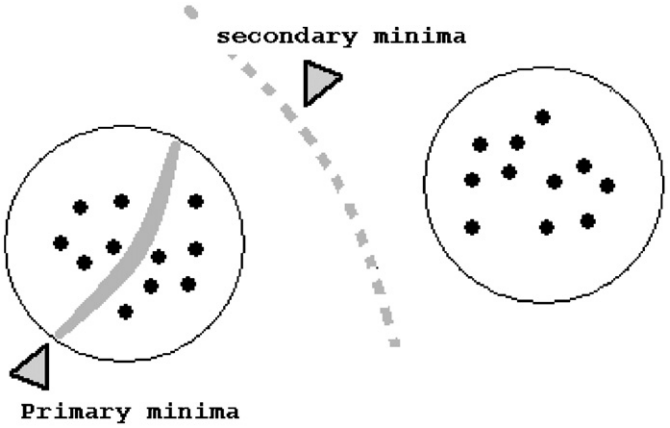

Fig. 7. Primary minima in the GBM may indicate local variation of allele frequencies (they correspond to white areas in GENBMAP outputs). Secondary minima may indicate the existence of structure at larger scales and may be less apparent.

GENBMAP is to perform separate runs at local scales in order to detect the presence of local variation. A separate analysis of the eight Pakistan samples (not reported) actually supported the hypothesis that the response observed in Fig. 6 might be a consequence of the multi-scale variation present in the data. The separate analysis concluded that low values of the GBM surrounded the Kalash locations.

A frequently reported issue in the recent literature is that computer programs seeking population structure may be confounded by irregular spatial sampling (Serre and Pääbo, 2004; Rosenberg et al., 2005). Perhaps the most widely used among these computer programs is the Bayesian clustering algorithm STRUCTURE (Pritchard et al., 2000), which probabilistically assigns individuals to K populations of origin. Because STRUCTURE puts a strong prior on the existence of clusters, it may be prone to errors when geographical sampling is uneven along clines. Recently François et al. (2006) dealt with the cline/cluster dilemma from a Bayesian perspective, and used hidden Markov random fields in order to attenuate the effect of uneven sampling in continuous populations. Although the GBM does not provide a direct solution to this very difficult issue, it yielded a reasonable answer when the problem happened in real data (Drosophila). Although geographical sampling was actually irregular in the Drosophila data, the GBM response was indeed relevant to clinal variation. This may be explained as both large-scale and local spurious structures contributed to the observed W-shaped response.

What does the GBM bring compared to the already existing methods? The answer requires a short review of methods used in spatial genetics. Spatial population genetics often relies on theoretical models and statistical methods for inference from genetic data in subdivided populations. Mainly, these methods are concerned with the estimation of migration or dispersal rates based on neutral models of evolution (Wright, 1943; Malécot, 1968; Rousset, 2004) that have originally demonstrated how *F*-statistics depends on spatial dispersal under simplified assumptions. These approaches assume predefined populations and they do not use spatial information explicitly. Another stream of theoretical works has traditionally been built on nonparametric spatial statistics. These works fall into three categories (see Manel et al., 2003): (1) matrix methods and the Mantel test (Mantel, 1967), (2) spatial autocorrelation statistics (Moran, 1950; Sokal and Oden, 1978), and (3) methods related to identification of boundaries (Monmonier, 1973). A criticism addressed to Mantel and autocorrelation methods is that they may indeed reveal the presence of specific structures, but they fail to identify their shape or their location precisely (Barbujani, 2000). In constrast, the Monmonier's algorithm (Monmonier, 1973) and the Wombling method include the analysis of local features. However, the Monmonier's algorithm requires fixing a number of intrinsic parameters, as the number of genetic boundaries, which are generally unknown.

Clustering methods like those based on principal component analysis or phylogenetic reconstruction are also popular in spatial genetics and phylogeography, but they may sometimes be difficult to interpret. In contrast, model-based clustering algorithms like the one implemented in the Bayesian computer program STRUCTURE have revealed powerful at detecting cryptic population structure. Although these clustering algorithms do not exploit spatial information explicitly, the identified clusters can be mapped on a landscape representation (e.g. Manel et al., 2004), and this generally provides satisfactory results.

In regard of all the previous approaches, GBM can be classified as nonparametric (as opposed to model-based) and uses spatial information explicitly. Like autocorrelation methods, it deals with genetic data and geographical coordinates simultaneously. But the main difference is that GBM enables locating the spatial genetic structures. Compared to methods that use the Monmonier's algorithm, GBM avoids using predefined number of populations and does not assume a particular measure of genetic distances. The results of GBM are not directly comparable to Bayesian clustering algorithms because the GBM does not assign any individual to a population of origin. With this respect, the GBM may deal with the cline/cluster dilemma more equitably than Bayesian clustering algorithms do. In addition, GBM is not subject to the convergence diagnosis issue often reported for MCMC programs.

Wombling has generated a great amount of applied and theoretical works since its introduction by Womble in 1951 (Womble, 1951; Barbujani et al., 1989; Bocquet-Appel and Bacro, 1994; Fortin et al., 2000; Jacquez et al., 2000; Fagan et al., 2003). So far, these works have focused on estimating systemic maps using various statistical procedures. This article adopted a distinct approach. Based on Wombling, it introduced a new structural parameter (the genetical bandwidth) and gave a natural interpretation to this

parameter. The common idea underlying studies of geographical diversity is that one can proceed from the observed pattern to the underlying evolutionary process (Barbujani, 2000). The first step is to assess the observed pattern of genetic variation. GBM has proved to be a powerful tool to address this question and to provide graphical insights on population structure when no prior information is available.

Acknowledgments

We are grateful to J. Rebreyend for his help with the GENBMAP program interface. This work was conducted while O.F. and S.M. were supported by the IMAG project AlpB and the French Ministery of research projects ACI project ImpBio and ANR MAEV.

Appendix. Mathematical details of the GBM

The estimates of derivatives needed for computing the systemic function were obtained according to a nonparametric method using *local polynomials* (Fan and Gijbels, 1996). Let us explain the method briefly. For sake of simplicity, assume the observation of a single locus, and let (z_i) denote the Bernoulli variables that indicate the presence or absence of a specific allele at each site. Here the subscript i refers to the individual location, and $z_i = 1$ means that an organism located at the site i carries the studied allele.

We denote by (x_i, y_i) the spatial coordinates of the individual i and by (x, y) the coordinates of an arbitrary grid site. In the local polynomial method, a function w weights observations according to their distance to the grid sites. The weight function usually consists of a nonnegative decreasing function of the Euclidean distance d_i between the observation (x_i, y_i) and the grid site (x, y). In nonparametric statistics, Gaussian weights are a standard choice:

$$g(d) = \exp(-d^2/2), \quad d > 0.$$

The bandwidth (denoted as h) is a scale parameter that enables the control of weight decay. This parameter influences the quality of estimation, and it is subject to the bias/variance dilemma. More specifically the observation i receives the weight

$$w_i(h) = g(d_i/h).$$

When the bandwidth is small, the decay can be fast compared to the scale of spatial data, and only the observations close to the grid site will be taken into account. In contrast, large bandwidths may lead to an underestimation of local structures.

The local polynomial method attempts to fit a polynomial function $P(x, y)$ to the (unknown) frequency of the studied allele at every grid site (x, y). The fitted polynomial takes the following form:

$$P(x,y) = \alpha_0 + \beta_x x + \beta_y y + 1/2\gamma_{xx} x^2 + \gamma_{xy} xy + 1/2\gamma_{yy} y^2,$$

and it solves the minimum square problem associated with the error function

$$E(P) = \sum_{i=1}^{n} w_i(h) \left\| P(x_i - x, y_i - y) - z_i \right\|^2, \tag{1}$$

where n is the sample size, and θ corresponds to the six-dimensional set of parameter defining P. The solution of the minimization problem (1) is given by the following matrix product:

$$\theta = ({}^tXWX)^{-1}({}^tXWZ), \tag{2}$$

where $Z = {}^t(z_1, \ldots z_n)$, W is a n-dimensional diagonal matrix such that $w_{ii} = w(d_i/h)$, and

$$X = \begin{pmatrix} 1 & x_1 - x & y_1 - y & (x_1 - x)^2 & (x_1 - x) & \times & (y_1 - y) & (y_1 - y)^2 \\ . & . & . & . & . & & . & . \\ . & . & . & . & . & & . & . \\ . & . & . & . & . & & . & . \\ 1 & x_n - x & y_n - y & (x_n - x)^2 & (x_n - x) & \times & (y_n - y) & (y_n - y)^2 \end{pmatrix}$$

Note that computing the estimated parameter θ requires a number of operations of order $O(n)$, which is generally fast on modern computers. The derivatives can be estimated as

$$\frac{\partial P}{\partial x}(x,y) \equiv \beta_x = \beta_x(x,y,h)$$

and

$$\frac{\partial P}{\partial y}(x,y) \equiv \beta_y = \beta_y(x,y,h).$$

An estimate of the gradient norm can then be given by

$$\left\| \nabla f \right\|^2 = \beta_x^2 + \beta_y^2$$

Note that Eq. (2) can be rewritten as

$$\theta = BZ,$$

where $B = ({}^tXWX)^{-1}({}^tXW)$.

The estimates are thus obtained by making the product of the two matrices B and Z. The matrix B depends on the spatial data only, whereas the genetic data are contained in Z. This remark is crucial for the implementation of the permutation tests. During the permutation test, the alleles are resampled without replacement, and the matrix B does not need to be recalculated. The derivatives are therefore reestimated from the application of a single matrix product. The reduced algorithmic cost of this method supports the choice of the linear regression method (instead of a logistic regression method for example).

References

Balloux, F., 2001. EASYPOP (version1.7): a computer program for the simulation of population genetics. J. Hered. 92, 301–302.

Bamshad, M.J., Wooding, S., Watkins, W.S., Ostler, C.T., Batzer, M.A., Jorde, L.B., 2003. Human population genetic structure and inference of group membership. Am. J. Hum. Genet. 72, 578–589.

Barbujani, G., 2000. Geographic patterns: how to identify them and why? Hum. Biol. 72, 133–153.

Barbujani, G., Oden, N.L., Sokal, R., 1989. Detecting regions of abrupt change in maps of biological variables. Syst. Zool. 38, 376–389.

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

Bocquet-Appel, J.P., Bacro, J.N., 1994. Generalized Wombling. Syst. Biol. 43, 442–448.

Cann, H.M., de Tomas, C., Cazes, L., et al., 2002. A human genome diversity cell line panel. Science 296, 261–262.

Csillag, F., Kabos, S., 2002. Wavelets, boundaries and the spatial analysis of landscape patterns. Ecoscience 9, 177–190.

Fagan, W.F., Fortin, M.-J., Soykan, C., 2003. Integrating edge detection and dynamic modelling in quantitative analysis of ecological boundaries. Bioscience 53, 730–783.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Fortin, M.-J., 1994. Edge detection algorithms for two-dimensional ecological data. Ecology 75, 956–965.

Fortin, M.-J., Dale, M., 2005. Spatial Analysis. A Guide for Ecologist. Cambridge University Press, Cambridge.

Fortin, M.-J., Olson, R.J., Ferson, S., Iverson, L., Hunsaker, C., Edwards, G., Levine, D., Butera, K., Klemas, V., 2000. Issues related to the detection of boundaries. Landscape Ecol. 15, 453–466.

François, O., Ancelet, S., Guillot, G., 2006. Bayesian clustering using hidden markov random fields in spatial population genetics. Genetics 174, 805–816.

Jacquez, G., Maruca, S., Fortin, M.-J., 2000. From fields to objects: a review of geographic boundary analysis. J. Geogr. Syst. 2, 221–241.

Kayser, M., Lao, O., Anslinger, K., Augustin, C., Bargel, G., Edelmann, J., Elias, S., Heinrich, M., Henke, J., Henke, L., et al., 2005. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. Hum. Genet. 117, 428–443.

Klopfstein, S., Currat, M., Excoffier, L., 2006. The fate of mutations surfing on the wave of a range expansion. Mol. Biol. Evol. 23, 482–490.

Malécot, G., 1968. The Mathematics of Heredity. Freeman & Company, New York (USA).

Manel, S., Schwartz, M., Luikart, G., Taberlet, P., 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18, 157–206.

Manel, S., Bellemain, E., Swenson, J., François, O., 2004. Assumed and inferred spatial structure of populations: the Scandinavian brown bears revisited. Mol. Ecol. 13, 1327–1331.

Mantel, N., 1967. The detection of disease clustering and a generalised regression approach. Cancer Res. 27, 173–220.

Marjanovic, D., Fornarino, S., Montagna, S., Primorac, D., Hadziselimovic, R., Vidovic, S., Pojskic, N., Battaglia, V., Achilli, A., Drobnic, K., et al., 2005. The peopling of modern Bosnia-Herzegovina: Y-chromosome haplogroups in the three main ethnic groups. Ann. Hum. Genet. 69, 757–763.

Monmonier, M., 1973. Maximum-difference barriers: an alternative numerical regionalization method. Geogr. Anal. 3, 245–261.

Moran, P.A., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

R Development Core Team, 2006. R Installation and Administration. R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran, S., Rosenberg, N., Zhivotovsky, L., Feldman, M., 2004. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Hum. Genomics 1, 87–97.

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

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

Rousset, F., 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton, NJ, USA.

Serre, D., Pääbo, S., 2004. Evidence for gradients of human genetic diversity within and among continents. Genome Res. 14, 1679–1685.

Silvermann, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York.

Sokal, R., Oden, N., 1978. Spatial autocorrelation in biology. I-Methodology. Biol. J. Linn. Soc. 10, 199–228.

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

Womble, W., 1951. Differential systematics. Science 28, 315–322.

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