The LINNEAN SOCIETY of London BIOLOGICAL Journal of the Linnean Society

Biological Journal of the Linnean Society, 2014, **113**, 1152–1161. With 4 figures

Pattern-oriented modelling of population genetic structure

JOSÉ ALEXANDRE FELIZOLA DINIZ-FILHO[1]*, THANNYA NASCIMENTO SOARES[2] and MARIANA PIRES DE CAMPOS TELLES[2]

[1]*Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás, CP 131 Campus II, 74001-970 Goiânia, GO, Brazil*
[2]*Departamento de Genética, Instituto de Ciências Biológicas, Universidade Federal de Goiás, CP 131 Campus II, 74001-970 Goiânia, GO, Brazil*

Received 18 April 2014; revised 29 June 2014; accepted for publication 30 June 2014

Although several statistical approaches can be used to describe patterns of genetic variation and infer stochastic differentiation, selective responses, or interruptions of gene flow due to physical or environmental barriers, it is worthwhile to note that similar processes, controlled by several parameters in theoretical models, frequently give rise to similar patterns. Here, we develop a Pattern-Oriented Modelling (POM) approach that allows us to determine how a complex set of parameters potentially driving empirical genetic differentiation among populations generate alternative scenarios that can be fitted to observed data. We generated 10 000 random combinations of parameters related to population size, gene flow and response to gradients (both driven by dispersal and selection) in a spatially explicit model, and analysed simulated patterns with F_{ST} statistics and mean correlograms using Moran's *I* spatial autocorrelation coefficients. These statistics were compared with observed patterns for a tree species endemic to the Brazilian Cerrado. For a best match with observed F_{ST} (equal to 0.182), the important parameters driving simulated scenario are mainly related to population structure, including low population size with closed populations (low N_m), strong distance decay of gene flow, in addition to a strong effect of the initial variance of allele frequencies. These scenarios present a low autocorrelation of allele frequencies. Best matching of correlograms, on the other hand, appears in simulations with a large population size, high N_m and low population differentiation and F_{ST} (as well as more gene flow). Thus, targeting the two statistics (correlograms and F_{ST}) shows that best matches with empirical data with two distinct sets of parameters in the simulations, because observed patterns involve both a relatively high F_{ST} and significant autocorrelation. This conflict can be resolved by assuming that initial variance in allele frequencies can be interpreted as reflecting deep-time historical variation and evolutionary dynamics of allele frequencies, creating a relatively high level of population differentiation, whereas current patterns in gene flow creates spatial autocorrelation. This make sense in terms of the previous knowledge on population differentiation in *D. alata*, especially if patterns are explained by a combination of isolation-by-distance and allelic surfing due to range expansion after the last glacial maximum. This reveals the potential for more complex applications of POM in population genetics. © 2014 The Linnean Society of London, *Biological Journal of the Linnean Society*, 2014, **113**, 1152–1161.

ADDITIONAL KEYWORDS: autocorrelation – correlograms – F-statistics – geographical genetics – Pattern-Oriented Modelling – population structure – simulation.

Geographical patterns of genetic variation among populations, within species, or in closely related species have been widely used to infer potential microevolutionary processes driving population differentiation and speciation (Epperson, 2003; Rousset, 2004). Traditionally, the first approach to infer such processes is to evaluate the spatial heterogeneity among individuals or populations (i.e. the magnitude of population differentiation) as well as variation among loci using F_{ST} and related estimators (see

*Corresponding author. E-mail: diniz@ufg.br; jafdinizfilho@gmail.com

Holsinger & Weir, 2009). However, more complex patterns arise under alternative evolutionary processes, and despite problems due to the lack of full correspondence between patterns and processes, a better understanding of potential evolutionary scenarios that drive population differentiation can be achieved using several forms of multivariate and spatial statistical analyses (Sokal & Oden, 1978a, b; Epperson, 1995, 1996; Hardy & Vekemans, 1999; Manel *et al.*, 2003; Jombart *et al.*, 2008; Balkenhol, Waits & Dezzani, 2009; Guillot *et al.*, 2009; Diniz-Filho & Bini, 2012; Wagner & Fortin, 2013). These approaches allow for a more explicit evaluation of the spatial structure of genetic variation and go beyond a simple evaluation of thee level of differentiation (see Pearse & Crandall 2004 for a general discussion).

Several forms of multivariate hierarchical clustering and, more recently, model-based clustering under a Bayesian framework have been widely used to assess population structure (see Guillot *et al.*, 2009). The overall idea is to find clusters of individuals who maximise genetic variation among groups; in some more recent methods, this is done under biologically realistic assumptions and allows for the testing of specific hypotheses about barriers and gene flow pathways. On the other hand, explicit spatial analyses, such as Mantel correlations between genetic and geographical distances among local populations, autocorrelation indices and autoregressive models, eigenvector mapping and discontinuity statistics (see Guillot *et al.*, 2009; Diniz-Filho & Bini, 2012; Wagner & Fortin, 2013; Diniz-Filho *et al.*, 2009, 2013a for recent reviews), can go a step further and be used to describe genetic variation in a continuous space, evaluating how the combined distribution of loci are potentially responding to environmental gradients or anthropogenic disturbances, creating barriers to gene flow, or are being driven by historical processes that are related, for example, to range expansion (Excoffier & Ray, 2008).

In addition to statistical analyses of genetic data, there is a long tradition in geographical genetics of using computer simulations (in many cases coupled with spatial statistics) to understand the mechanisms underlying population differentiation and to evaluate the statistical performance of methods used to describe patterns of genetic variation and their ability to disentangle patterns and processes (see Epperson *et al.*, 2010 for a review). In a pioneering application, R. R. Sokal's research group (see Sokal & Oden, 1978a, b; Sokal & Wartenberg, 1983; Sokal, Harding & Oden, 1989; Sokal & Jacquez, 1991; Sokal, Oden & Thomson, 1997) used simulation experiments to show that spatial autocorrelation analyses would be useful for recovering evolutionary and ecological processes underlying population differentiation, including, for instance, Wright's (1943) Isolation-by-Distance (IBD) process (see also Epperson, 1995, 1996, Barbujani 1987 and Hardy & Vekemans 1999 for theoretical discussions and further methodological approaches).

One of the main difficulties in using simulation approaches to understand complex patterns of population differentiation is the difficulty (or even impossibility, in many cases) of parameterising the simulations with realistic biological data on the ecology and life-history traits of a species that could drive genetic variation. In several of Sokal's papers, simulations were used in a heuristic context to understand how analyses work and how they could be potentially used to recover processes under relatively simple scenarios. When dealing with real datasets, the situations were always more complicated, and the results were not interpretable in a straightforward manner (e.g. Sokal & Riska, 1981; Sokal & Menozzi, 1982; Sokal, Smouse & Neel, 1986; Sokal, Jacquez & Wooten, 1989). Thus, the observed data are usually so complex that simple models (such as IBD) would hardly fit the data, and a wide range of possibilities of parameter combinations would appear.

Although advances in likelihood and Bayesian theory and application would allow a direct evaluation of matches between data and models, the complexity of ecological and evolutionary systems, especially in terms of the need for evaluating multiple responses simultaneously at several hierarchical levels, usually makes this strategy very difficult (see Hartig *et al.*, 2011 for a recent review and discussions). One possible way to couple simulations and empirical data in complex systems is what ecologists have termed 'Pattern-Oriented Modelling' (POM hereafter; see Grimm *et al.*, 2007; Grimm & Railsback, 2012). The idea of POM is to use computer simulations to generate, for a large range of variation in several parameters, many possible realisations of the processes under study. Then, it is possible to evaluate and biologically or ecologically interpret the set of parameter values that gives the best match with empirical data. Thus, POM addresses complex ecological and evolutionary processes by creating a potentially infinite set of scenarios and subsequently using an empirical and computationally intense strategy to find matches between simulated and observed patterns under a likelihood-based epistemological reasoning.

Here, we used a POM approach to analyse the geographical patterns of genetic variation in microsatellite loci of *Dipteryx alata* (Fabaceae), a tree species with a large geographic range that is endemic to the savanna environments ('Cerrados') in Central Brazil. Previous studies (see Soares *et al.*, 2008; Collevatti *et al.*, 2010, 2013; Diniz-Filho *et al.*, 2012a, b, 2013a, b) showed significant spatial patterns of genetic variation in this species at distinct spatial scales, which seem to be related to a combination of broad-scale IBD processes and southeastern range expansion after the last glacial maximum. We used computer simulations to generate patterns of population differentiation that can be directly associated with distinct adaptive and neutral evolutionary processes driven by several parameters. It is then possible to search for the combinations of parameters that generate geographic patterns in allelic frequencies that produce the best match between simulated and empirical F_{ST} statistics and Moran's *I* spatial correlograms.

METHODS

Basic simulations

A script written in R (R Development Core Team, 2013) (see Appendix S1) was used to simulate deme-level dynamics of allele frequencies as a Markov process with exponential decrease of gene flow with geographical distances for several alleles and loci. The parameters used in the simulations and their range of variation are described in Table 1, and the source code is available from the main author upon request.

We started with initial allele frequencies *A* for a given number of loci, with mean 0.5 and a given standard deviation (sd_0) for setting the initial conditions (mimicking the effect of deep-time historical processes that set the initial distribution of allele frequencies) of each population, with *N* individuals each. At each generation, a given proportion of migrants (*m*) is used as a starting point to define the amount of gene flow among populations as a function of geographical distance. The decrease of gene flow is modelled as an exponential function of geographical distances among populations as

$$N_{out} < -N_m * (\exp((\alpha^* - 1) * \mathrm{D}))$$

where N_{out} is the effective number of individuals arriving in a given population, N_m is the initial theoretical number of individuals dispersing into the population and α is the distance decay parameter in respect to geographic distances **D** (these distances **D** are standardised to vary between 0 and 1). Deterministically, the allele frequency of a population in a given time is then given by a weighted mean of the allele frequencies in all populations based on the equation above, where the weights are $1 - N_m$ and N_m for the contributions to the allele frequencies from within the population and from all other populations, respectively. This last component is, in turn, given by the allele frequencies in all other populations in the system weighted by the N_{out}. In summary, the allele frequency of a given population is given by the geographically weighted average of the other neighbouring populations, akin to a simultaneous autoregressive process (see Epperson, 2003; Fortin & Dale, 2005), thus forming a migration matrix. The stochastic component of allele frequency variation, which expresses genetic drift within populations, is given by adding a value sampled from a binomial distribution with mean *A* and variance equal to $A^*(1 - A)/N$.

The process described above will approximate Wright's (1943) IBD at population scale, with the genetic differentiation exponentially increasing with geographical distances. However, it is also possible that some (or all) allele frequencies are simultaneously affected by other processes that generate directional clines. These clines are usually interpreted as evidence of selection (i.e. response to environmental factors) or migration and demic diffusion among populations with different initial allele frequencies (see Sokal *et al.*, 1989; Excoffier & Ray, 2008; Manel *et al.*, 2010a, b). We added a geographic gradient to the stochastic gradient by adding a deterministic value given by the relative position along a direction gradient (with coefficient $P = 1/(2^*t)$, so that an allele frequency would be fixed after *t* generations starting from a mean of 0.5) to the allele frequencies at each time step. The gradient can have different orientations in respect to latitude and longitude, so that the direction of the cline is given by a weighted mean of latitude and longitude, with coefficients equal to *g1* and *g2*, respectively. A high value of *g1* will generate a more latitudinal gradient, whereas a high *g2* will give a longitudinal gradient. The gradient is

Table 1. Parameters used in the simulation of population genetic structure of *D. alata* in the context of pattern-oriented modelling, including population size (*N*), number of immigrants at minimum geographic distance (*Nm*), distance decay of migration rate (α), weights of latitudinal and longitudinal gradients (*g1* and *g2*), number of alleles responding as gradients (*NAG*) and initial standard deviation of allele frequencies within loci (sd_0)

Model parameters	Simulation range		Realised range	
	Minimum	Maximum	Mean	sd
N	25	500	257.48	138.01
Nm	0.01	0.25	0.13	0.07
α	−2	−50	−29.42	12.12
g1	0	1	0.51	0.29
g2	0	1	0.50	0.29
NAG	0	16	8.53	4.33
sd_0	0.05	0.25	0.15	0.06

standardised to vary between 0 and 1 and then centred so that the allele frequencies will increase and decrease in the extreme positions along this gradient. In the simulations, we allowed the number of alleles responding the gradient (*NAG*) to vary among simulations, which can appear because selective agents drive variation in these alleles or because of long-distance migration connecting populations that differ in these alleles (see Diniz-Filho & Bini 2012; Diniz-Filho *et al*., 2012c for a recent review).

We run the simulation for 1000 generations for a given combination of parameters, and matching between simulated and empirical patterns was then performed (see below). A total of 10 000 random combinations of parameter values, with values randomly sampled from a uniform distribution within the limits given in Table 1, were generated.

Matching empirical data

The idea in POM is to compare simulated and empirical patterns using several criteria simultaneously. We performed the simulations above to match the empirical patterns of genetic variation of *Dipteryx alata* as best as possible. We analysed the empirical data of 54 allele frequencies from eight microsatellite loci in 644 individuals of *D. alata* from 25 populations in the Brazilian Cerrado (see Soares *et al*., 2012; Diniz-Filho *et al*., 2012a, b, 2013a). We fixed the number of populations and their geographical distribution (pairwise distances) in the simulations to match the empirical patterns for *D. alata*, and we allowed parameters to vary according to Table 1. Although this is only a sample of local populations within a species range, we expect that the dynamics of the microevolutionary processes driving genetic variation are captured by the parameters of the simulation process based on the migration matrix.

We then calculated, for each simulation, the F_{ST} statistics and the mean spatial correlogram (see Sokal & Oden, 1978a, b; Epperson, 2003) obtained by calculating Moran's *I* coefficients for five geographic distance classes with limits that were defined to maximise the similarity in the number of pairs of populations connected. The F_{ST} and correlograms were also obtained for each allele, and an average correlogram was computed for empirical microsatellite data. We obtained, for each simulation, the difference between observed and simulated F_{ST} values and the Manhattan distance between observed and simulated mean correlograms (i.e. the sum of absolute differences between Moran's *I* in the correlograms).

The parameters of the simulations that better match the observed F_{ST} and correlograms were then interpreted by comparing the best values (as well as the 5% combination of parameter values that generated results closest to the observed statistics) with simulation ranges and overall means. We also applied multiple regressions to the simulated data to evaluate which combination of parameters best explains the population differentiation measured by F_{ST} and the similarity (Manhattan distance) between observed and simulated correlograms. Notice that because we are targeting the two statistics independently, the simulated F_{ST} was used as an explanatory variable when analysing patterns in correlograms, whereas simulated spatial patterns (Moran's *I* in the first and last distance classes) can be used to explain variation in F_{ST} (see Table 2).

RESULTS

The observed F_{ST} for *D. alata* was equal to 0.182, and the mean correlogram follows a strong linear gradient, with a positive Moran's *I* in the first distance class coupled with a negative Moran's *I* in the last distance class (i.e. close populations are similar and populations that are geographically separated tend to be different) (Fig. 1). It is then possible to use the POM approach to evaluate which parameters are best matched with these observed patterns (Table 2).

Distances from the observed and simulated F_{ST} ranged from close to 0 (minimum equal to 0.02) to 0.35, with a skewed distribution (Fig. 2). Most simulations generated values much lower than the observed F_{ST} (in general, close to zero). On the other hand, the distribution of Manhattan distances (i.e. the differences between mean correlograms) was much less skewed, with mean distances of approximately 0.6 (Fig. 3) and a minimum distance equal to 0.047 (each simulated Moran's *I* in the correlogram is within 0.01 of the observed value on average). Notice that the differences in F_{ST} and Manhattan distances are not linearly correlated (Fig. 4), although a pattern can be observed in the bivariate space (i.e. when simulated correlograms are similar to the observed one, there is a wide range of possible F_{ST} values, but when simulated correlograms are different from the observed one, F_{ST} is always large), so the simulation parameters that produce the best empirical matches for these two sets of statistics are different.

Multiple regression coefficients allow for the determination of which parameters in the simulations best explain distance between the observed and simulated statistics. For F_{ST}, the variation in simulation parameters explained almost 70% of the difference between observed and simulated statistics. The highest coefficients indicate that the combination of simulation parameters that best explained distance to observed F_{ST} was a low population size (approximately 100, in the range of 25 to 500), lower N_m (approximately 0.01, in the range of 0 to 0.25, but varying widely even in

Table 2. Results of POM analysis matching observed F_{ST} values and mean spatial correlograms for *D. alata* in Brazilian Cerrado. Results show which values for simulation parameters give the BEST and the 5% best match for observed spatial correlograms (Fig. 1) and between simulated and observed F_{ST}. The model parameter and values include population size (*N*), number of immigrants at minimum geographic distance (*Nm*), distance decay of migration rate (α), weights of latitudinal and longitudinal gradients (*g1* and *g2*), number of alleles responding as gradients (*NAG*) and initial standard deviation of allele frequencies within loci (sd_0) (the simulation parameters) and the observed statistics obtained in the simulations, including Moran's I coefficients in the first and fifth distance classes and deviation of F_{ST} (D_FST) and Manhattan distance correlograms (Dist_CRLG) for observed and simulated data. Also shown are the t-values of multiple regression of the deviation of D_FST and Dist_CRLG against all other simulation parameters (see text for detail), with bold values indicating significant ($P < 0.05$) values

Model parameters and variables	Correlograms			F_{ST}			Multiple regressions	
	5%	(± sd)	BEST	5%	(± sd)	BEST	Correlograms	F_{ST}
N	160.077	130.199	335.000	101.955	98.009	150.000	**22.093**	**43.385**
Nm	0.084	0.066	0.072	0.094	0.070	0.016	**21.269**	**26.806**
A	−40.302	7.638	−48.746	−38.141	8.077	−31.458	**21.072**	**48.288**
Moran 1	0.077	0.067	0.084	0.001	0.005	0.000		**−2.387**
Moran 5	−0.078	0.020	0.078	−0.132	0.086	−0.079		**−28.057**
D_FST	−0.119	0.018	−0.116	−0.172	0.063	−0.160	**−13.657**	
g1	0.496	0.281	0.022	0.515	0.278	0.985	−0.135	−1.34
g2	0.520	0.277	0.867	0.502	0.307	0.960	**−2.227**	0.676
Dist_CRLG	0.081	0.012	0.047	0.242	0.157	0.146		
NAG	5.363	2.602	7.000	10.170	3.938	11.000	**41.083**	**−49.784**
sd_0	0.151	0.056	0.178	0.191	0.051	0.233	**2.53**	**−49.792**

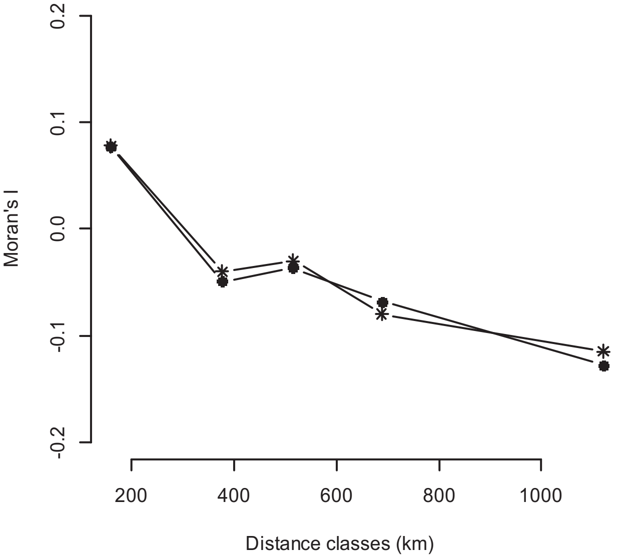

Figure 1. Observed mean correlogram (circles) and simulated mean correlogram (stars) for the best combination of parameters (see Table 2) using a POM approach for *D. alata* in Brazilian Cerrado.

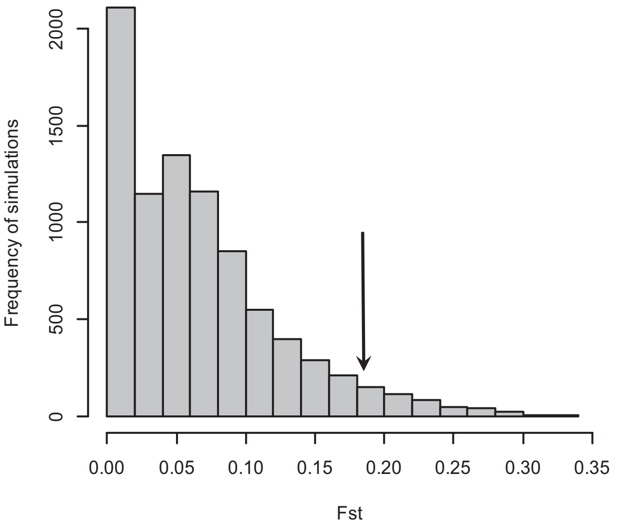

Figure 2. Distribution of 10 000 simulated F_{ST} values under random combination of the parameters in Table 1. Arrow indicates observed F_{ST} for *D. alata* in Brazilian Cerrado.

the best 5%), more negative distance decay of gene flow (−38, in the range of −2 to −50), the absence of autocorrelation (both in first and last distance class, but particularly low Moran's *I* in the first distance class), a high number of linear gradients (approximately ten in the range of 0 to 16, but without any direction effects) and a high initial variation in allele frequencies (deviation from 0.5 initial conditions equal to 0.23, in the range of 0.05 to 0.25).

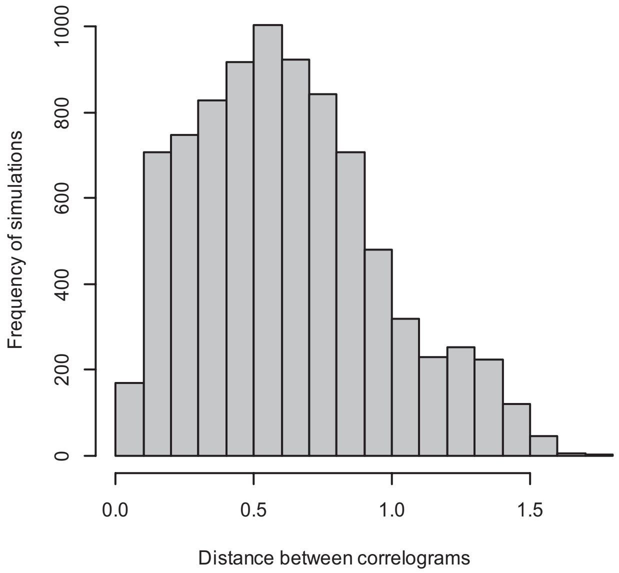

Figure 3. Distribution of 10 000 Manhattan distances between observed and simulated mean correlograms under random combination of the parameters in Table 1.

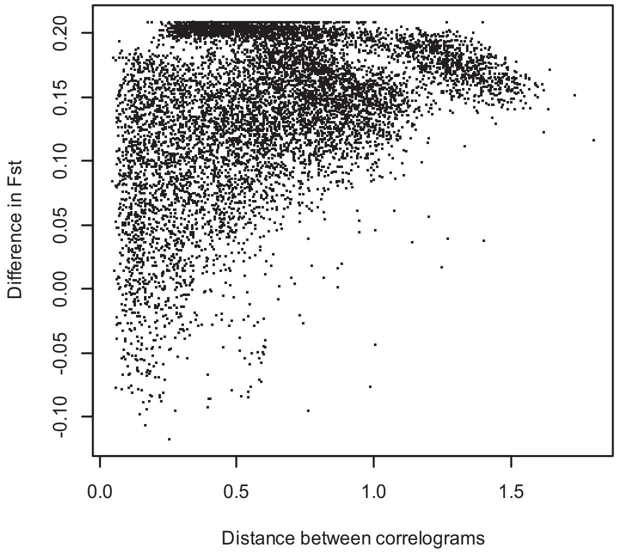

Figure 4. Relationship between difference between simulated and observed F_{ST} values for *D. alata*) and Manhattan distance between mean simulated and observed correlograms across 10 000 simulations using random combination of the parameters in Table 1.

The explanatory power of Manhattan distances between correlograms was not as high as for F_{ST}, with an adjusted R^2 of 0.386. Again, high coefficients were recorded for population size, but the opposite was observed when compared to the effect for F_{ST} with a higher population size (a mean of 160, although the best match was achieved when *N* equals 335). The best match also occurs at lower than the mean N_m (approximately 0.08, in the range of 0 to 0.25) and at a very high negative distance decay of gene flow (−48.7, in the range of −2 to −50). The parameter with the highest regression coefficient was the number of gradients in the simulation, with the best match achieved when this number equals 7 (in the range 0 to 16, again with low effects of the direction of the gradient).

The parameters that produce the best matches for both F_{ST} and the mean correlogram are, in some sense, an average of the patterns described above (see arrow in Fig. 4; Table 2). The population size is low (37, close to minimum parameter allowed in the simulation), with low N_m, high distance decay of gene flow and low spatial autocorrelation as well as 7 (out of 55) allele frequencies, with a slight effect of the direction of the gradient, mainly in the longitudinal (east–west) component.

DISCUSSION

Pattern-oriented modelling and geographical patterns of population differentiation

Inferring evolutionary processes from empirical patterns of genetic differentiation among populations in geographical space has always been a challenge because of the lack of correspondence in this relationship (e.g. Sokal & Oden, 1978a, b; Epperson, 2003; Rousset, 2004). The chief difficulty is that the same observed pattern can be generated by several processes, and of course this mismatch increases as patterns become more complex. So, our goal here is to use the POM approach to explain an observed pattern by investigating what patterns may arise under varying parameter combinations in computer simulations. In a more practical context, the simulations performed here within the POM framework, as proposed by Grimm *et al.* (2007; see also Rangel, Diniz-Filho & Colwell, 2007), also reveal that patterns observed for the same species and targeting distinct statistics (i.e. the magnitude of population differentiation estimated by F_{ST} and the spatial pattern measured by the spatial correlograms) provide complementary evidence about the processes underlying genetic variation.

The multiple regression model shows that similar patterns in F_{ST} and correlograms appear under distinct scenarios and parameter combinations. For replicating F_{ST} patterns, the processes are mainly related to population structure (as expected) and include a low population size with closed populations (low N_m) and strong distance decay, in which gene flow will occur only at short distances. There is also a strong effect of the initial variance of allele frequencies, reflecting the importance of strong stochasticity under the initial conditions. The combination of parameters described above will generate low levels of autocorrelation in allele frequencies, even at smaller geographical distances, because even if gene flow occurs at short geographical distances, high genetic drift and strong variation under the initial conditions will tend to increase randomness in the allele frequencies and eliminate autocorrelation (Sokal & Jacquez, 1991; see also Epperson, 1995, 1996, 2005). This is in some sense compatible with Sewall Wright's 'island model', which was the basis for defining *F*-statistics.

However, despite the relatively high observed F_{ST}, low levels of spatial autocorrelation are not observed in our empirical data, and this explains why a different set of parameters than those found for F_{ST} will appear when best matches are sought on the basis of correlograms. These parameters include high population size and higher N_m (suggesting more gene flow) and some variables with directional gradients. At the same time, under the POM approach, simulations will generate the spatial patterns observed here only if population differentiation is low, as expected by relatively large local populations linked by gene flow from neighbours. Thus, targeting the two statistics will provide the best matches with empirical data with two distinct sets of parameters in the simulations, and of course, this conflict must somehow be resolved (if, for example, the idea is to find a unique and integrated explanation for the patterns).

When trying to merge these results and identify which combinations of parameters produce results that best match both statistics, the patterns found for F_{ST} appear with more intensity. This is expected, considering that the explanatory power of simulation parameters with respect to empirical patterns is much higher for F_{ST}, according to multiple regression-adjusted R^2. However, it is interesting that, in this case, the direction of the gradient has a slightly larger effect.

Despite conflicting results, the parameter combinations found here make sense in terms of theoretical expectations about population differentiation, in which high F_{ST} values are expected for highly isolated populations, whereas autocorrelation will arise if drift is not that high, and higher levels of gene flow among neighbouring populations chiefly determine allele frequencies (see Sokal & Jacquez, 1991; Sokal *et al*., 1997). Additionally, some gradients will improve the autocorrelation pattern, mainly by generating long-distance (negative) autocorrelation (Sokal *et al*., 1989). More importantly, one of the most important drivers of the F_{ST} pattern is the high variance in initial allele frequencies, and this may help in understanding the two sets of parameters obtained when targeting the two statistics. By assuming that strong stochasticity in initial conditions is important for understanding, F_{ST} can be directly interpreted as reflecting deep-time historical variation and evolutionary dynamics of allele frequencies (explicitly adding this historical dimension into analyses is one of the difficulties involved in coupling phylogeographical patterns and population structure – see Felsenstein, 1982; Avise, 2009; Holsinger & Weir, 2009; Chan, Brown & Yoder, 2011; Croucher, Oxford & Gillespie, 2011). Thus, we understand that F_{ST} values reveal older patterns that reinforce deep divergences between populations, whereas correlograms reflect more recent spatial dynamics of allele frequencies.

Patterns in dipteryx alata

Biologically, the results of our POM also agree with previous spatial analyses of genetic variation in plants (e.g. Hardy *et al*., 2006), particularly within and among local populations of *D. alata* in the Brazilian Cerrado (see Collevatti *et al*., 2010, 2013; Diniz-Filho *et al*., 2012a, 2013a). The 'baru' tree *D. alata* is a widely distributed species, and its geographic range covers most of the Cerrado region in Central Brazil. Despite this, it is usually restricted to particular environments within this range, being more frequently found in seasonal savannas and growing in eutrophic, drained soils. Thus, the local density is not expected to be very high, although precise estimates are not available. Previous autocorrelation analyses revealed the spatial genetic structure at local spatial scales within local populations, as expected for a species with its reproductive and life-history characteristics (Collevatti *et al*., 2010). The species is hermaphroditic and self-incompatible, and pollination is performed by large/medium-sized bees. The seeds have a woody endocarp and an edible nut inside, which makes the seed attractive to bats and monkeys and allow for its dispersal. These dispersal patterns, coupled with restricted distributions in a wide geographic range, tend to generate isolation among populations when expanded to broader spatial scales (forming a pattern akin to an 'island model') or an exponential decrease in genetic similarity among populations with increasing geographic distances. This will generate, in the long term, a high level of population differentiation (i.e. high F_{ST}), as observed in our simulations.

However, in these simulations, strong differentiation is usually not coupled with a clinal geographic pattern in allele frequencies. Thus, to understand the observed pattern (i.e. a high level of population differentiation coupled with clinal patterns in allele frequencies), it is necessary to add historical processes of range expansion that could generate clines. Previous analyses using variance partition of genetic similarity into turnover and nestedness components of allele richness (Diniz-Filho *et al*., 2012a, 2013) suggest that clines in *D. alata* allele frequencies can be explained by allele surfing due to southward range expansion after the last glacial maximum, tracking environmental changes in Central South America (at the same time that populations are maintained in several stable areas of the Cerrado). This suggestion is also supported by evaluating historical levels of gene flow, which are not linked with simple geographic distances (Collevatti *et al*., 2013). This also allows an understanding of the effect of the longitudinal component found when matching the spatial correlograms.

The importance of historical effects is also reinforced because F_{ST} patterns are recovered by the POM approach only when there is high variation in the initial allele frequencies, as expected in a species with a long history in which alleles have different frequencies because of their long divergence time (i.e. the evolutionary dynamics of alleles will drive a strong variation in initial conditions, with newer alleles occurring at lower frequencies). In a sharp contrast with deep-time historical effects, there would also be more recent effects due to human occupation (i.e. Soares *et al*., 2008; Telles *et al*., 2014), which are much more difficult to analyse using the POM developed here. However, a more thorough exploration of the effect of initial conditions and the evolution of alleles requires a much more complex POM.

CONCLUDING REMARKS

The analyses performed here represent a first step in using POM in population genetics, producing patterns that are both consistent with theoretical expectations from population genetics and with previous studies of genetic variation at distinct geographical scales in *D. alata*. Further improvements can be made to the POM approach used here, both in the sense of making more realistic simulations and in the statistics evaluated. In the former, it would be possible to simulate a particular type of molecular marker (rather than using SSR allele frequencies as overall surrogates, as performed here), with its own characteristics in terms of evolutionary models and rates. Deep-time phylogeographical patterns could also be investigated, with simulations also involving range shifts, perhaps based on ecological niche models and alternative environmental hindcasting (see Nogués-Bravo, 2009; Collevatti *et al*., 2013). Despite introducing more complexity in the simulation model, incorporating deep-time phylogeographical patterns in the simulations may also help integrate the different results obtained when targeting the distinct statistics in POM, as previously discussed. An even more complex approach, which actually better fits the original tradition of using POM in ecological analyses, is to perform analyses at the individual level rather than at the population level, as performed here. Each of these additions may require more statistics to be evaluated under a POM approach. Of course, adding these multiple new sets of variables and statistics will increase the complexity of the simulation and analyses, so a careful evaluation of parsimony (the 'Medawar zone', as noted by Grimm *et al*., 2007), in terms of balancing complexity and gain of knowledge, should be performed.

ACKNOWLEDGEMENTS

We thank Thiago Fernando Rangel for many discussions about POM and Helene Wagner for critical reading of a previous version of the manuscript. This work was supported by several grants and fellowships to the research network GENPAC (Geographical Genetics and Regional Planning for natural resources in Brazilian Cerrado) from CNPq/MCT/CAPES (projects # 564717/2010-0 and 563624/2010-8) and by the 'Núcleo de Excelência em Genética e Conservação de Espécies do Cerrado' – GECER (PRONEX/FAPEG/CNPq CP 07-2009). Field work has been supported by Systema Naturae Consultoria Ambiental LTDA. Our work has been also continuously supported by productivity fellowships from CNPq.

REFERENCES

Avise JC. 2009. Phylogeography: retrospect and prospect. *Journal of Biogeography* **36:** 3–15.

Balkenhol N, Waits LP, Dezzani RJ. 2009. Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* **32:** 818–830.

Barbujani G. 1987. Autocorrelation of gene frequencies under isolation-by-distance. *Genetics* **177:** 772–782.

Chan LM, Brown JL, Yoder AD. 2011. Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Molecular Phylogenetics and Evolution* **59:** 523–537.

Collevatti RG, Lima JS, Soares TN, Telles MPC. 2010. Spatial genetic structure and life-history traits in Cerrado tree species: inferences for conservation. *Natureza & Conservacao* **8:** 54–59.

Collevatti RG, Telles MPC, Nabout JC, Chaves LJ, Soares TN. 2013. Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical savannas. *Heredity* **111:** 97–105.

Croucher PJP, Oxford GS, Gillespie RG. 2011. Population structure and dispersal in a patchy landscape: nuclear and mitochondrial markers reveal area effects in the spider *Theridion californicum* (Araneae: Theridiidae). *Biological Journal of the Linnean Society* **104:** 600–620.

Diniz-Filho JAF, Bini LM. 2012. Thirty-five years of spatial autocorrelation analysis in population genetics: an essay in honour of Robert R. Sokal (1926–2012). *Biological Journal of the Linnean Society* **105:** 721–736.

Diniz-Filho JAF, Collevatti RG, Soares TN, Telles MPC. 2012a. Geographical patterns of turnover and nestedness-resultant components of allelic diversity among populations. *Genetica* **140:** 189–195.

Diniz-Filho JAF, Diniz JV, Rangel TF, Soares TN, Telles MPC, Collevatti RG, Bini LM. 2013a. A new eigenfunction spatial analysis describing population genetic structure. *Genetica* **141:** 479–489.

Diniz-Filho JAF, Melo DB, Oliveira G, Collevatti RG, Soares TN, Nabout JC, Lima JS, Dobrovolski R, Chaves LJ, Naves RV, Loyola RD, Telles MPC. 2012b. Planning for optimal conservation of geographical genetic variability within species. *Conservation Genetics* **13:** 1085–1093.

Diniz-Filho JAF, Nabout JC, Telles MPC, Soares TN, Rangel TFLVB. 2009. A review of techniques for spatial modeling in geographical, conservation and landscape genetics. *Genetics and Molecular Biology* **32:** 203–211.

Diniz-Filho JAF, Siqueira T, Padial AA, Rangel TF, Landeiro VL, Bini LM. 2012c. Spatial autocorrelation allows disentangling the balance between neutral and niche processes in metacommunities. *Oikos* **121:** 201–210.

Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, Telles MPC, Rangel TF, Bini LM. 2013b. Mantel test in population genetics. *Genetics and Molecular Biology* **36:** 475–485.

Epperson BK. 1995. Spatial distribution of genotypes under isolation by distance. *Genetics* **140:** 1431–1440.

Epperson BK. 1996. Measurement of genetic structure within populations using Moran's I spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences of the United States of America* **93:** 10528–10532.

Epperson BK. 2003. *Geographical genetics*. Princeton: Princeton University press.

Epperson BK. 2005. Estimating dispersal from short distance autocorrelation. *Heredity* **95:** 7–15.

Epperson BK, McRae B, Scribner K, Cushman SA, Rosenberg MS, Fortin M-J, James PMA, Murphy M, Manel S, Legendre P, Dale MRT. 2010. Utility of computer simulations in landscape genetics. *Molecular Ecology* **19:** 3549–3564.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution* **23:** 347–351.

Felsenstein J. 1982. How can we infer geography and history from gene frequencies. *Journal of Theoretical Biology* **96:** 9–20.

Fortin M-J, Dale MRT. 2005. *Spatial analysis: a guide for ecologists*. Cambridge: Cambridge University Press.

Grimm V, Railsback SF. 2012. Pattern-oriented modelling: a 'multi-scope' for predictive systems ecology. *Philosophical Transactions of Royal Society B: Biological Science* **367:** 298–310.

Grimm V, Revilla E, Berger U, Jeltsch F, Mooij WM, Railsback SF, Thulke H-H, Weiner J, Wiegand T, DeAngelis DL. 2007. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science* **310:** 987–991.

Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. *Molecular Ecology* **18:** 4734–4756.

Hardy OJ, Maggia L, Bandou E, Breyne P, Caron H, Chevallier MH, Doligez A, Dutech C, Kremer A, Latouche-Halle C, Troispoux V, Veron V, Degen B. 2006. Fine-scale genetic structure and gene dispersal inferences in 10 neotropical tree species. *Molecular Ecology* **15:** 559–571.

Hardy OJ, Vekemans X. 1999. Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Genetics* **83:** 145–154.

Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. 2011. Statistical inference for stochastic simulation models: theory and applications. *Ecology Letters* **14:** 816–827.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Review Genetics* **10:** 639–650.

Jombart T, Devillard S, Dufour A-B, Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101:** 92–103.

Manel S, Joost S, Epperson BK, Holderegger R, Storfer A, Rosenberg MS, Scribner K, Bonin A, Fortin MJ. 2010a. Perspective on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology* **19:** 3760–3772.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R. 2010b. Common factors drive adaptive genetic variation at different scale in *Arabis alpina*. *Molecular Ecology* **19:** 2896–2907.

Manel S, Schwartz MK, Luikart G, Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution* **15:** 189–197.

Nogués-Bravo D. 2009. Predicting the past distribution of species climate niche. *Global Ecology and Biogeography* **18:** 521–531.

Pearse DE, Crandall KA. 2004. Beyond F_{ST}: analysis of population genetic data for conservation. *Conservation Genetics* **5:** 585–602.

R Development Core Team. 2013. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.r-project.org

Rangel TFLVB, Diniz-Filho JAF, Colwell RK. 2007. Species richness and evolutionary niche dynamics: a spatial pattern-oriented simulation experiment. *American Naturalist* **170:** 602–616.

Rousset F. 2004. *Genetic structure and selection in subdivided population*. Princeton: Princeton Univ. Press.

Soares TN, Chaves LJ, Telles MPC, Diniz-Filho JAF, Resende LV. 2008. Landscape conservation genetics of *Dipteryx alata* (‘“baru”’ tree: Fabaceae) from Cerrado region of central Brazil. *Genetica* **132:** 9–19.

Soares TN, Melo DB, Resende LV, Vianello RP, Chaves LJ, Collevatti RG, Telles MPC. 2012. Development of microsatellite markers for the Neotropical tree species *Dipteryx alata* (Fabacea). *American Journal of Botany* **99:** e72–e73.

Sokal RR, Harding RM, Oden NL. 1989. Spatial patterns of human gene frequencies in Europe. *American Journal of Physical Anthropology* **80:** 267–294.

Sokal RR, Jacquez GM. 1991. Testing inferences about microevolutionary processes by means of spatial autocorrelation analysis. *Evolution* **45:** 152–168.

Sokal RR, Jacquez GM, Wooten MC. 1989. Spatial autocorrelation analysis of migration and selection. *Genetics* **121:** 845–855.

Sokal RR, Menozzi P. 1982. Spatial autocorrelation of HLA frequencies in Europe support demic diffusion of early farmers. *American Naturalist* **119:** 1–17.

Sokal RR, Oden N, Thomson BA. 1997. A simulation study of microevolutionary inferences by spatial autocorrelation analysis. *Biological Journal of the Linnean Society* **60:** 73–93.

Sokal RR, Oden NL. 1978a. Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society* **10:** 199–228.

Sokal RR, Oden NL. 1978b. Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society* **10:** 229–249.

Sokal RR, Riska B. 1981. Geographic variation in *Pemphigus populitransversus* (Insecta: Aphididae). *Biological Journal of the Linnean Society* **15:** 201–233.

Sokal RR, Smouse P, Neel JV. 1986. The genetic structure of a tribal population, the Yanomama indians. XV. Patterns inferred by autocorrelation analysis. *Genetics* **114:** 259–287.

Sokal RR, Wartenberg DE. 1983. A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105:** 219–237.

Telles MPC, Dobrovolski R, Souza KS, Lima JS, Collevatti RG, Soares TN, Chaves LJ, Diniz-Filho JAF. 2014. Disentangling landscape effects on population genetic structure of a neotropical savanna tree. *Natureza & Conservacao* **12:** 65–70.

Wagner H, Fortin MJ. 2013. A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics* **14:** 253–261.

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

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

Appendix S1. R-script for simulations of population genetic structure using a pattern-oriented modelling approach.