Title
Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees

Journal
Genetica, 124(2-3)

ISSN
0016-6707

Authors
Kuo, Chih-Horng
Avise, John C

Publication Date
2005-07-01

DOI
10.1007/s10709-005-2095-y

Genetica (2005) 124: 179–186
DOI 10.1007/s10709-005-2095-y

Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees

Chih-Horng Kuo & John C. Avise
Department of Genetics, University of Georgia, Athens, GA 30602, USA (Phone: +1-706-542-8133; Fax: +1-706-542-3910; E-mail: chkuo@uga.edu)

Received 9 October 2004 Accepted 18 January 2005

Key words: coalescent processes, gene flow, geographic barriers, intraspecific phylogeography, population size

Abstract

Computer simulations were used to investigate population conditions under which phylogeographic breaks in gene genealogies can be interpreted with confidence to infer the existence and location of historical barriers to gene flow in continuously distributed, low-dispersal species. We generated collections of haplotypic gene trees under a variety of demographic scenarios and analyzed them with regard to salient genealogical breaks in their spatial patterns. In the first part of the analysis, we estimated the frequency in which the spatial location of the deepest phylogeographic break between successive pairs of populations along a linear habitat coincided with a spatial physical barrier to dispersal. Results confirm previous reports that individual gene trees can show 'haphazard' phylogeographic discontinuities even in the absence of historical barriers to gene flow. In the second part of the analysis, we assessed the probability that pairs of gene genealogies from a set of population samples agree upon the location of a geographical barrier. Our findings extend earlier reports by demonstrating that spatially concordant phylogeographic breaks across independent neutral loci normally emerge only in the presence of longstanding historical barriers to gene flow. Genealogical concordance across multiple loci thus becomes a deciding criterion by which to distinguish between stochastic and deterministic causation in accounting for phylogeographic discontinuities in continuously distributed species.

Introduction

One of the most common and striking empirical patterns in phylogeography is the occurrence of sharp breaks in mitochondrial (mt) DNA genealogy that often distinguish regional sets of populations across a species' range (Avise, 2000). These geographic arrays of highly differentiated matrilines are sometimes referred to as 'intraspecific phylogroups' (Avise & Walker, 1999) or as provisional 'evolutionarily significant units' (Ryder, 1986; Moritz, 1994; Bernatchez, 1995). If a succession of populations along a linear transect were to be genetically sampled across the geographic range of such a species, a sharp spike in the magnitude of mtDNA sequence divergence would characterize adjacent pairs of demes on alternate sides of each phylogeographic break.

It is tempting, and also quite customary, to invoke specific deterministic factors (associated with Pleistocene refugia, for example, in conjunction with contemporary barriers to gene flow at the current suture zone) to account for such phylogeographic discontinuities. However, as first shown by Neigel and Avise (1986, 1993; see also Saunders, Kessler & Avise, 1986), discontinuities in individual gene trees can also arise haphazardly (i.e., without geographic barriers to gene flow) within any continuously distributed low-dispersal species simply as a consequence of idiosyncratic

lineage sorting and stochastic coalescent processes operating on non-recombining stretches of DNA (such as mtDNA, the Y-chromosome, or tightly linked nuclear sequences). Irwin (2002) confirmed and extended these ideas through computer simulations which showed that such gene-tree phylogeographic breaks, often quite deep, can readily arise in continuously distributed species if mean dispersal distances of individuals and/or population sizes in a species are low. Although Irwin's (2002) approach has some limitations (Templeton, 2004; see Discussion), it does offer a relatively straightforward and intuitively simple way to address possible competing evolutionary sources of phylogeographic breaks at an individual locus.

All of these findings beg the question: When can salient splits in particular gene trees be correctly used to infer the presence of salient historical barriers to gene flow?

Here, using computer simulations like those employed by Irwin (2002), we investigate the conditions under which phylogeographic breaks can be interpreted with confidence to signify the existence and location of historical barriers to gene flow (as opposed to mere stochastic effects of lineage sorting under the coalescent).

Materials and methods

Simulation model

Our simulation algorithm for generating gene genealogies is logically similar to the coalescent model described in Irwin (2002). This model uses a backward-in-time approach from coalescent theory (Kingman, 1982; Harding, 1996) and is adapted from the lattice model by Slatkin and Maddison (1990). The study organism is assumed to have non-overlapping generations and is continuously distributed along a linear geographic range (such as a coastline or a river). That range consists of a one-dimensional array of N points, each point representing one individual. In other words, N refers both to the size of the population and the magnitude of its spatial distribution, which are assumed to remain constant through time. Genes chosen for reconstructing genealogies are assumed to be selectively neutral.

Based on settings specifying the number of locations sampled and the number of individuals sampled per location, a collection of individuals is chosen for reconstructing a gene genealogy. For the individuals sampled, each gene tree reflects the actual hereditary pathways for a particular gene (uncomplicated by empirical limitations that would attend any estimate of a gene tree from actual DNA sequence data, for example). The sampling locations were evenly distributed along the geographic distribution and in this study the two outer locations were treated as reflective boundaries. Thus, if the determined movement would take an individual across a boundary or geographic barrier, the organism first moved toward the boundary and then was forced to bounce back the appropriate distance in the opposite direction. [The computer program is more flexible, however, permitting the user also to specify absorbing boundaries or ring habitats.] The physical distance between an individual and its parent was determined by drawing a random magnitude of movement from a normal distribution with mean = 0 and standard deviation $= \sigma_{disp} * N$. The variance in reproductive success across parents implicitly stems from the spatial variance in dispersal, because only one offspring can occupy a given site.

A coalescent event occurs when two or more individuals shared the same parent. In other words, individuals coalesce when the random drawing process determines that their parents are located at the same point on the one-dimensional array. Thus, each set of hereditary pathways can be thought of as analogous to a haploid transmission route (even in a sexual species) such as that traversed by mtDNA, a Y-linked gene, or a haploid nuclear analogue of those as defined and described by Avise and Wollenberg (1997). By moving backward in time through the generations, a genealogy emerges in which all lineages eventually trace back to a single ancestral individual. Each simulated genealogy constitutes one realization of a gene tree under the population conditions specified.

The simulation and data analysis programs are written in C++ programming language. Simulations and data analysis were executed on a computer cluster running Red Hat Linux at the University of Georgia. Source code and sample input/output files are available at http://chkuo.name/.

Simulation settings and data analysis

Following Irwin (2002), three variables were chosen for investigation: level of dispersal ($\sigma_{disp} = 0.00125$, 0.005 and 0.02), population size ($N = 400$, 3200 and 25,600) and temporal duration of the geographic barrier (0, $N/16$, $N/8$, $N/4$, $N/2$, N, $2N$, $4N$, $8N$ and $16N$ generations). The geographic barrier was arbitrarily positioned at the center of the geographic range and it totally blocked gene flow between the two sides. Temporal duration refers to how far back in time the geographic barrier was formed. For each parameter combination, 1000 iterations were performed to generate a collection of independent gene genealogies. Each genealogy was constructed by sampling 10 individuals from each of 40 evenly distributed locations across the species' geographic range.

To measure the strength of the phylogeographic break, we calculated the mean coalescent time of individuals from successive pairs of population sampled along the linear habitat. Two steps of data analysis were performed. First, the location of the deepest phylogeographic break between successive pairs of adjacent populations was recorded for each gene tree to generate a frequency distribution of break locales. Second, for multiple gene trees generated under the same parameter settings, all such possible pair-wise population comparisons were performed to investigate their level of agreement with regard to the spatial locations of the deepest phylogeographic breaks.

Results

All of the simulations yielded qualitatively consistent outcomes, representative examples of which are illustrated in Figures 1 and 2. Figure 1 shows how the location of the deepest phylogeographic break between adjacent populations is influenced by the presence and duration of an historical geographic barrier. In the absence of such a barrier (top panel in each column), the deepest phylogeographic break can appear with appreciable probability at almost any location along the species' distributional range. [Such probabilities were, however, lower near the edges and higher near the center of the range, probably because the reflective boundaries in effect tend to promote higher genetic exchange among adjacent demes that are nearer the range edges].

In the presence of a geographic barrier to gene flow (successive durations of which are shown in panels 2–10 in each column of Figure 1), the probability distribution of the deepest phylogeographic breaks gradually shifted to a sharply unimodal spike centered precisely on the spatial location of the historical barrier to dispersal. In the simulations shown, this transition from a relatively uniform probability distribution (top panel) to a strongly spiked one (lower panels) invariably occurred in less than $2N$ generations since origin of the barrier. [We suspect that the slower transition to a spiked distribution in the smaller populations may be due to greater stochasticity of coalescent processes in those populations, thus making it more likely that for a time, the largest genetic break might occur elsewhere than at the physical dispersal barrier.]

Figure 2 shows the frequency distribution of geographic distances between the deepest genealogical breaks from pairs of independent gene trees generated under the same sets of parameter conditions. When the locations of the deepest genetic breaks in two gene trees are identical, by definition the geographic distance is zero. In the absence of a barrier to gene flow (top panel in each column), the probability distributions of spatial distances between gene-tree breaks were broad and relatively uniform (albeit tapering off at larger geographic distances). In the presence of a geographic barrier to gene flow (successive durations of which are shown in panels 2–10 of each column), the frequency distributions of spatial distances gradually shifted to lower mean values, eventually spiking at zero. In the simulations shown, this spike at zero distance always began to emerge within about $4N$–$16N$ generations or less.

Figure 3 shows comparable outcomes under a set of parameter conditions that includes much higher magnitudes of individual dispersal. This diagram also illustrates the point that the types of pattern changes in Figures 1 and 2 were qualitatively consistent across all of our computer simulations.

Figure 4 plots the outcome of all of our simulations in a different format. In the left column are diagrammed the probabilities that the deepest genealogical break between adjacent populations in a gene tree spatially coincides with an historical

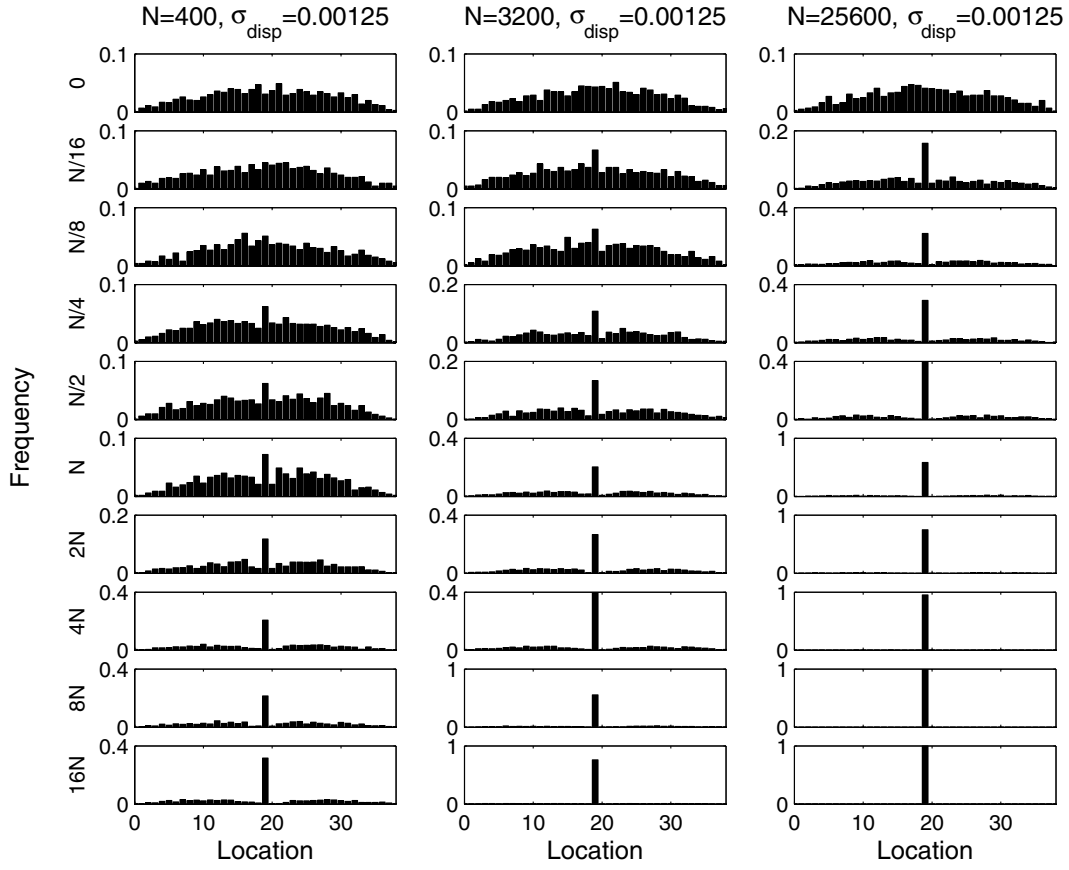

Figure 1. Representative examples of the frequency distributions of locations along a linear habitat where the deepest breaks in intraspecific gene genealogies occurred. The historical barrier that completely blocked gene flow between its two sides was located at the center of the species' range (between locations 19 and 20). Simulation parameters in this case were as follows: number of independent gene trees = 1000; $\sigma_{disp} = 0.00125$; N = 400, 3200, and 25,600 (in columns from left to right, respectively); and duration (in generations) of the geographic barrier (0, $N/16$, $N/8$, $N/4$, $N/2$, N, $2N$, $4N$, $8N$ and $16N$ in successive rows from top to bottom). Note that the frequency scales along the Y-axis vary across different panels in this presentation.

barrier to gene flow, under various sets of parameter conditions in the simulations. These probabilities consistently increase as the duration of the barrier increases and they tend to do so much faster when, under a given population size, the magnitudes of individual dispersal are higher. This pattern probably emerges because higher gene flow regimes tend to spatially homogenize lineages in the gene tree, thereby lessening the probability that stochastic lineage sorting would result in a placement of the deepest genealogical break at a position other than at the dispersal barrier.

In the right column of Figure 4 are diagrammed the probabilities that pairs of independent gene trees share the same location of the deepest genealogical break. These probabilities consistently increase as the duration of the barrier increases, and they usually tend to do so much faster when, for a given population size, the magnitudes of individual dispersal are higher (for the same reasons as mentioned above). By comparing the right-hand and the left-hand columns in Figure 4, it also becomes apparent that given the same duration of an historical barrier, the probability that pairs of independent gene genealogies share the same location of the deepest phylogeographic break is always lower than the frequency that the location of that deepest break coincides with the barrier in any one gene tree.

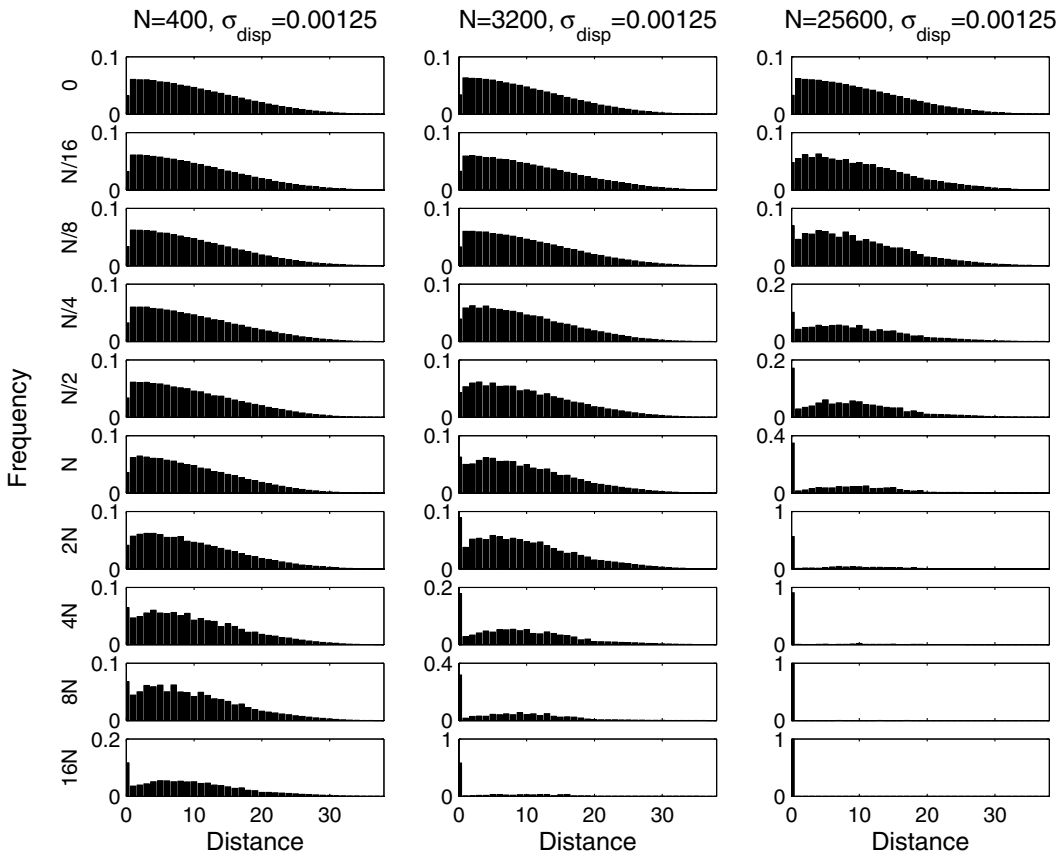

Figure 2. Sample frequency distributions of the spatial distance between the deepest phylogeographic breaks from pairs of independent gene genealogies. Simulation parameters are identical to those described in Figure 1.

Discussion

Our results complement those of earlier simulations by Niegel and Avise (1986, 1993) and Irwin (2002) by again showing that in the absence of dispersal barriers, the deepest population breaks in a gene tree can arise haphazardly at essentially any location along a species' linear distribution. This phenomenon must be viewed as potentially highly troubling for researchers who might wish to use phylogeographic discontinuities in individual gene trees (such as from mtDNA) as unequivocal genetic signatures of firm historical impediments to gene flow.

Our simulations also go beyond those presented earlier by explicitly comparing the spatial positions of genealogical breaks in independent pairs of gene trees as a function of presence (versus absence) and duration of a geographic barrier to dispersal. Our findings demonstrate that in the absence of dispersal barriers, the spatial positions of the deepest population breaks in independent gene genealogies are likely to contradict one other. Such contradictions gradually shift to agreement or concordance, however, when geographic barriers are present and have been in place for more than approximately $2N$–$16N$ generations (depending on the sets of parameters simulated). Thus, concordance or lack thereof in the spatial positions of genealogical breaks across multiple unlinked loci becomes itself a deciding criterion by which to distinguish between stochastic and deterministic causation in accounting for phylogeographic discontinuities in continuously distributed species.

In commenting on Irwin's (2002) computer simulations, and on empirical genealogical surveys based on mtDNA, Templeton (2004) pointed out that phylogeographic breaks in a gene tree can sometimes appear simply as an artefact of sparse

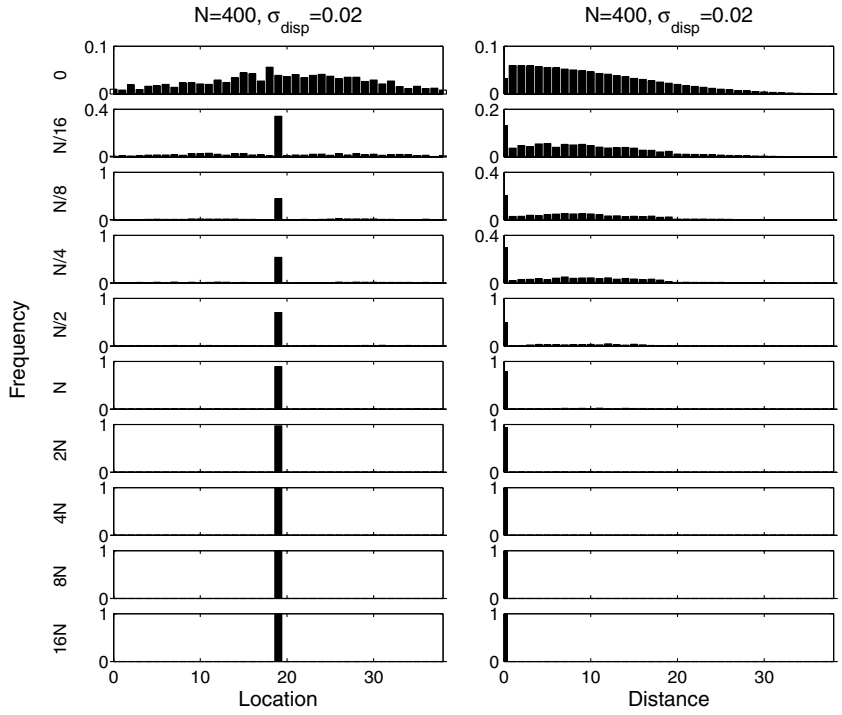

Figure 3. Sample frequency distributions of the locations of the deepest phylogeographic breaks in intraspecific gene genealogies. Simulation conditions here are exactly as described in Figure 1 except for the following: $\sigma_{disp} = 0.02$; $N = 400$.

sampling of geographic locales. However, such sampling artefacts would be unlikely to account for phylogeographic breaks that are evidenced as spatially concordant patterns across multiple unlinked loci, given that the number of locales examined is not unduly low.

In the current simulations, we used information only on the location (and not the absolute depth) of the most salient phylogeographic breaks. We feel this is justified in the current context for several reasons. First, in empirical studies, gene genealogies reconstructed from molecular data may not be entirely accurate with regard to true branch lengths. Second, even if gene trees reconstructed from molecular data were fully accurate, the depths of their phylogeographic breaks have high stochastic variation under a given set of population conditions (Irwin, 2002). Third, from empirical experience, molecular genealogies often appear more likely to agree on their spatial locations than on the apparent absolute depths of their major phylogeographic breaks (several examples are summarized in Chapter 5 of Avise, 2000). Finally, as shown in the current study, agreement between independent genealogies with regard to the location of the deepest phylogeographic breaks can provide strong support for concluding that an historical barrier actually caused the genealogical discontinuities (and conversely, that disagreements between the genealogies of unlinked loci normally indicate that the phylogeographic breaks probably did not result from a longstanding historical barrier).

This study helps to formalize and quantify a more intuitive notion that has been termed 'aspect II' (among a total of four distinct aspects) of 'genealogical concordance' (Avise & Ball, 1990; Avise, 1996, 2000). Aspect II refers to the agreement across independent loci with respect to the geographic positions of salient genealogical breaks within a species. In phylogeographic inference, aspect II concordance has been deemed a sufficient if not necessary condition (depending on the strength of empirical support for the other aspects of genealogical concordance) for properly inferring the presence of a longstanding historical barrier to gene flow.

Merely for the sake of simplicity, our current simulations involved linear habitats (as did those of Neigel & Avise, 1986; Irwin, 2002). In the

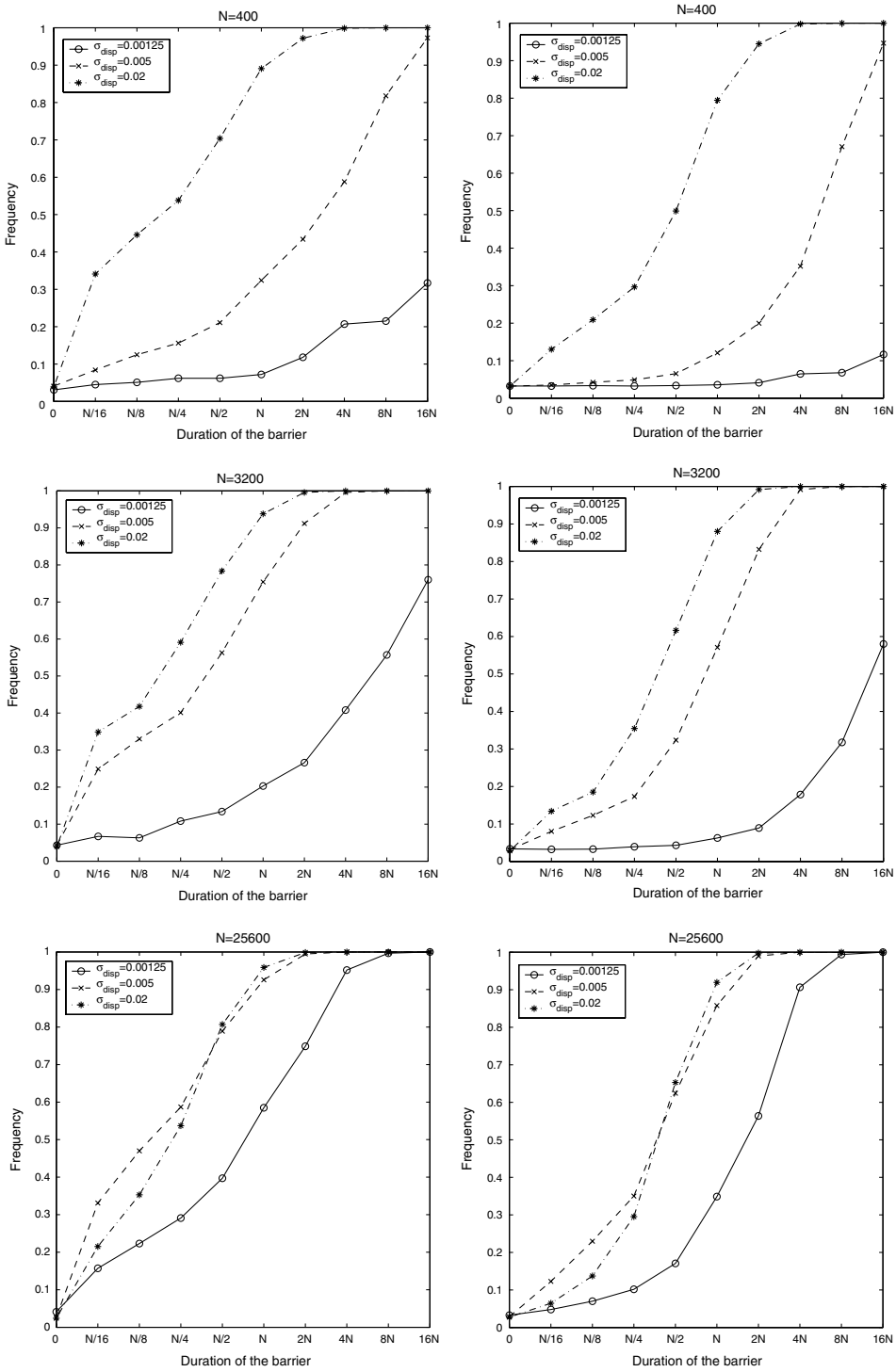

Figure 4. Results for all simulated parameters of population size, dispersal magnitude, and duration of a geographic barrier to gene flow. *Left column*: frequency in which the location of the deepest phylogeographic break coincides perfectly with the historical barrier. *Right column*: frequency in which pairs of independent gene genealogies share the same location of the deepest phylogeographic break.

future, it may be desirable to model two-dimensional environments in similar fashion, although this will probably be a much more challenging bookkeeping task. Our models also involved relatively low-dispersal species only. However, the results summarized in Figure 4 suggest that geographic barriers would be even more effective in rapidly generating concordant phylogeographic breaks across multiple gene trees in species with higher magnitudes of dispersal.

Our current models assume selective neutrality for the loci surveyed. Conclusions may not always hold (or may do so with diminished force) if strong balancing selection has acted on the loci in question. In such cases, major genealogical breaks would be less likely to localize to the position of any historical geographic barrier that in truth did exist. Conversely, strong disruptive selection could cause the emergence of concordant genetic clines or genealogical breaks across multiple loci even without benefit of longstanding population separations (Endler, 1977). For example, in the grass species *Agrostis tenuis* and *Anthoxanthum odoratum*, sharp and concordant clinal variation at several loci clearly was promoted by disruptive selection for metal tolerance and flowering times across the ecotones between normal pastures and lead–zinc mines (McNeilly & Antonovics, 1968; Antonovics & Bradshaw, 1970).

However, it seems unlikely that either balancing or disruptive selection of such high intensities would normally characterize mitochondrial genomes or most other molecular marker systems normally employed in phylogeographic surveys. Thus, we conclude that genealogical concordances (or lack thereof) should continue to be a key consideration when inferring phylogeographic processes from empirical phylogeographic patterns.

Acknowledgements

This work was supported by funds from the University of Georgia. We thank J. Kissinger and P. Brunk for technical support and we thank J. Bouzat, M. Mackiewicz, J. Mank, D. Promislow, A. Tatarenkov and an anonymous reviewer for helpful comments that improved the manuscript.

References

Antonovics, J. & A.D. Bradshaw, 1970. Evolution in closely adjacent plant populations VIII. Clinal patterns at a mine boundary. Heredity 25: 349–362.

Avise, J.C., 1996. Toward a regional conservation genetics perspective: phylogeography of faunas in the southeastern United States, pp. 431–470 in Conservation Genetics: Case Histories from Nature, edited by J.C. Avise & J.L. Hamrick. Chapman & Hall, New York.

Avise, J.C., 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge.

Avise, J.C. & R.M. Ball, 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxford Surv. Evol. Biol 7: 45–67.

Avise, J.C. & D. Walker, 1999. Species realities and numbers in sexual vertebrates: perspectives from an asexually transmitted genome. Proc. Natl. Acad. Sci. USA 96: 992–995.

Avise, J.C. & K. Wollenberg, 1997. Phylogenetics and the origin of species. Proc. Natl. Acad. Sci. USA 94: 7748–7755.

Bernatchez, L., 1995. A role for molecular systematics in defining evolutionarily significant units (ESU) in fishes. Am. Fish. Soc. Symp. 17: 114–132.

Endler, J.A., 1977. Geographic Variation, Speciation, and Clines. Princeton University Press, Princeton.

Harding R.M., 1996. New phylogenies: an introductory look at the coalescent, pp. 15–22 in New Uses for New Phylogenies, edited by P.H. Harvey, A.J. Leigh Brown, J. Maynard Smith & S. Nee. Oxford University Press, Oxford.

Irwin, D.E., 2002. Phylogeographic breaks without geographic barriers to gene flow. Evolution 56: 2383–2394.

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13: 235–248.

McNeilly, T. & J. Antonovics, 1968. Evolution in closely adjacent plant populations. IV. Barriers to gene flow. Heredity 23: 205–218.

Moritz, C., 1994. Applications of mitochondrial DNA analysis in conservation: a critical review. Mol. Ecol. 3: 401–411.

Neigel, J.E. & J.C. Avise, 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation, pp. 515–534 in Evolutionary Processes and Theory, edited by E. Nevo & S. Karlin. Academic Press, New York.

Neigel, J.E. & J.C. Avise, 1993. Application of a random-walk model to geographic distributions of animal mitochondrial DNA variation. Genetics 135: 1209–1220.

Ryder, O.A., 1986. Species conservation and the dilemma of subspecies. Trends Ecol. Evol. 1: 9–10.

Saunders, N.C., L.G. Kessler & J.C. Avise, 1986. Genetic variation and geographic differentiation in mitochondrial DNA of the horseshoe crab, *Limulus polyphemus*. Genetics 112: 613–627.

Slatkin, M. & W.P. Maddison, 1990. Detecting isolation by distance using phylogenies of genes. Genetics 126: 249–260.

Templeton, A.R., 2004. Statistical phylogeography: methods of evaluating and minimizing inference errors. Mol. Ecol. 13: 789–809.