Comparing Phylogeographic Hypotheses by Simulating DNA Sequences under a Spatially Explicit Model of Coalescence

Simon Dellicour,*[†,‡,1] Chedly Kastally,[†,1] Olivier J. Hardy,[1] and Patrick Mardulyn[1]

[1]Evolutionary Biology and Ecology, Université Libre de Bruxelles, Brussels, Belgium

[†]These authors contributed equally to this work.

[‡]Present address: Department of Zoology, University of Oxford, Oxford, United Kingdom

***Corresponding author:** E-mail: simon.dellicour@zoo.ox.ac.uk.

Associate editor: Anna Di Rienzo

Abstract

Computer simulations of genetic data are increasingly used to investigate the impact of complex historical scenarios on patterns of genetic variation. Yet, in most empirical studies, relatively large portions of species ranges are often treated as panmictic populations, ignoring the underlying spatial context. In some cases, however, a more accurate spatial model is required. We use a spatially explicit model of coalescence (easily constructed by overlaying a two-dimensional grid on maps displaying an estimate of past and current species ranges) to evaluate the potential of several summary statistics to differentiate three typical phylogeographic scenarios. We first explore the variation of each summary statistic within the boundaries of each phylogeographic scenario, and identify those that appear most promising for a comparison of historical scenarios and/or to infer historical parameters. We then combine a selected set of summary statistics in a single chi-square statistic and evaluate whether it can be used to differentiate past geographic fragmentation or range expansion from a simple scenario of isolation by distance. We also investigate the benefits of using a spatially explicit model by comparing its performance to alternative models that are less spatially explicit (lower geographic resolution). The results identify conditions in which each summary statistic is useful to infer the evolution of a species range, and allow us to validate our spatially explicit model of coalescence and our procedure to compare simulated and observed sequence data. We also provide a detailed description of the spatially explicit model of coalescence used, which is currently lacking.

Key words: **DNA sequences, coalescence simulations, phylogeography, summary statistics, PHYLOGEOSIM 1.0.**

Introduction

Phylogeographic studies rely increasingly on models of coalescence to analyze DNA sequence variation in a historical context, either through the derivation of a likelihood or probability function (e.g., Hey and Nielsen 2004; Hey 2010), or through simulation of data (Arenas 2012; Hoban et al. 2012), to compare hypotheses and/or to estimate historical parameters under the constraint of a given evolutionary hypothesis. In most empirical studies, large portions of a species range, if not the entire range, are treated as single panmictic populations, thereby ignoring the limited capacity of movement of individuals within it (but see exceptions in, e.g., Currat and Excoffier 2005, 2011; Estoup et al. 2010). Although this classic approach offers a good approximation for evaluating many evolutionary hypotheses, in some cases, it is desirable to rely on models integrating a more detailed description of the geographic distribution of natural populations (e.g., Excoffier et al. 2009; Meirmans 2012). Indeed, a species range (or subsets of species ranges) characterized by a continuous distribution is (are) not naturally subdivided into separate panmictic populations, and delimiting populations in that context will rely on arbitrary decisions for the placement of population boundaries. Moreover, when estimating the evolution of a species range over time, for example in response to climate changes, it becomes crucial to include a more accurate description of that range in the model, and a way of accounting for limited dispersal of individuals within continuous portions of the range. A spatially explicit model of coalescence will not only allow evaluation of alternative hypotheses regarding the detailed geographic distribution of a species, but will also allow the use of features from the current geographic distribution of genetic variation for the evaluation process.

The increase in model complexity associated with using a spatially explicit model of coalescence, however, forces the use of computer simulations to infer population histories from patterns of genetic variation (as opposed to full likelihood or Bayesian approaches). Genetic data are simulated according to specified evolutionary hypotheses and compared with observed genetic data through the calculation of one or more summary statistics to estimate the probability that the evolutionary history associated with these hypotheses could have generated them (e.g., Tavaré et al. 1997). For each hypothesis tested, simulations need to be computed for a large array of parameter values, thereby spanning a vast number of combinations of possible parameter values in the most efficient possible way. For this task, the Approximate Bayesian Computation (ABC) approach has become the method of choice, and is used to compare alternative historical hypotheses and/or to estimate specific historical parameters (Beaumont et al. 2002; Bertorelle et al. 2010).

Specifically for studying the evolution of a species range, we previously developed a simple spatially explicit model of coalescence, represented by a two-dimensional grid that can be overlaid on a map, and on which we identify grid cells accessible to individuals over time (Dellicour, Mardulyn, et al. 2014; Dellicour, Fearnly, et al. 2014). In this model, each grid cell is treated as a separate panmictic population, and the probability of migration between adjacent cells is specified. The size of the grid cells is chosen by relying on prior assumptions over the ability of the studied organism to disperse, but also to limit the total amount of cells to a reasonable number. Here, we rely on this model to investigate the usefulness of several summary statistics for estimating historical parameters, and for comparing alternative detailed hypotheses, related to the evolution of a species range. For this purpose, we evaluate an approach introduced in Dellicour, Mardulyn, et al. (2014) that combines several statistics to compare empirical and simulated data. We also formally evaluate the added value of using a spatially explicit model of coalescence, by comparing it to alternative models that are less spatially explicit (lower geographic resolution). Finally, we provide a detailed description of our model, which is currently lacking.

More specifically, we use this model to simulate DNA sequences, with our program PHYLOGEOSIM 1.0, under three basic geographically explicit historical scenarios, but keeping the same pattern of sequence sampling for all scenarios (fig. 1): 1) a species characterized by a geographic distribution stable over time and isolation by distance ("IBD" scenario), 2) a typical history of geographic fragmentation ("GF" scenario), and 3) a typical history of range expansion ("RE" scenario). These sequences are used to investigate the variation of several standard and/or promising summary statistics under varying model conditions, and to test whether it is possible to rely on a set of such statistics to identify the past demographic scenario that generated them. For comparing the alternative phylogeographic hypotheses, we combine a selected set of summary statistics calculated from the data into a single chi-square statistic, and use a nonparametric test that returns a single *P*-value, as proposed in Dellicour, Mardulyn, et al. (2014). To demonstrate the benefits of a spatially explicit model, we estimate divergence time from pseudo-observed data simulated under GF or RE, by comparing these with data simulated under the same scenario, but gradually shifting from a fully spatially explicit model toward models with lower geographic resolution (lower number of cells), resulting in much larger portions of the range considered as panmictic populations.

New Approaches

This article investigates historical inference in a geographic context, using a recently introduced spatially explicit model of coalescence (Dellicour, Mardulyn, et al. 2014; Dellicour, Fearnly, et al. 2014). We evaluate a procedure (Dellicour, Mardulyn, et al. 2014) combining a selected set of statistics summarizing the main features of the data into a single chi-square statistic for the purpose of comparing observed and simulated data. The chi-square statistic is computed for a set of model parameter values chosen a priori from, and evenly distributed within, their prior range. Although this approach is simpler than the more thorough random exploration of model parameter space implemented in the popular approximate Bayesian computation (ABC) approach, it is suggested here as a useful alternative because simulations along spatially explicit models are too slow to allow the high number of simulations required by the ABC process. A detailed description of the spatially explicit model of coalescence used is also provided.

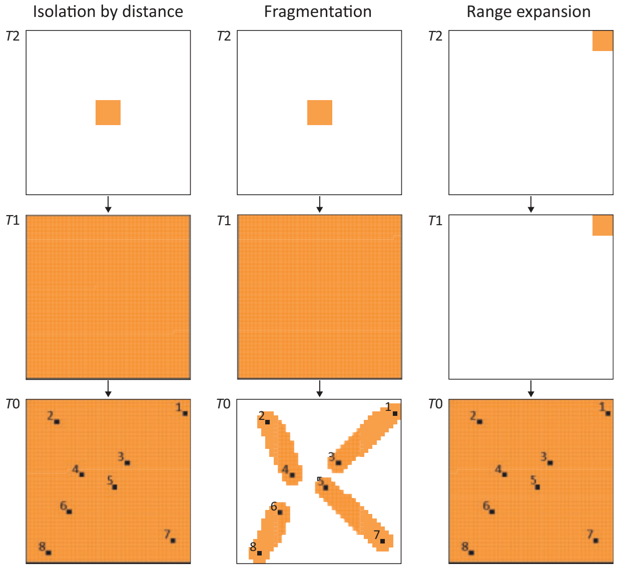


FIG. 1. Three theoretical phylogeographic scenarios under which we explore the behavior of several summary statistics. Each scenario is represented by three successive grids, the third one corresponding to the current time ($t = 0$). For all scenarios, we conducted coalescence simulations of sequences sampled in the same eight grid cells (ten sequences per cell), each colored in black and identified by a different number. Note that in the RE scenario, while the available range is instantaneously modified at *T*1 (forward simulation), colonization of new cells takes time, the number of generations involved depending on the migration rate (*mf*).

Results and Discussion

Summary Statistics Evaluation

Variation in summary statistics calculated for different model parameter values within the three historical scenarios is shown in figure 2. For the RE scenario, only graphs generated with a reproduction rate t_R (reproduction rate) = 2 are presented in figure 2, whereas those generated under t_R = 1.1, 4, 7, and 10 are available as supplementary material (supplementary figs. S4–S7, Supplementary Material online). Our detailed interpretations of these results, given separately for each statistic and each scenario, are summarized in supplementary table S1, Supplementary Material online.

Those statistics that display the largest variation within a graph are those that are most promising for inferring historical parameters from DNA sequence data. The variation displayed by each statistic is explained not only by the variation

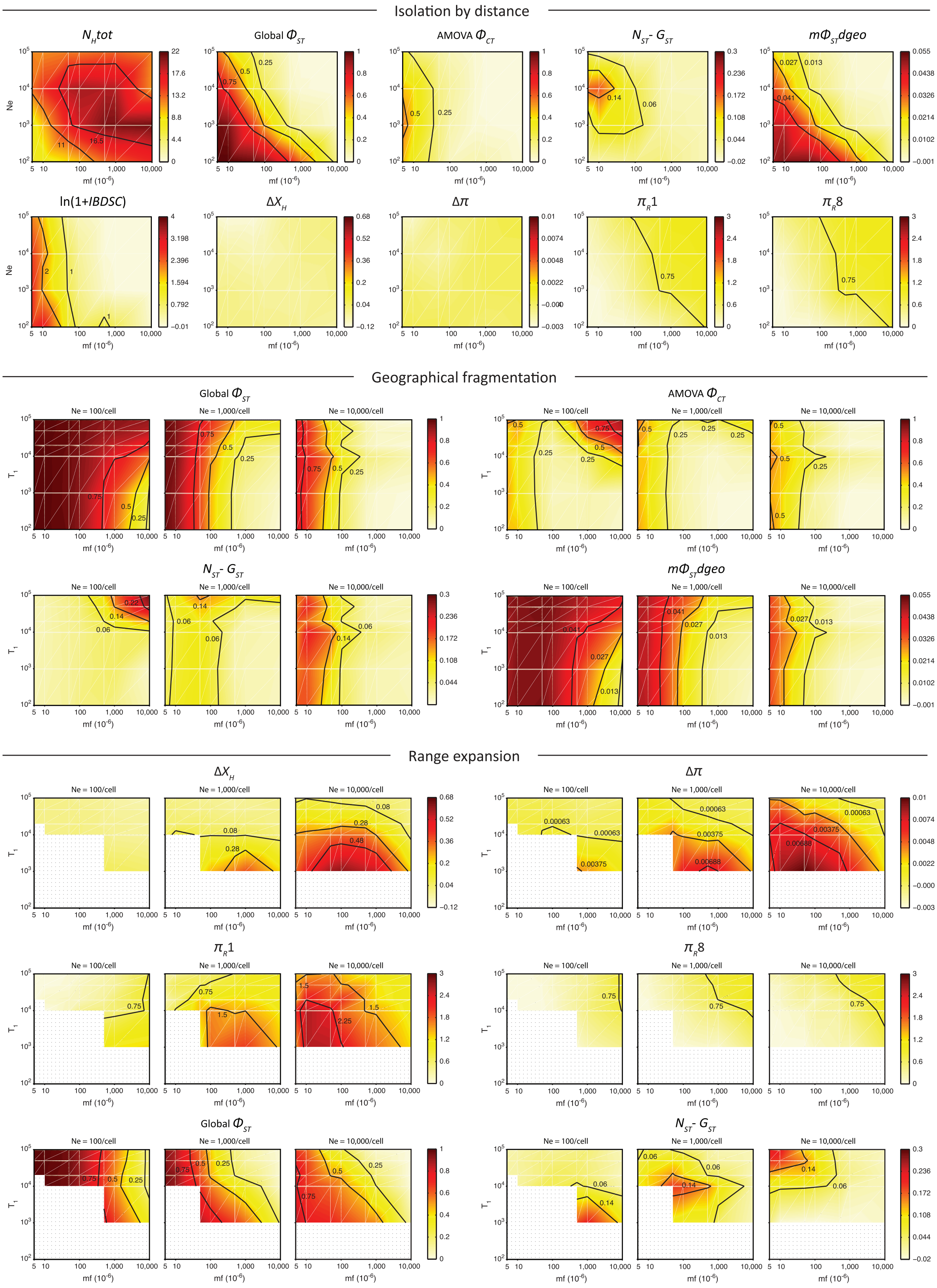


Fig. 2. Variation of the mean of summary statistics within the boundaries of the IBD, GF, and RE scenarios described in figure 1, with a reproduction rate $t_R = 2$. Statistics were computed from 100 simulations for each tested set of simulation parameters (see table 4). For the IBD scenario, the maximal grid cell effective sizes *Ne* vary along the *y* axis and the forward migration rate *mf* along the *x* axis; for the GF scenario, the time *T*1 at which the range fragmentation occurs varies on the *y* axis and the forward migration rate *mf* along the *x* axis; and for the RE scenario, the time *T*1 at which the RE occurs varies on the *y* axis and the forward migration rate *mf* along the *x* axis. In the RE graphs, white areas with dots correspond to sets of parameter values for which we could not compute the statistic because they did not allow the RE to reach all sampled cells. See supplementary figures S1–S3, Supplementary Material online, for the associated standard deviations of these statistics.

in model parameters but also by the stochastic nature of the evolutionary process (see standard deviation of each statistic in supplementary figs. S1–S3, Supplementary Material online). To highlight the impact of each source of variation, we performed linear regressions for each summary statistic 1) against all simulation parameters and 2) against each simulation parameter taken separately. The adjusted coefficients of determination (R^2) estimated for these linear regressions are reported in table 1 and provide an estimate of the proportion of variation explained only by variation in model parameters within each scenario. Although several of the studied statistics show little variation in the range of parameter values tested (e.g., *IBDSC*), others display clear patterns of variation within and/or among scenarios (e.g., global Φ_{ST}, ΔX_H), and thus appear promising, either to infer historical parameters assuming a specific historical scenario or to compare the likelihood of alternative historical scenarios given observed DNA sequence variation data. Overall, each summary statistic appears useful only on a portion (large or small, depending on the statistic) of the parameter space investigated that can be identified on figure 2.

It is worth noting that $N_{ST} - G_{ST}$, usually interpreted as a measure of "phylogeographic signal" (also called "phylogeographic structure"; Pons and Petit 1996) and widely used in phylogeographic studies, varies only for a restricted range of parameter values within the IBD scenario: the value of this statistic increases only when associated with a low forward migration rate (*mf*) and a maximal cell effective size (*Ne*) around 10,000. In fact, when populations approach fixation, both N_{ST} and G_{ST} approach one and their difference cannot reveal phylogeographic patterns. Over most of the graph depicted for this statistic, it appears insensitive to the implemented level of isolation by distance. In other words, to allow for different parts of the distribution to become differentiated by mutation events, migration among these regions needs to be very small. On the other hand, global Φ_{ST} statistics (Excoffier et al. 1992), as well as the related $m\Phi_{ST}dgeo$ statistic (average ratio between Φ_{ST}'s and geographical distances), varied much more across the graph associated with the IBD scenario. The relationship between Φ_{ST} and *Ne***mf* is already well known (e.g., Wright 1969) and is confirmed by our simulations. The *IBDSC* statistic (isolation by distance slope coefficient based on Rousset 1997; see table 2) strongly varies with migration rates in the IBD scenario when migration is weak, but its increase correlates with a strong increase of the associated standard deviation (supplementary fig. S1, Supplementary Material online), which limits its usefulness to estimate migration rates. The pattern of variation shown by $m\Phi_{ST}dgeo$ is always similar to that of Φ_{ST}, so that taking geographic distances among localities into account (as is the case for calculating $m\Phi_{ST}dgeo$) appears to offer little extra information, at least under the investigated sampling design. In contrast, in a phylogeographic study of the flat-tailed horned lizard in which sampled populations are less evenly spaced, Mulcahy et al. (2006) identified a nonproportional relationship between Φ_{ST} and a statistic similar to $m\Phi_{ST}dgeo$, suggesting that this statistic could be more useful under other sampling conditions.

Under the RE scenario, the reproduction rate parameter t_R would intuitively appear important, because it determines the proportion of new individuals in a cell that comes from reproduction (of individuals in that cell) compared with those entering the cell through migration. In a newly colonized cell, for which the effective size increases at each generation (before reaching its maximum capacity), a high reproduction rate combined with low migration should result in low diversity per cell but high differentiation among cells, whereas a low reproduction rate combined with high migration should result in low differentiation and high diversity. However, despite these considerations, it appears, based on figure 2 and supplementary figures S4–S7, Supplementary Material online, that the reproduction rate has only a weak influence on the calculated summary statistics. The only noticeable difference among graphs generated with different reproduction rates lies mainly in the size of the portion of the graph for which the summary statistic can be calculated (which increases with the reproduction rate). This difference is related to the time taken for the RE to complete during the forward simulation, which is compulsory for calculating summary statistics associated with that simulation. When a large number of forward simulations do not result in a completed RE, the corresponding value for the summary statistic is unavailable (depicted by the absence of color on these graphs). The weak influence of the reproduction rate on calculated summary statistics is

Table 1. Results of the Linear Regression Analyses Performed on Simulated Data Sets to Characterize the Variation of Each Summary Statistic.

		Overall	Adjusted R^2 Estimated from the Linear Regression against Each Parameter Taken Separately			
Scenario	Statistic	Adjusted R^2	*Ne*	*mf*	*T*1	t_R
IBD	$N_H tot$	0.749	0.254	0.271	–	–
	Global Φ_{ST}	0.955	0.339	0.531	–	–
	AMOVA Φ_{CT}	0.512	0.018	0.464	–	–
	$N_{ST} - G_{ST}$	0.378	0.098	0.121	–	–
	$m\Phi_{ST}dgeo$	0.953	0.407	0.445	–	–
	IBDSC	0.196	0.007	0.154	–	–
	ΔX_H	0.155	0.043	0.025	–	–
	$\Delta\pi$	0.113	0.026	0.030	–	–
	$\pi_R 1$	0.564	0.126	0.394	–	–
	$\pi_R 8$	0.579	0.139	0.399	–	–
	Global Φ_{ST}	0.949	0.376	0.443	0.019	–
	AMOVA Φ_{CT}	0.533	0.350	0.237	0.056	–
GF	$N_{ST} - G_{ST}$	0.470	0.019	0.041	0.036	–
	$m\Phi_{ST}dgeo$	0.945	0.449	0.374	0.019	–
	ΔX_H	0.714	0.201	0.008	0.367	0.018
	$\Delta\pi$	0.531	0.155	0.033	0.217	0.013
RE	$\pi_R 1$	0.318	0.155	0.006	0.076	0.005
	$\pi_R 8$	0.555	0.026	0.349	0.038	0.015
	Global Φ_{ST}	0.874	0.129	0.548	0.009	0.017
	$N_{ST} - G_{ST}$	0.283	0.012	0.073	0.017	<0.001

NOTE.—Values reported are adjusted coefficients of determination R^2 estimated from the linear regressions of each statistic against all the different simulation parameters (overall adjusted R^2) and against each simulation parameter taken separately. For these analyses, simulation parameters were treated as factors (see the text). All the R^2 values were significant (P-value < 0.05).

Table 2. List, Brief Description, and Reference of the Different Summary Statistics Analyzed in This Study.

Statistic	Description	Reference
$N_{\mathrm{H}}tot$	Total number of haplotypes in the data set.	
Global Φ_{ST} (for $K = 1$)	Global Φ_{ST} among populations/sampled cells (AMOVA Φ_{ST} for $K = 1$, i.e., when considering only one overall group of sampled cells): correlation among random sequences within populations/sampled cells, relative to that of random pairs of sequences drawn from the entire data set.	Excoffier et al. (1992)
$N_{\mathrm{ST}} - G_{\mathrm{ST}}$	G_{ST} among populations/sampled cells is a measure of populations/sampled cells differentiation based on allelic frequency differences only (i.e., without taking genetic distances between sequences into account); N_{ST} among populations/sampled cells, similar to the Φ_{ST} defined above, is a measure of populations/sampled cells differentiation taking genetic distances between sequences into account. The difference between these two statistics, $N_{\mathrm{ST}} - G_{\mathrm{ST}}$, is commonly used as a measure of overall phylogeographic signal.	Pons and Petit (1995, 1996)
IBDSC	Isolation by distance slope coefficient: slope coefficient of the linear regression between $\Phi_{\mathrm{ST}}/(1 - \Phi_{\mathrm{ST}})$ and $\ln(x)$, where Φ_{ST} is the pairwise Φ_{ST} between two populations/sampled cells and x the geographic distance between these two populations/sampled cells.	Rousset (1997)
$m\Phi_{\mathrm{ST}}dgeo$	Average ratio between Φ_{ST} estimators (Excoffier et al. 1992) and geographical distances between all pairwise populations/sampled cells.	See Mulcahy et al. (2006) for a similar statistic
Φ_{CT}, an AMOVA Φ-statistics[a]	Φ_{SC}, Φ_{ST}, and Φ_{CT} are Φ-statistics of the AMOVA, computed for estimating the population structure associated with user-defined groups of populations/cells. Φ_{SC} is a measure of the proportion of variation among populations/cells within groups, Φ_{ST} a measure of the proportion of variation among populations/cells, and Φ_{CT} a measure of the proportion of variation among groups.	Excoffier et al. (1992)
ΔX_{H}[a]	X_{H} is the ratio between the number of haplotypes in a user-defined group of populations/sampled cells and the total number of haplotypes ($N_{\mathrm{H}}tot$); ΔX_{H} is defined here as the difference between X_H estimated in sampled cell no. 1, located in the area of origin in the RE scenario, and the average X_H estimated in the other sampled cells not located in this area of origin.	
$\Delta\pi$[a]	π is the nucleotide diversity in a user-defined group of populations/cells (i.e., the average number of nucleotide differences per site between two sequences in this group), and $\Delta\pi$ is defined here as the difference between π estimated in sampled cell no. 1, located in the area of origin in the RE scenario, and the average π estimated in the other sampled cells not located in this area of origin.	Nei and Li (1979)
$\pi_{\mathrm{R}}1$[a], $\pi_{\mathrm{R}}8$[a]	π_{R} is the relative nucleotide diversity (i.e., the ratio between the nucleotide diversity within a given user-defined group of populations/cells and the nucleotide diversity within the virtual group formed by the populations/cells belonging to all other defined groups of populations). $\pi_{\mathrm{R}}1$ is here defined as the relative nucleotide diversity estimated in sampled cell no. 1, located in the area of origin in the RE scenario, and $\pi_{\mathrm{R}}8$ the relative nucleotide diversity estimated in sampled cell no. 8, the sampled cell furthest from this area of origin. Note that these statistics depend on the a priori definition of groups. For IBD and RE scenarios, 16 overlapping groups were defined (see text) so that these two statistics correspond to the ratios between nucleotide diversity estimated within cell no. 1 or 8 and nucleotide diversity estimated for the entire range (including cell no. 1 and 8). This decreases the range of possible values for the statistic but avoids having to deal with infinite values in extreme cases where nucleotide diversity calculated for the denominator equals zero (i.e., when there is only one distinct haplotype outside the group for which the relative nucleotide diversity is computed).	Mardulyn et al. (2009)

[a] Summary statistics based on the user-defined groups of sampled cells.

further shown by linear regression analyses that yielded significant but very small adjusted R^2 values (table 1, RE scenario).

All three summary statistics estimating the difference in genetic diversity between the area of origin and newly colonized regions (i.e., ΔX_{H}, $\Delta\pi$, and $\pi_{\mathrm{R}}1$) strongly vary with the different simulation parameter values, and appear particularly promising for investigating RE patterns. This is further confirmed by the linear regression analyses, in which the overall adjusted R^2 of ΔX_{H} and $\Delta\pi$ are notably higher under an RE than under an IBD scenario. Adjusted R^2 estimated for each parameter taken separately highlights parameters $T1$ (time at which the expansion occurred) and maximum cell effective size as those that best explains the variation of these statistics (table 1). These statistics should thus be useful to estimate these two parameters if one is wiling to assume an RE hypothesis.

ΔX_{H}, a statistic based on allele frequencies and measuring the difference in genetic diversity between the origin of an RE and the remaining of the range (see table 2), shows highest values in the case of relatively recent expansion events (short $T1$), before gene flow reduces the resulting diversity gradient between the origin of the expansion and the newly colonized portion of the range. In the case of recent expansions, ΔX_{H} is clearly lower when migration rates are too small or too high. If migration is too low, going backward in time, all sampled gene copies in a cell will have a high probability to coalesce before gene flow occurs and, most sampled cells will include a single distinct haplotype, leading to similar levels of diversity between newly colonized cells and those in the area of origin. If, on the contrary, migration is too high, gene flow will homogenize diversity between the area of origin and the newly colonized fraction of the range. $\Delta\pi$ is an analog of ΔX_H, but based on nucleotide diversity (Nei and Li 1979; table 2), therefore taking genetic distances among alleles into account. It displays a pattern very similar to that of ΔX_H, but is associated with a higher standard deviation (supplementary figs. S3–S7, Supplementary Material online), and is therefore characterized by a smaller adjusted R^2 in the linear regression analyses (table 1). For the parameter conditions investigated here, it thus appears redundant and less useful (since associated with more statistical noise). These results can be related to the recent character of the expansion: allele frequencies will mostly be affected by the recent RE, whereas variability associated with allelic distances does not provide interesting information.

$\pi_R 1$, the ratio between nucleotide diversity of population 1 (located in the area of origin) and that of the remaining of the distribution, displays a pattern of variation similar to the one observed for $\Delta\pi$: its standard deviation also increases with its mean value, and it appears also less useful than ΔX_H in this context. In fact, the initial purpose of the π_R statistic was to investigate diversity hotspots, and to infer whether they corresponded to ancient refuges or secondary contact zones (this application of the statistic is not tested here; see Mardulyn et al. 2009). It could therefore be useful and complementary to $\Delta\pi$ for cases not investigated here. Compared with $\pi_R 1$, $\pi_R 8$ acts as a negative control, because sampled cell no. 8 is located outside the origin of the RE. As expected, $\pi_R 8$ is thus almost constant across all surfaces. Finally it is interesting to note that global Φ_{ST} can also provide an interesting tool to explore RE, in particular to estimate the time of the expansion and the migration rate during the expansion. The overall adjusted R^2 estimated for $\pi_R 1$ under the RE scenario is clearly smaller than under IBD, which reflects the higher standard deviation associated with this statistic in the RE simulations.

A clear overall trend that can be inferred from the graphs showing patterns of variation in summary statistics is that the possibility to differentiate RE from IBD decreases with the age of the expansion. In other words, the genetic variation signal that reveals an RE is maximal just after the expansion, but fades out progressively afterwards. This confirms results obtained in a theoretical study of RE through mismatch distributions by Ray et al. (2003). Moreover, the speed with which the historical signal for RE decreases seems to depend on the extent of gene flow (migration rates) occurring among neighbor cells. For example, using this spatially explicit model to investigate the very recent RE (demonstrated by field observations) of a solitary bee across western Europe (Dellicour, Mardulyn, et al. 2014) highlighted the occurrence of high migration across the species range through the absence of a significant RE signal in DNA sequence data (a signal that would be expected if migration was low).

Comparison of Phylogeographic Scenarios

Thanks to the identical sampling design (sampled cells and number of sequences sampled per cell) in all computer simulations along three phylogeographic scenarios (fig. 1), it was possible to investigate whether sequence variation data (those generated by the simulations) can be used to identify the historical scenario that produced them. For this purpose, we have compared the sequence data simulated initially (i.e., the pseudo-observed data) along the IBD scenario with a series of data sets simulated along the GF or RE ($t_R = 2$) scenario (100 simulations per scenario and tested set of parameter conditions), through a set of selected computed summary statistics. Each comparison between pseudo-observed and simulated data is performed by combining computed summary statistics into a single chi-square statistic, as suggested in Dellicour, Mardulyn, et al. (2014). Results are displayed on figure 3. Additional figures presenting comparisons between IBD and RE simulations associated with $t_R = 1.1$, 4, 7, and 10 can be found in supplementary figure S9, Supplementary Material online.

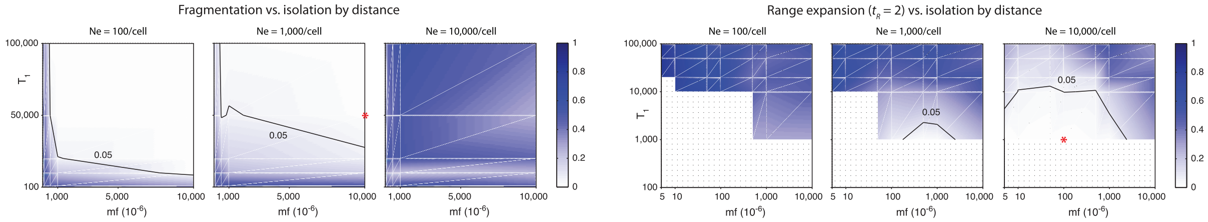


Fig. 3. Variation in the mean of *P*-values generated from pairwise scenario comparisons, 1) between IBD and GF scenarios, and 2) between IBD and RE scenarios, for all simulations conducted with reproduction rate $t_R = 2$. Comparisons were only performed between data sets simulated with the same maximal cell effective sizes *Ne* and forward migration rates *mf*. The lighter portion of the diagram identifies areas where the two scenarios are well differentiated (mean *P*-value close to 0) and in the case of RE simulations, white areas with dots correspond to sets of parameter values for which we could not compute the statistic because they did not allow the RE to reach all sampled cells. The portion of the space in which the two considered scenarios are significantly differentiated by a *P*-value ≤ 0.05 is delimited by a continuous black line. See supplementary figure S8, Supplementary Material online, for the corresponding representations of the standard variation associated with these *P*-values. Stars indicate the parameter combinations and associated averaged *P*-values selected for investigating the benefits of the spatially explicit model (see text and fig. 2).

Distinguishing a case of GF from IBD is notoriously difficult, because the effect of geographic isolation increasing population structure (through genetic drift) is compensated by migration decreasing population structure (by homogenizing genetic diversity). Therefore, similar levels of genetic structure can be reached with various combinations of divergence times and levels of migration. Indeed, summary statistics used to measure genetic differentiation among populations (e.g., global Φ_{ST} and $m\Phi_{ST}dgeo$) appear useful to distinguish between GF and IBD only in cases of old fragmentation events combined with relatively high migration (occurring within the continuous portions of the range) and small maximum cell effective size (fig. 2). Only extreme cases of population fragmentation should thus be easily identified. This is also reflected in figure 3, in which a significant difference between GF and IBD is found essentially when cell effective sizes are not too large ($Ne = 100$ and 1,000) and when combining a sufficiently high divergence time (i.e., time since the beginning of fragmentation $T1$) with a sufficiently high forward migration rate *mf* (allowing migration to occur within continuous portions of the species range, whereas no migration occurs among isolated regions in the GF scenario). For example, in a study focusing on historical range fragmentation of a cold-adapted leaf-beetle in North America (Dellicour, Fearnly, et al. 2014), GF could be favored over IBD because the isolation was sufficiently ancient (estimated at >50,000 years) that phylogeographic structure was strong (with maximum differentiation displayed by the mitochondrial locus, for which alleles found in each region can potentially form a monophyletic group). On the other hand, it is worth noting that we have explored a large range of *Ne***fm* values, including extreme values that are unlikely to be realistic, for example, corresponding to a number of effective migrant between two adjacent cells and per generation less than 0.001 or, at the other extreme, up to 100 (as most cells share eight neighbors, amounting to 800 effective migrants per cell and per generation). Thus, a portion of the space of parameter values investigated for which scenarios are indistinguishable likely corresponds to unrealistic cases that will never be encountered with real biological data.

Comparing RE with IBD, a significant difference is observed when cell effective size is large (*Ne* of at least 1,000) and under a specific range of forward migration rate *fm*, depending on the cells effective size (fig. 3). Indeed, even when considering a recent expansion, a high forward migration rate will logically cancel the difference in genetic diversity between the area of origin and the newly colonized regions. Conversely, a too small forward migration rate associated with a relatively small cell effective size (e.g., $Ne = 1{,}000$) can lead to the coalescence of all gene copies sampled within each newly colonized cell, before migration separates them. In this case, all sampled cells will contain a single distinct haplotype, and no difference in genetic diversity will be detected among them.

Overall, the results show that a spatially explicit model of coalescence, coupled with our method for comparing simulated and observed sequences, has the potential to detect past historical events such as GF or RE, but within a limited range of parameter values (at least with the set of summary statistics investigated here). Furthermore, a multilocus version of this method can easily be implemented (by simulating the evolution of several unlinked loci and performing independent comparisons with the real DNA sequence alignments). In this case, the *P*-values obtained for the different loci can be combined using the method of Fisher (1948). A unique *P*-value can thus be generated for each tested scenario (associated with a set of simulation parameter values). Two studies have implemented this approach already, along with the use of our spatially explicit model of coalescence, one investigating the colonization of a recent RE of a solitary bee in Europe (Dellicour, Mardulyn, et al. 2014), and another inferring the past and present connectivity across the range of a North American leaf beetle (Dellicour, Fearnly, et al. 2014). Although the use of an ABC framework to explore the space of parameter values for each scenario would be more desirable, coalescence simulations performed under such spatially explicit model are usually too time consuming. This is probably the reason why this kind of complex model has only been sporadically used in ABC analyses so far (e.g., Hamilton et al. 2005; Estoup et al. 2010; Ray and Excoffier 2010). The more rudimentary alternative presented here consists in manually choosing combinations of parameter values for the simulations, while trying to sample the space of parameter values being considered as homogeneously as possible. Although this approach will not be as effective as the automated random sampling of model parameter space implemented in an ABC analysis, it allows the investigation of more complex (in this case, geographically explicit) evolutionary scenarios.

Benefits of the Spatially Explicit Model

One obvious benefit of a spatially explicit model is the possibilities it offers to compare hypotheses over the evolution of a species range, through the integration of their detailed geographic description, and its possible combination with a niche modeling approach. In addition, under specific circumstances, failing to integrate the IBD nature of the data can lead to biased evolutionary inferences. To demonstrate this, we analyzed some of our simulated data generated under a spatially explicit model (referred hereafter as pseudo-observed data or POD), with simulations conducted following models more similar to a classic model. More specifically, we compared POD generated under a GF or RE scenario with the initial grid of 40 × 40 cells to data simulated under the same scenario but gradually decreasing the resolution of the grid (number of cells of 40 × 40, 20 × 20, 10 × 10, and 5 × 5, while keeping the overall maximum effective size for each region and effective migration [*Ne***mf*] constant). The POD were generated for different values of parameter $T1$ (corresponding to divergence time for GF or beginning of the expansion for RE), and we tested the ability of our comparison analyses to identify the correct $T1$. Resulting *P*-values are summarized in table 3. Under GF, the true $T1$ is only correctly identified as the most-likely scenario when using the highest grid resolution (40 × 40 cells), and its relative probability

Table 3. Impact of Grid Resolution on Historical Inference.

	Averaged *P*-Value (SD)			
T1 Number of Cells	40 x 40 (*mf* = 0.01)	20 x 20 (*mf* = 0.0025)	10 x 10 (*mf* = 0.0005)	5 x 5 (*mf* = 0.0001)
POD: GFs with *Ne* (for 40 x 40 cells) = 1,000 and *T1* = 50,000 generations ago				
5,000	0.003 (0.008)	0.002 (0.008)	0.000 (0.000)	0.000 (0.000)
10,000	0.009 (0.040)	0.009 (0.027)	0.000 (0.000)	0.000 (0.000)
50,000	0.505 (0.290)	0.256 (0.217)	0.087 (0.096)	0.008 (0.024)
100,000	0.408 (0.264)	0.301 (0.179)	0.151 (0.098)	0.144 (0.135)
POD: REs with *Ne* (for 40 x 40 cells) = 10,000; T1 = 1,000 generations ago and $t_R = 2$				
500	0.413 (0.312)	0.500 (0.263)	0.401 (0.297)	0.355 (0.296)
1,000	0.505 (0.290)	0.508 (0.282)	0.443 (0.296)	0.393 (0.282)
5,000	0.405 (0.283)	0.341 (0.254)	0.297 (0.253)	0.335 (0.282)
10,000	0.251 (0.193)	0.251 (0.202)	0.172 (0.159)	0.332 (0.274)

NOTE.—*P*-values obtained from the comparison between 100 sets of POD generated under a fully spatially explicit model (40 × 40 cells) and a specific time *T1* (i.e., divergence time for GF and beginning of expansion for RE; *T1* = 50,000 under GF, 1,000 under RE) and simulations of the same scenario but with various grid resolutions and times *T1*.

strongly decreases with decreasing grid resolution. Although the true *T1* is always associated with the highest *P*-value under RE, its relative probability also decreases with decreasing grid resolution, making the identification of the correct value more difficult.

Like for any theoretical study investigating the behavior of a method or summary statistics under varying conditions, we had to limit our exploration to a restricted set of conditions. We chose those that appeared most relevant for phylogeographic studies, and spanned a range of conditions that seemed realistic in natural conditions. However, the effect of some important parameters associated with the sequences themselves was not tested, for example, the sequence length, mutation rate (or number of mutations), and number of loci. For our study, we have obviously set the number of mutations and sequence length to values that allowed a sufficient level of polymorphism for measuring genetic diversity and population structure on the simulated data sets. Reducing these too much would result in levels of polymorphism too low to be useful, whereas increasing the mutation rate without increasing the sequence length would lead to saturated historical signal. In practice, most biologists choose their DNA markers by selecting those providing a good level of polymorphism for their studied organism/range. On the other hand, increasing the number of loci should definitely improve the inferences performed here on a single locus. Another important factor that has been poorly investigated here is sampling design. Indeed, a single sampling design, relatively symmetric, has been conducted across scenarios (fig. 1). This was necessary for our purpose of evaluating the potential of the method to distinguish among phylogeographic scenarios. Although beyond the scope of this study, it would probably be interesting in future studies to test the impact of various sampling designs on historical inference.

Conclusions

Our study illustrates the possibility of discriminating among phylogeographic hypotheses using a spatially explicit model of coalescence, by simulating DNA sequence data under three typical phylogeographic scenarios and comparing data sets through classic summary statistics. One strong observation made in our study is the relatively narrow range of parameter values under which each summary statistic can be useful, especially in the context of a geographical fragmentation, either for inferring parameter estimates within a scenario assumed a priori or for comparing alternative phylogeographic hypotheses. This underlines the importance of the choice of appropriate summary statistics, which will likely strongly vary among studies. Selection of appropriate statistics is already discussed in several articles, mainly in the context of ABC methods (e.g., Joyce and Marjoram 2008; Nunes and Balding 2010; Fearnhead and Prangle 2012; Blum et al. 2013). Prior simulations along alternative spatially explicit scenarios can be performed, as was done here, to identify the most promising summary statistics for the purpose of discriminating among corresponding historical hypotheses. One option is to perform a PCA (principal component analysis) on summary statistics calculated from simulated data, to select those associated with interscenarios variation, that is, those that appear most efficient to discriminate among the proposed hypotheses (see, e.g., Veeramah et al. 2012).

Although the summary statistics assessed here are relatively common indices used in population genetic studies, most are relatively crude indices for measuring spatial distribution of genetic variation. Indeed, they simply measure overall genetic diversity within populations or population structure given a predefined set of geographical groups, each taken as a panmictic population. However, the use of a spatially explicit model of coalescence opens the possibility to define new summary statistics describing the distribution of genetic variation more accurately, that is, taking patterns of isolation by distance within regions into account. Such summary statistics may have the potential to use better the information on spatial distribution of genetic variation, and probably deserve to be developed and investigated in future studies.

Model Description

Overview

The geographic structure of the studied range at any point in time is defined by a two-dimensional grid in which each cell is considered a panmictic population. This is easily done by overlaying a grid of appropriate size and resolution on a map showing the current or past distribution of the range. Cells accessible to individuals (i.e., those included in the range) are thus identified on each grid, and each sampled sequence is attributed to one cell on the grid that displays the current range. This geographic information is easily transferred to the input file, along with a set of demographic parameters (e.g., maximal cell effective sizes and migration rates between adjacent cells). The simulation begins at the current time $t = 0$ and, going backward in time, ends when all gene copies (DNA sequences) for the considered locus have coalesced. At each generation g, a given gene copy has the opportunity to move to adjacent cells and to coalesce with another gene copy located in the same cell. At the end of the simulation, a genealogy is built based on the recorded coalescence events. The total number of mutations added on the genealogy can be specified a priori or is determined by a stochastic mutation process for which a mutation rate is defined. Similarly to the method developed by Currat et al. (2004; see also Ray et al. 2010), a forward presimulation is performed to estimate parameters (cells size and migration rates) for the coalescence simulation (details below). A representation of the general workflow is available in figure 4.

Coalescence Simulation

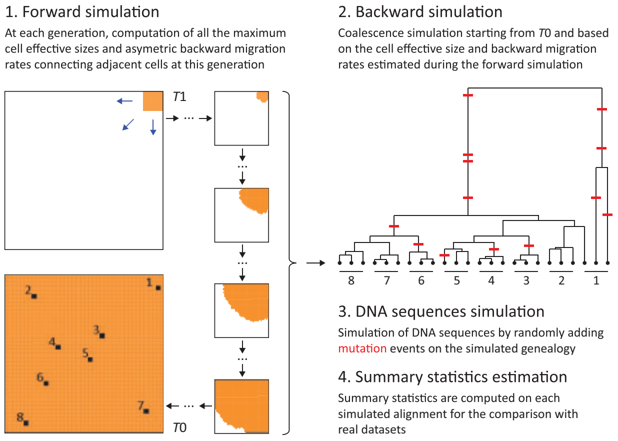
A simulation begins at $t = 0$ (i.e., sampling time) and is finished when all gene copies of a given locus have coalesced. For multiple loci, the simulations are independent from each other (assuming maximum recombination among loci). Going backward in time, at each generation g, a given gene copy has the opportunity: 1) to coalesce with another gene copy located in the same cell and 2) to migrate to one of the adjacent cells on the grid. The probability of coalescence of a given gene copy located in cell j is noted $P_c(j,g)$:

$$P_c(j, g) = \frac{n_j(g) - 1}{N_j(g)}$$

with $N_j(g)$ the effective size (total number of gene copies), and $n_j(g)$ the number of sampled gene copies, in cell j at $t = g$. The probability of migration for a given gene copy in cell j is determined by "backward" migration rates with adjacent cells and is noted $P_m(j,g)$:

$$P_m(j, g) = \sum_{j'=1}^{K} m_{jj'}(g)$$

with $m_{jj'}(g)$, the backward migration rate from cell j to cell j' at generation g, that can be retrieved from the backward migration matrix (see below); and K, the total number of cells on the grid. The simulation continues until a single gene copy remains. At the end of the simulation, the program builds a genealogy based on the recorded events of coalescence. Mutations are then added to the genealogy according to a Jukes–Cantor model of DNA substitution (Jukes and Cantor 1969).


1. Forward simulation

At each generation, computation of all the maximum cell effective sizes and asymetric backward migration rates connecting adjacent cells at this generation

2. Backward simulation

Coalescence simulation starting from *T0* and based on the cell effective size and backward migration rates estimated during the forward simulation

3. DNA sequences simulation

Simulation of DNA sequences by randomly adding mutation events on the simulated genealogy

4. Summary statistics estimation

Summary statistics are computed on each simulated alignment for the comparison with real datasets

Fig. 4. Workflow for one simulation (RE scenario) using the spatially explicit model of coalescence, leading to specific values for computed summary statistics. In this fictive example, three gene copies have been "sampled" in each of the eight sampled populations (represented by black squares on the grid).

Two Coalescence Simulation Modes

A coalescence simulation is conducted in a "generation-by-generation" mode (i.e., one generation at a time), unless two conditions are met: 1) if the effective population size of each cell on the grid has reached its maximal value (which is determined by the forward presimulation) and 2) if the estimated time to the next coalescence or migration event is larger than one generation. In that case, the algorithm switches to a faster mode, a "time to the next event" mode, in which the time to the next migration or coalescence event is simulated. Both modes are described below.

Generation-by-Generation Mode.

The implementation of this mode is justified by the fact that the effective population size of a cell, as determined by the preforward simulation, varies at each generation until it reaches its maximum value (defined by the user), or because the estimated time to the next migration or coalescent event is not more than a single generation. At each generation and in each cell, the algorithm begins by determining whether coalescence(s) occur(s) as follows: 1) it computes the probability that there is at least one coalescence event in the considered cell j,

$$P_{\text{coalescence},j} = 1 - P_{\text{noCoalescence},j} = 1 - \prod_{i=1}^{n_j - 1} \left(1 - \frac{i}{Ne_j}\right)$$

with n_j, the number of gene copies in cell j; and Ne_j, the effective size (haploid case) of cell j; 2) it generates a random number between 0 and 1. If this number is smaller than the computed probability, there is at least one event of coalescence in this cell and the algorithm goes to step (3). 3) This step determines how many events of coalescence will occur. There are three distinct cases:

(3.1) if $n_j = 2$: Coalescence of these two gene copies.
(3.2) if $n_j > 2$ and if $Ne_j \geq 10{*}n_j$: Coalescence of only two random gene copies in cell *j*.
(3.3) if $n_j > 2$ and if $Ne_j < (10{*}n_j)$: Possibility of more than one event of coalescence in cell *j*.

In case (3.3), the number of coalescence events is determined as follows: n_j gene copies are drawn with replacement within the Ne_j gene copies present in cell *j* at this generation. The number of coalescence events is determined by the number of times that a gene copy is drawn more than once.

After the coalescence step, the algorithm determines whether migration(s) occur(s) by considering that the probability of migration for a given gene copy equals the "backward" migration rate from the sink cell to the source cell. This probability is calculated as described above.

Time-to-the-Next-Event Mode.

In this mode, the time to the next coalescence or migration event is simulated using an extended version of the structured coalescent model of Hudson (1991), initially developed for two populations of effective size *Ne* that exchange gene copies at rate *m*. In the two equal-size populations model of Hudson, the time until the next event (coalescence or migration) is distributed according to an exponential distribution with mean:

$$\frac{1}{\left(\frac{1}{Ne}\binom{n_1}{2}+\frac{1}{Ne}\binom{n_2}{2}+(n_1+n_2)m\right)},$$

where n_j is the number of remaining gene copies in population *j*. When an event does occur, the probability of a coalescence event among the n_j gene copies remaining in population *j* is

$$\frac{\frac{1}{Ne}\binom{n_j}{2}}{\left(\frac{1}{Ne}\binom{n_1}{2}+\frac{1}{Ne}\binom{n_2}{2}+(n_1+n_2)m\right)},$$

whereas the probability of a migration event for a gene copy from population *j* is

$$\frac{n_j m}{\left(\frac{1}{Ne}\binom{n_1}{2}+\frac{1}{Ne}\binom{n_1}{2}+(n_1+n_2)m\right)}.$$

Extending this model to *K* populations/cells of different effective sizes and connected with different backward migration rates, the time until the next coalescence or migration event is still distributed according to a geometric distribution, this time with mean:

$$\frac{1}{\left(\sum_{j=1}^{K}\left(\frac{1}{Ne_j}\binom{n_j}{2}\right)\right)+\sum_{j_1}^{K}\sum_{j_2}^{K}\left(n_{j_1}m_{j_1j_2}\right)},$$

where $m_{jj'}$ is the backward migration rate from cell *j* to cell *j'*. When an event does occur, the probability of a coalescence event among the n_{k} gene copies remaining in cell k is

$$\frac{\frac{1}{Ne_{\mathrm{k}}}\binom{n_{\mathrm{k}}}{2}}{\left(\sum_{j=1}^{K}\left(\frac{1}{Ne_j}\binom{n_j}{2}\right)\right)+\sum_{j_1}^{K}\sum_{j_2}^{K}\left(n_{j_1}m_{j_1j_2}\right)}.$$

The probability of a migration event for a gene copy from cell k_1 to cell k_2 is

$$\frac{n_{\mathrm{k}_1}m_{\mathrm{k}_1\mathrm{k}_2}}{\left(\sum_{j=1}^{K}\left(\frac{1}{Ne_j}\binom{n_j}{2}\right)\right)+\sum_{j_1}^{K}\sum_{j_2}^{K}\left(n_{j_1}m_{j_1j_2}\right)}.$$

The Need for a Preliminary Forward Simulation

Each coalescence simulation uses "backward" migration rates, that is, the probability that a gene copy has migrated from another specific cell at a given generation. Natural dispersal processes, however, define "forward" migration rates, that is, the probability that a gene copy migrates to a given location. In some cases, forward and backward migration rates are not identical. For example, during an RE, the probability that gene copies from a cell *A* located at the margin of the colonization front migrate to a newly colonized (i.e., previously empty) cell *B* could be set to a forward migration rate of 0.001. If a few gene copies do colonize cell *B* from cell *A* at a given generation, the backward migration rate from *B* to *A* at the same time will be equal to 1. Because backward migration rates depend both on forward migration rates and grid cell effective sizes, and that the latter can change over the course of the simulation (e.g., simulating a geographic expansion, a newly colonized cell will increase its effective size until it reaches the maximum value defined in the model), backward migration matrices are generated for each generation prior to the coalescence simulation by performing a forward stochastic simulation that records the changes in cell effective sizes. For this purpose, the model requires that users specify the actual effective sizes on the most ancestral grid, the maximum effective sizes on all grids (and the time interval during which each grid should be implemented), an intrinsic reproduction rate (t_{R}, indicating the speed at which a population grows thanks to reproduction), and two forward migration rates (characterizing the dispersal capabilities of an organism): A short-distance migration rate *mf*1, between adjacent cells (including cells connected by a corner), and a long-distance migration rate *mf*2, between cells separated by two cells on the grid. Hence, the probability that a gene copy migrates to another cell is $mf_{\mathrm{total}} = 8{*}mf1 + 16{*}mf2$, if located in a cell at least two cells apart from any border.

The model assumes that the dispersal ability and reproduction rate of the studied organism are constant across the entire grid. Only the carrying capacity of each cell is allowed to change by defining the maximum effective size of each cell separately. Note that the presence of a barrier to migration (e.g., a mountain range or a river) can be modeled by assigning a small (or null) maximal effective size to one or more cells, thereby reducing the probability of migration through them. The preliminary forward simulation is performed as follows: starting with the matrix of actual effective sizes for the most ancestral grid, at a specified time, the simulation proceeds forward until $t = 0$. During the time interval dedicated to one grid, the simulation proceeds one generation at a time until each cell has reached its maximum effective size. During this phase, at each generation: 1) all actual cell effective sizes are recorded; 2) migration events among cells are simulated, using the two predefined forward migration rates and the cell effective sizes of the preceding generation; these migration events are recorded and used to define the backward migration rates at this generation; 3) the effective size of each cell increases as individuals reproduce and new migrants are brought in; and 4) if the effective size of a cell has exceeded its maximum value, it is reduced to this maximum.

When two cells have reached their maximal effective size, the two backward migration rates connecting them are directly computed according to the following deterministic formula (instead of being based on simulating an exchange of migrants):

$$m_{jj'}(g) = \frac{Ne_{j'}(g-1)\cdot M_{j'j}}{\sum\limits_{p=1}^{P}\left(Ne_{p}(g-1)\cdot M_{pj}\right)}$$

with $m_{jj'}(g)$, the backward migration rate from cell j to cell j' at generation g; $Ne_j(g-1)$, the effective size (haploid case) of cell j at generation $g - 1$; $M_{jj'}$, the forward migration rate from cell j to cell j'; and P, the total number of cells on the grid. If $j = j'$, then

$$M_{jj'} = 1 - \sum_{p\neq j} M_{jp}$$

When all effective sizes on the grid have reached their maximum, the forward simulation stops until the next change of grid or until $t = 0$. From then, all backward migration rates are fixed values computed with the deterministic formula above. The use of a deterministic formula to estimate backward migration rates when cell effective sizes are constant allows a significant increase of the forward simulation speed. Cell effective sizes and backward migration rates are recorded for the coalescence (backward) simulations. Because of the stochastic character of the preliminary forward simulation (migration events occur according to probabilities defined by forward migration rates), the program can renew the forward simulation after any number of backward simulations (this number being defined by the user).

Note that if the backward simulation reaches the most ancestral generation of the forward simulation, the model simply uses the most ancestral cell effective sizes and backward migration rates are then estimated according to the following deterministic formula:

$$m_{\mathrm{a},jj'} = \frac{Ne_{\mathrm{a},j'}\cdot M_{j'j}}{\sum\limits_{p=1}^{P}(Ne_{\mathrm{a,p}}\cdot M_{\mathrm{p}j})}$$

with $m_{a,jj'}$, the backward migration rate from cell j to cell j' estimated during the most ancestral generation of the forward simulation; and $Ne_{a,j}$, the effective size of cell j (haploid case) estimated during the most ancestral generation of the forward simulation.

Model Implementation

A simulation program implementing this model, PHYLOGEOSIM 1.0 (for "phylogeographic simulator"; java executable compatible with any operating system equipped with a Java Virtual Machine) was developed for the purpose of this study. It is available from ebe.ulb.ac.be/ebe/Software.html (last accessed October 13, 2014) (with a manual and example input files), and can be used for many applications that involve simulating DNA sequences in a geographic context. It can also compute several summary statistics describing DNA sequence variation directly from the simulated data sets. These summary statistics can be used either to conduct theoretical studies evaluating the impact of different evolutionary scenarios on patterns of genetic variation (as was the purpose here) or to compare observed and simulated DNA sequence data to evaluate different historical hypotheses. A description of all summary statistics computed by the software and used in this study is available in table 2.

Materials and Methods

Simulations of Typical Phylogeographic Scenarios

Spatially explicit simulations of past demographic events were performed to investigate the behavior of common summary statistics: 1) While varying parameter values of each historical scenario (i.e., within the boundaries of this scenario) and 2) among historical scenarios, to attempt to identify statistics that could be used to differentiate them a posteriori. Furthermore, for comparing scenarios, that is, for estimating the relative probability that each of them has generated the observed data, we evaluated the possibility of combining a set of selected summary statistics into a single chi-square statistic, used to compare the observed (or pseudo-observed) and simulated data sets.

We designed three spatially explicit models representing each a typical historical scenario commonly inferred in phylogeographic studies (fig. 1): 1) IBD within a large continuous range (IBD scenario), 2) GF scenario, and 3) RE scenario. Three successive spatial grids, each defined by a matrix of maximal cell effective sizes, represent each scenario. The sampling design was identical in all three cases, so that the resulting pattern of genetic variation could be easily compared among scenarios. The most ancestral grid was characterized in all cases by a strong reduction of the range. Failing to do so resulted in extremely large TMRCA (i.e., time to the most recent common ancestor) values, because the last remaining gene copies at the end of the backward simulation took an excessively long time to coalesce. In that case, the generated genealogies were dominated by long ancestral branches on which most of the mutations occurred, regardless of the implemented scenario, and the history depicted on the two most recent grids did not influence the spatial distribution of genetic diversity found at $t = 0$. Also, such an ancestral restricted range is found in many natural systems, as many species experienced one or more bottlenecks and/or a decrease in their geographic range (e.g., during the last glaciation for many temperate climate species) during their more or less recent history.

The following parameters were set for all simulations across all three scenarios: DNA sequences of 800 bp, 25 mutation events randomly added on the branches of the simulated genealogy, and ten sampled gene copies in each of the eight sampled cells. In addition, long distance forward migration rate (*mf*2, see Model Description section) was systematically set to zero in order to decrease the number of simulation parameters to study/explore. As a consequence, only migration between adjacent cells (*mf*1) was allowed. As shown on figure 1, the eight sampled cells were distributed on a square grid of 1,600 cells such that four sampled cells were more peripheral and the four others were located in the center of the range. Other parameters varied among simulations in order to test the effect of their variation on the summary statistics investigated. Two parameters were varied in all scenarios: the cell effective size *Ne* and the forward migration rate *mf*. In addition, specific parameters of the GF and of the RE simulations varied across simulations: the time of the fragmentation event (GF) and the rate and time of the expansion event (RE). The tested values for all simulation parameters are summarized in table 4. For each combination of explored parameter values, a total of a 100 backward simulations were performed and one preliminary forward simulation was run for every ten backward simulations. For comparing IBD with GF (see step 2 below), four distinct groups of sampled cells were defined: Sampled cells 1 and 3, 2 and 4, 5 and 7, and 6 and 8; and for comparing IBD with RE, each sampled cell (no. 1–8) was considered a separate group. In this last case, we compared sampled cell no. 1, located inside the area of origin of the RE, with all other sampled cells. When comparing IBD with RE, eight additional groups were also defined: the four groups defined for the comparison between IBD and GF, two groups, respectively, gathering the "Western" (no. 2, 4, 6, 8) and "Eastern" sampled cells (no. 1, 3, 5, 7), and two groups gathering the central population (no. 3, 4, 5, 6) and external sampled cells (no. 1, 2, 7, 8). These additional groups were involved in estimating relative nucleotide diversity (see table 2). Note that the a priori definition of groups of sampled cells does not affect the simulation itself, only the calculation of some related summary statistics on simulated data.

Evaluating Summary Statistics

Table 2 lists the summary statistics investigated. Except for $m\Phi_{ST}dgeo$, that we define here as the average ratio between Φ_{ST} estimators (Excoffier et al. 1992) and geographical distances calculated between all population (i.e., sampled cell) pairs, all other tested summary statistics were already used in previous population genetic and/or phylogeographic studies. N_Htot, Global Φ_{ST}, $N_{ST} - G_{ST}$, *IBDSC*, or $m\Phi_{ST}dgeo$ are overall measures of the genetic variability across the range of a species. A second group of statistics are based on the definition, by the user, of groups of sampled cells: The analysis of molecular variance (AMOVA) Φ statistics, or ΔX_H, $\Delta\pi$, and π_R1, three measures of genetic diversity difference between sampled cell no. 1, located in the area of origin in the RE scenario, and the other cells located elsewhere. We also estimated π_R8,

Table 4. Simulation Parameters Explored for Three Evolutionary Scenarios.

Scenario	Fixed Parameter	Tested Values for Fixed Parameter	*X* Axis	Tested Values on *X* Axis	*Y* Axis	Tested Values on *Y* axis	Tested Summary Statistics
IBD	—	—	*Ne*	10^2, 10^3, 10^4, 10^5	*mf*	5×10^{-6}, 10^{-5}, 5×10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 10^{-2}	N_Htot, global Φ_{ST}, AMOVA Φ_{CT}, $N_{ST} - G_{ST}$, $m\Phi_{ST}dgeo$, *IBDSC*, ΔX_H, $\Delta\pi$, π_R1, π_R8
GF	*Ne*	10^2, 10^3, 10^4	*T*1	10^4, 2×10^4, 5×10^4, 10^5	*mf*	5×10^{-6}, 10^{-5}, 5×10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 10^{-2}	N_Htot, global Φ_{ST}, AMOVA Φ_{CT}, $N_{ST} - G_{ST}$, $m\Phi_{ST}dgeo$
RE	*Ne* t_R	10^2, 10^3, 10^4 1.1, 2, 4, 7, 10	*T*1	10^2, 10^3, 10^4, 2×10^4, 5×10^4, 10^5	*mf*	5×10^{-6}, 10^{-5}, 5×10^{-5}, 10^{-4}, 5×10^{-4}, 10^{-3}, 10^{-2}	N_Htot, global Φ_{ST}, $N_{ST} - G_{ST}$, ΔX_H, $\Delta\pi$, π_R1, π_R8

NOTE.—Parameters are either fixed, or vary along the *x* axis or *y* axis of the graphs displaying summary statistics variation (fig. 2) generated under the three scenarios. *Ne* refers to maximal cell effective size, *mf* to forward migration rate, and *T*1 to the time (in number of generations) when the fragmentation or range expansion begins. See text and table 2 for further details on the different tested summary statistics.

the relative nucleotide diversity estimated in sampled cell no. 8, the sampled cell geographically most distant from this area of origin. Although statistics like AMOVA Φ statistics, $N_{ST} - G_{ST}$, *IBDSC*, and $m\Phi_{T}dgeo$ should allow to measure and compare the levels of genetic differentiation among sampled cells (belonging or not to the same group in the case of GF scenario), statistics like ΔX_H, $\Delta\pi$, and π_R1 should allow to quantify the loss of genetic diversity due to potential founder effects in newly colonized cells for the RE scenario.

For each summary statistic, a graph displaying its variation for different model parameter values was generated for each scenario. In the case of GF and RE scenarios, a graph was generated per tested value of maximal cell effective size *Ne*, and for the RE scenario alone, a different graph was also generated for different values of the reproduction rate t_R. The tested sets of model parameter values are summarized in table 4. All summary statistic graphs were generated in MATLAB (The MathWorks, Inc) with the "surf" function. Note that we show variation of ln(1+*IBDSC*), because such a logarithmic transformation can facilitate the interpretation of the variation pattern. In addition to these graphs, we also conducted linear regression analyses and calculated the related coefficients of determination R^2 as an estimate of the proportion of summary statistics variation explained by each parameter. For all summary statistics studied within each scenario, we performed two kinds of linear regressions: 1) an overall linear regression of each statistic against all different simulation parameters and 2) linear regressions of each statistic against each simulation parameter taken separately. In all regression analyses, the variables of the linear model, although continuous, were categorized to avoid the assumption of linear relationship between the statistics and simulation parameters. However, both analyses were also carried out using continuous variables for each model, and similar results were obtained (not shown). We then report, per statistic, an overall R^2, as well as an R^2 value associated with each simulation parameter (*Ne* and *mf* for all scenarios, *T*1 for GF and RE, and also t_R for RE). All linear regressions were performed with the "lm" function available in R (R Development Core Team 2013).

Comparing Phylogeographic Scenarios

To compare real and simulated data, for the purpose of discriminating among different hypothesized historical scenarios, all summary statistics computed for one data set are combined into a single chi-square statistic as follows:

$$\chi^2 = \sum_j \left[\frac{(St_j - m_j)^2}{\sigma_j^2} \right],$$

where St_j is the value of the *j*th summary statistic and m_j and σ_j are, respectively, the average and the standard deviation of the *j*th statistic over the *n* simulations generated under the same scenario (e.g., GF) and set of parameter values. We can thus generate a distribution of *n* chi-square statistics for a given scenario and a given set of parameter values, from simulated data. This distribution can be compared with the chi-square statistic estimated on an observed, or in the case of this study, a POD set, yielding a nonparametric test that returns a *P*-value corresponding to the proportion of simulated values greater than or equal to the observed value.

We used this "chi-square" method to perform several pairwise scenario comparisons, each time between the IBD, considered here to be our null model (absence of historical event), and the GF scenario or the RE scenario. Each pairwise comparison is performed with identical *Ne* and *mf* values for the two scenarios involved. Each comparison with the GF scenario is made for several values of divergence time and with the RE scenario for several values of the time at which the expansion begins and of the reproduction rate. For each set of selected parameter values within a scenario, we performed 100 simulations. For every comparison, each of the generated 100 simulated data sets under one scenario (thus considered as the POD sets) was compared with the distribution of chi-square statistics generated with the other scenario. As each single comparison is associated with a *P*-value, each set of simulation parameters was associated with a mean *P*-value and its standard deviation, calculated from 100 comparisons. Based on a prior exploration of different summary statistics, we selected four promising summary statistics for each pairwise comparison of two scenarios: Φ_{ST}, AMOVA Φ_{CT}, $N_{ST} - G_{ST}$ and $m\Phi_{ST}dgeo$ for comparing the GF and IBD scenarios; and ΔX_H, $\Delta\pi$, π_R1 and π_R8 for comparing the RE and IBD scenarios.

Benefits of the Spatially Explicit Model: Impact of Grid Resolution

In order to explore the impact of the model grid resolution (i.e., the number of cells defining the grid) on historical inference, we analyzed POD generated using a GF or RE model (100 sets of POD per scenario), by comparing them with data simulated while gradually shifting from a fully spatially explicit model toward a more classic model (i.e., simply decreasing the total number of cells on the grid). This comparison was conducted with various *T*1 values (divergence time for GF, time of beginning of expansion for RE; see table 3 for tested values). For this test, we selected simulation parameter values that allowed a clear differentiation between GF or RE and IBD scenarios. For simulations involving a lower number of cells (20 × 20, 10 × 10, and 5 × 5), maximal cell effective sizes were increased to maintain the same overall maximum effective sizes per region (as well as across the entire range) and forward migration rates were proportionally decreased in order to maintain similar *Ne***mf* values (see table 3).

Supplementary Material

Supplementary figures S1–S9 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Caroline Hardy and François Mayer for help in testing the simulation program. Computational resources have been provided by the High

Performance Computing Centre co-funded by ULB and VUB (HPC cluster "Hydra") and by the *Consortium des Équipements de Calcul Intensif* (CÉCI, HPC cluster "HMEM") funded by the Fonds de la Recherche Scientifique de Belgique (FRS-FNRS) under convention 2.5020.11. They also thank Raphaël Leplae from the ULB/VUB Computing Centre, Damien François and David Collignon from the CÉCI for their precious help and advices. This research project was funded by the Belgian *Fonds pour la Recherche Scientifique* (FNRS) and the *Fonds pour la Recherche dans l'Industrie et l'Agriculture* (FRIA).

References

Arenas M. 2012. Simulation of molecular data under diverse evolutionary scenarios. *PLoS Comput Biol.* 8(5):e1002405.

Beaumont MA, Zhang WY, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol.* 19:2609–2625.

Blum MGB, Nunes MA, Prangle D, Sisson SA. 2013. A comparative review of dimension reduction methods in approximate Bayesian computation. *Stat Sci.* 28:189–208.

Currat M, Excoffier L. 2005. The effect of the Neolithic expansion on European molecular diversity. *Proc R Soc Lond B Biol Sci.* 272: 679–688.

Currat M, Excoffier L. 2011. Strong reproductive isolation between humans and Neanderthals inferred from observed patterns of introgression. *Proc Natl Acad Sci U S A.* 108:15129–15134.

Currat M, Ray N, Excoffier L. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. *Mol Ecol Notes.* 4:139–142.

Dellicour S, Fearnly S, Lombal A, Heidl S, Dahlhoff E, Rank NE, Mardulyn P. Forthcoming 2014. Inferring the past and present connectivity across the range of a North American leaf beetle: combining ecological-niche modeling and a geographically explicit model of coalescence. *Evolution* 68:2371–2385.

Dellicour S, Mardulyn P, Hardy OJ, Hardy C, Roberts SPM, Vereecken NJ. 2014. Inferring the mode of colonization of the rapid range expansion of a solitary bee from multilocus DNA sequence variation. *J Evol Biol.* 27:116–132.

Estoup A, Baird SJE, Ray N, Currat M, Cornuet JM, Santos F, Beaumont MA, Excoffier L. 2010. Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad Bufo marinus. *Mol Ecol Res.* 10:886–901.

Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. *Annu Rev Ecol Evol Syst.* 40:481–501.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479–491.

Fearnhead P, Prangle D. 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J R Stat Soc B.* 74:419–474.

Fisher RA. 1948. Questions and answers #14. *Am Stat.* 2(5):30–31.

Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, Excoffier L. 2005. Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* 170:409–417.

Hey J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol.* 27:905–920.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.

Hoban S, Bertorelle G, Gaggiotti OE. 2012. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet.* 13: 110–122.

Hudson R. 1991. Gene genealogies and the coalescence process. *Oxford surreys in evolutionary biology* 7:1–44.

Joyce P, Marjoram P. 2008. Approximately sufficient statistics and Bayesian computation. *Stat Appl Genet Mol Biol.* 7, article 26.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Mardulyn P, Mikhailov YE, Pasteels JM. 2009. Testing Phylogeographic hypotheses in a Euro-Siberian cold-adapted leaf beetle with coalescent simulations. *Evolution* 63:2717–2729.

Meirmans PG. 2012. The trouble with isolation by distance. *Mol Ecol.* 21: 2839–2846.

Mulcahy DG, Spaulding AW, Mendelson JR, Brodie ED. 2006. Phylogeography of the flat-tailed horned lizard (*Phrynosoma mcallii*) and systematics of the *P. mcallii–platyrhinos* mtDNA complex. *Mol Ecol.* 15:1807–1826.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76: 5269–5273.

Nunes MA, Balding DJ. 2010. On optimal selection of summary statistics for approximate Bayesian computation. *Stat Appl Genet Mol Biol.* 9, article 34.

Pons O, Petit RJ. 1995. Estimation, variance and optimal sampling of genetic diversity. I. Haploid locus. *Theor Appl Genet.* 91: 122–130.

Pons O, Petit RJ. 1996. Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* 144:1237–1245.

R Development Core Team. 2013. R: a language and environment for statistical computing.

Ray N, Currat M, Excoffier L. 2003. Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol.* 20:76–86.

Ray N, Excoffier L. 2010. A first step towards inferring levels of long-distance dispersal during past expansions. *Mol Ecol Res.* 10: 902–914.

Ray N, Currat M, Foll M, Excoffier L. 2010. SPLATCHE2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics* 26:2993–2994.

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

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145: 505–518.

Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, Destro-Bisol G, Soodyall H, Louie L, Hammer MF. 2012. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol Biol Evol.* 29:617–630.

Wright S. 1969. Evolution and the genetics of populations: the theory of gene frequencies. Chicago (IL): The University of Chicago Press.