Syst. Biol. 63(5):743–752, 2014

DOI:10.1093/sysbio/syu040
Advance Access publication June 13, 2014

Modelling Competition and Dispersal in a Statistical Phylogeographic Framework

Louis Ranjard[1], David Welch[2], Marie Paturel[3], and Stéphane Guindon[3,4,*]

[1]*Bioinformatics Institute;* [2]*Department of Computer Sciences;* [3]*Department of Statistics, The University of Auckland, New Zealand and* [4]*LIRMM, CNRS, Montpellier, France;*

[*]*Correspondence to be sent to: Department of Statistics, The University of Auckland, New Zealand, E-mail: s.guindon@auckland.ac.nz*

Received 16 March 2014; reviews returned 27 March 2014; accepted 14 May 2014
Associate Editor: Edward Susko

Abstract.—Competition between organisms influences the processes governing the colonization of new habitats. As a consequence, species or populations arriving first at a suitable location may prevent secondary colonization. Although adaptation to environmental variables (e.g., temperature, altitude, etc.) is essential, the presence or absence of certain species at a particular location often depends on whether or not competing species co-occur. For example, competition is thought to play an important role in structuring mammalian communities assembly. It can also explain spatial patterns of low genetic diversity following rapid colonization events or the "progression rule" displayed by phylogenies of species found on archipelagos. Despite the potential of competition to maintain populations in isolation, past quantitative analyses have largely ignored it because of the difficulty in designing adequate methods for assessing its impact. We present here a new model that integrates competition and dispersal into a Bayesian phylogeographic framework. Extensive simulations and analysis of real data show that our approach clearly outperforms the traditional Mantel test for detecting correlation between genetic and geographic distances. But most importantly, we demonstrate that competition can be detected with high sensitivity and specificity from the phylogenetic analysis of genetic variation in space. [Competition; dispersal; phylogeography.]

Introduction

Deciphering the processes that generated the current spatial distribution of organisms is central to our understanding of biodiversity (Crisp et al. 2010). Beside the many evolutionary forces taking place at various spatial and time scales, the observed distribution results from a series of past successful colonization events (MacArthur and Wilson 1967; MacDonald 2003; Knowles 2009). The success or otherwise of colonizations is governed by the migrants' ability to disperse, their suitability to the new habitat, and competition with already established populations at the new locations. Phylogeographic models that have been proposed to date explicitly account for dispersal but all ignore competition. The model introduced in the present study overcomes these limitations and defines a sound statistical framework to accommodate for both dispersal and competition.

Dispersal is the movement from a birthplace to a new site (Brown and Lomolino 1998). Whether organisms and their propagules propel themselves or are carried by wind, water, or other organisms, it is regularly assumed that the probability of dispersal decreases with distance from the point of origin (Ōkubo and Levin 2001). In cases where the dispersal range is limited and the mutation rate sufficiently high, we expect genetic and geographic distances to be correlated (Clobert 2001; Begon et al. 2006). Therefore, incorporating spatial information into the analysis of genetic patterns can shed light on the underlying dispersal processes.

The competitive exclusion (CE) principle states that two species competing for the same resource in a constant environment cannot coexist (Gause 1932; Hardin 1960; Ayala 1971). Recent support for the CE principle comes from, for instance, Cooper et al. (2008) who used a phylogenetic approach to demonstrate that competition determines the assembly of mammal communities. Their study confirmed that CE among closely related species sharing a habitat eliminates inferior competitors, causing unrelated species to assemble into phylogenetically over-dispersed communities (Hutchinson 1959).

The definition of CE was extended recently (Waters 2011; Waters et al. 2013) to account for evidence that competition also takes place between conspecific populations. For instance, the grasshopper *Chorthippus parallelus* displays high level of genetic homogeneity due to rapid dispersal events following the last glacial maximum in high-latitude Northern Hemisphere (Hewitt 1996). The prevention of secondary colonization provides the most likely explanation for these patterns. On a different scale, Hallatschek et al. (2007) demonstrated experimentally that neutral mutations can spread through large populations of bacteria and segregate the gene pool into well-defined, sector-like regions of reduced genetic diversity. Excoffier and Ray (2008) were able to reproduce these results *in silico* under a simple model of population expansion in two dimensions. They then argued that "single range expansion can create very complex patterns at neutral loci, mimicking adaptive processes and resembling postglacial segregation of clades from distinct refuge areas." Hence, evidence is strong that CE is frequent and occurs at widely different time and geographic scales (but see Tilman (2011) for a differing view).

Landscape genetics (Manel et al. 2003; Wang 2010) and phylogeography (Avise 2009; Riddle 2009; Knowles 2009) are two relatively young disciplines that combine genetic and geographic data to study the forces shaping

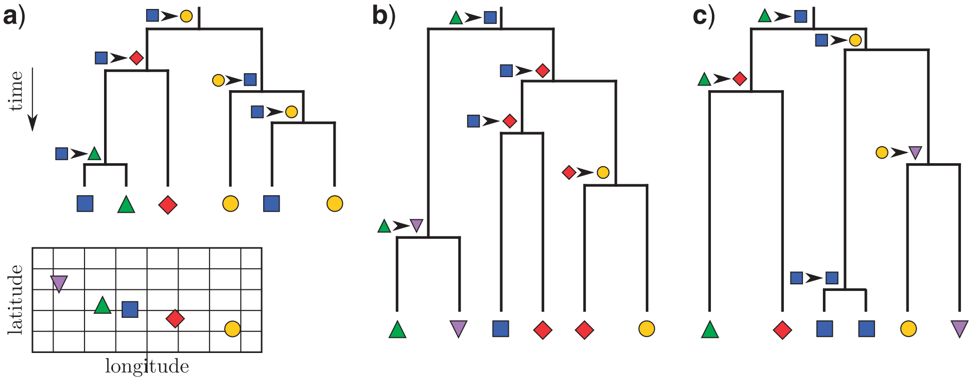

FIGURE 1. Phylogeographic signal of competition and biased dispersal. The three scenarios are based on the landscape and colored locations displayed at the bottom left. a) Competition is weak and dispersal is not biased toward short distances. b) Competition is weak and dispersal is biased. c) Competition is strong and dispersal is not biased. Node heights correspond to the times at which speciation/dispersal events occur (see text). When conditioning on the phylogenies having the same number of tips, the height of the trees with competition (c) is greater than that without competition as the rate of speciation/dispersal decreases with the diminishing number of empty locations. Node heights are also greater when dispersal is biased (b) compared to the unbiased case and speciation/dispersal events consist in rare "jumps" between adjacent locations.

biodiversity. Landscape genetics focuses on short time scales and relies on traditional population genetics techniques. Features of the landscape that correlate with the observed genetic patterns can be identified, for example, using a Mantel test (Mantel 1967; Sokal and Rohlf 1995). Phylogeography generally deals with longer time scales, using data and techniques from a range of disciplines including phylogenetics, climatology, geology, evolutionary genomics, and ecology to explain the geographic distribution of organisms (Avise et al. 1987; Hickerson et al. 2010; Chan et al. 2011; Ronquist and Sanmartín 2011).

Coalescent-based phylogeographic methods that accommodate for factors influencing the spatial variation of genetic patterns across closely related organisms are increasingly gaining interest (e.g., Beerli and Felsenstein 1999; Ewing et al. 2004; Lemey et al. 2010). Such approaches account for uncertainties in estimates of the genealogy and rely on measurably evolving markers such as viral sequences (Lemey et al. 2010) or human languages (Walker and Ribeiro 2011). Coalescent-based methods differ from traditional phylogeographic methods in that they rely on stochastic models to explain the observed spatial distribution of organisms. Their main focus is on estimating posterior distributions of parameters of interest, including genealogies and possible migration paths, rather than mapping particular events onto a fixed genealogy.

None of the aforementioned approaches explicitly model dispersal, competition and their interaction. A possible explanation for this methodological gap is that competition violates the assumption of lineage independence in a genealogy and therefore presents various technical challenges. The present study overcomes these limitations and introduces a phylogeographic approach that explicitly models dispersal and CE in a statistical framework. Our model defines a landscape as a set of vacant locations that are colonized through a series of dated dispersal events. Although we define CE as any process preventing secondary colonization, we do not identify the specific causal mechanism involved (e.g., adaptation, sexual selection, or competition for resource).

Using simulations, we show that dispersal and CE parameters can be recovered with high accuracy, even from small datasets. If genetic and geographic distances are dependent, our method detects it in approximately 80% of the cases, which compares favorably to 40% with the Mantel test. If they are independent, both methods perform well and display similar specificity. Most importantly, our model detects CE with high accuracy: when competition does occur, our approach detects it with a probability generally greater than 0.9; when competition does not occur, our approach correctly fails to detect it with a probability also greater than 0.9 in most cases. The analysis of data from the *Banza* genus in Hawaii, for which prior evidence suggests that CE plays an important part, illustrates the relevance of our approach.

MODEL

The proposed model takes genetic sequences and their spatial coordinates as data in order to date dispersal events and estimate dispersal and CE parameters. Before we present the details of our model, we will argue that, in principle, genetic and geographic data contain information about competition.

Figure 1 illustrates our reasoning. We assume that speciations are the immediate consequence of dispersal events. Thus, internal nodes in a phylogenetic tree correspond to coupled speciations/dispersals. The height of a node represents the time at which the corresponding event occurred. In this section of the article, node heights are considered as known. In practice, these heights are estimated from homologous molecular sequences provided some calibration information is available.

Figure 1a displays a phylogeny with node heights matching those expected under weak CE and no dispersal bias toward short distances. Dispersal is biased toward short distances and CE is strong in Figure 1b and c respectively. When dispersal is biased, locations are generally colonized by populations established nearby and long-range dispersals are rare. For the same landscape, dispersal events occur through long- and short-range "jumps" at similar rates if dispersal is not biased (Fig. 1a), whereas only short-range jumps and/or longer waiting times between them are observed in the biased situation (Fig. 1b). When competition is absent or weak, the per-lineage rate of dispersal is constant during the course of evolution and across lineages. The expected time to the next dispersal event along a given lineage does therefore not depend on which locations in the landscape are already occupied. Comparison between Figure 1a and c shows that, when CE is strong, the time between successive speciation events increases as the number of lineages grows. This observation is a consequence of the number of unoccupied locations decreasing and CE preventing the dispersal to already occupied locations.

The Dispersal and Colonization Process

The evolutionary unit (EU) of interest corresponds here to a population. An internal node in the phylogeny has a part of a source population establishing a new population at a postdispersal location, whereas the source population remains in the original location. Such an event can lead to reproductive isolation.

According to our model, multiple populations occupy a fixed number of locations. Each population produces migrants at some constant rate which attempt to colonize a new location (that may be the same as the original location). The success of the attempt depends on the distance between the two locations and whether or not the new location is occupied. The complete process is described as follows. Each population is represented by a lineage labeled with the corresponding location information. A lineage branches whenever there is a successful dispersal originating from that population. The rate at which each lineage branches out at a given point in time depends on its location label as well as the labels of other lineages at that particular time. Also, there is no extinction of lineages in our model.

We now formally describe this process, starting with the definition of the landscape. A landscape consists of m distinct geographic locations whose spatial coordinates are observed. Write $l_i=(l_{i1},\ldots,l_{ic})$ for the i-th location defined in c dimensions. c is typically two, though the inclusion of altitude or nonspatial dimensions, such as temperature, may have $c>2$. A location is either unoccupied or occupied by one or more populations. Let u be a vector indicating the number of populations at each location, so that $u_i\in\{0,1,2,3,\ldots\}$ and $u_i=0$ if and only if l_i is unoccupied. The process starts with a single population at a random location, l_i, where $i\sim U(1,m)$.

Each population produces dispersal attempts according to a Poisson process with constant rate, τ. If dispersal range is unlimited and CE plays no part, every dispersal attempt would be successful and the new colony would establish at location i with probability $1/m$, for all i. This would give a dispersal rate originating from a population at l_i and establishing at l_j of τ/m for all j.

In general, we allow that dispersal range is limited and CE may have some effect. The dispersal range of a migrant is given by the dispersal kernel, which we choose to be a normal distribution centered at the current location. Let $f(x,y)=f(x;y,\Sigma)=\exp\left(\sum_{i=1}^{c}-\frac{(x_i-y_i)^2}{2\sigma_i^2}\right)$ be the (unnormalized) density function of a multivariate normal distribution centered on y with covariance matrix $\Sigma=\mathrm{diag}(\sigma_1,\ldots,\sigma_c)$. Define F to be the $m\times m$ matrix with entries $F_{ij}=f\left(l_i,l_j\right)/mf(l_i,l_i)$. So $F_{ii}=1/m$ and $0<F_{ij}<1/m$ for $i\neq j$.

We account for the differential probability of successfully colonizing occupied versus unoccupied locations by attaching weight λ to the dispersal rate when the new location is occupied. Let Λ be a vector of length m with entries $\Lambda_i=\lambda$ if $u_i>0$ and $\Lambda_i=1$ if $u_i=0$ for $i=1,\ldots,m$. The total rate of (successful) dispersal for a population at location i to location j is $R_{ij}=\tau F_{ij}\Lambda_j$.

Note that when $\lambda=1$, there is no distinction between occupied and unoccupied locations at the same distance. When $\lambda<1$, unoccupied locations are preferred, indicating some form of CE, while $\lambda>1$ means already colonized locations are easier to colonize than unoccupied locations. The rate that any population establishes new colonies in its current location, i say, is $R_{ii}=\tau\lambda/m$, since $F_{ii}=1/m$ and $\Lambda_i=\lambda$ when l_i is occupied.

For a given set of locations, the above process is completely defined when values have been assigned to τ,σ and λ. We call these parameters the overall dispersal rate, dispersal parameter, and competition parameter, respectively.

Likelihood

In this section, we provide an expression for the likelihood $f(g,l|\tau,\sigma,\lambda)$, of observed and (imputed) ancestral locations, l, and tree, $g=(V,E,t)$, where V is the vertex set of g, E is the edge set and t is a vector of times of length $|V|$. Suppose that g has n leaves, so that $|V|=2n-1$ and each vertex $v\in V$ is associated with a time t_v. Each edge $e=(v_i,v_j)$ in the tree is associated with a location, $l(e)$ say. For $v\in V$, let $p(v)\in E$ denote the parent (or in-) edge of v and $c_1(v),c_2(v)\in E$ denote the children (or out-) edges of v, where they exist. Use the convention that $c_1(v)$ represents the population remaining in the ancestral location so that $l(c_1(v))=l(p(v))$. Each nonleaf vertex in the tree represents a dispersal event which occurs at time t_v and involves individuals from $l(p(v))$

establishing a colony at $l(c_2(v))$. Label the vertices according to time with vertex 1 being the root, vertex $n-1$ being the most recent dispersal event and vertices $n,\dots,2n-1$ being the leaves. Let t_k be the time of the k-th vertex. We assume that all samples were taken at the same time so all leaves have time t_n. For a given time, t, $t_1 \leq t \leq t_n$, define $E(t)$ to be the set of edges extant at t, so that $E(t)=\{e : e=(v_i,v_j)\in E \text{ and } t_{v_i} \leq t \leq t_{v_j}\}$. We extend the notation from the previous section so that, for a given time t, $u(t)=(u_1(t),\dots,u_m(t))$ is the occupancy status of locations at t. Note that $\sum_{i=1}^{m} u_i(t)=|E(t)|$ is the number of lineages in the tree at time t. Let $R(t)$ be the total rate of dispersal in the system at t. Each population in location i sends migrants to location j at rate R_{ij} and, since there are $u_i(t)$ populations at location i at time t

$$R(t)=\sum_{i=1}^{m}\sum_{j=1}^{m} R_{ij}u_i(t).$$

Thus, for $1<i<n$, the waiting time between the $(i-1)$-th and i-th vertices in the tree contributes a factor of

$$R(t_{v_i})\exp\left(-R(t_{v_i})(t_{v_i}-t_{v_{i-1}})\right)$$

to the likelihood where $R(t_{v_i})$ denotes the total rate of dispersal in the $[t_{v_{i-1}},t_{v_i}]$ interval. If $i<n$, so that v_i corresponds to a dispersal event, the probability of this dispersal event occurring at this point out of all possible dispersal events is

$$\frac{R_{l(p(v_i)),l(c_2(v_i))}}{R(t_{v_i})}.$$

The contribution to the likelihood for concurrent leaf vertices is 1. The likelihood is thus

$$f(g,l|\tau,\sigma,\lambda) = \frac{1}{m}\frac{w(v_1)}{\sum_{j=1}^{m} R_{l(c_1(v_1)),j}} \times \prod_{i=2}^{n} w(v_i)\exp(-R(t_{v_i})(t_{v_i}-t_{v_{i-1}})), \quad (1)$$

where $w : V \to \mathbb{R}$ is the function defined by

$$w(v)=\begin{cases} 1 & \text{if } v \text{ is a leaf,} \\ R_{l(p(v)),l(c_2(v))} & \text{otherwise} \end{cases}$$

and we have accounted for the special case at the root, which contributes the factors before the main product. Figure 2 provides a graphical illustration of the key ingredients of the likelihood calculation, that is the vector u and the matrix R, and how these change along the tree. In particular, as the likelihood calculation proceeds down the tree, the number of locations that are available as departure points increases, hence the changing dimension of R. The total dispersal rate τ and the competition parameter λ are confounded in the last two time slices.

Parameters of the model are estimated in a Bayesian framework using Markov chain Monte Carlo. The algorithm in this study uses Metropolis–Hastings moves to sample from the joint posterior distribution of λ, σ and τ and the unknown geographical locations on internal edges. The joint prior distribution of these three parameters is uniform with the lower and upper bounds for λ and τ set to 10^{-3} and 10^{3} respectively. The lower bound for σ is set to 10^{-3} while the upper bound for that parameter depends on the distribution of distances between pairs of locations in the landscape under study (Supplementary Appendix 2, http://dx.doi.org/10.5061/dryad.9pq70)

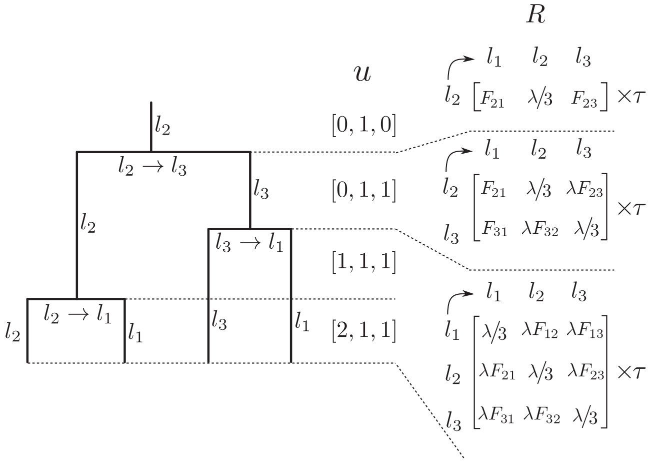
R
u
l_1 l_2 l_3
l_2 $[F_{21}\ \lambda/3\ F_{23}]\times\tau$
l_2
$[0,1,0]$
$l_2 \to l_3$
l_1 l_2 l_3
l_2 F_{21} $\lambda/3$ λF_{23}
l_3 F_{31} λF_{32} $\lambda/3$
$\times\tau$
l_3
$[0,1,1]$
l_2
$l_3 \to l_1$
$[1,1,1]$
l_1 l_2 l_3
$l_2 \to l_1$
l_2
l_1
l_3
l_1
$[2,1,1]$
l_1 $\lambda/3$ λF_{12} λF_{13}
l_2 λF_{21} $\lambda/3$ λF_{23}
l_3 λF_{31} λF_{32} $\lambda/3$
$\times\tau$

FIGURE 2. Elements of the likelihood calculation. Each edge in the tree is labeled with a corresponding location, l_1, l_2, or l_3. The dispersal events are indicated underneath each vertex. By convention, the child edge to the left of the parent inherits the parental location. The occupancy vector u and the dispersal rate matrix R are displayed to the right of the tree for each relevant time slice. The final two time slices share the same rate matrix since the list of occupied locations is the same in those intervals.

RESULTS

Recovering the Model Parameters

We simulated data according to our model, added noise to the obtained phylogeny, and estimated the dispersal and competition parameters (see Supplementary Appendices 1 and 2 for a more detailed description of the simulation settings). The "true" values of these parameters were chosen so as to cover a broad range of dispersal and competition conditions. In particular, values of the competition parameter λ were chosen such that about 65% of them were smaller than 1.0 to emphasize cases where CE takes place. Also, values of the dispersal parameter, σ, were selected such that approximately half of them corresponded to "uniform dispersal", that is when ignoring competition, locations at short and long distances from the current position of a dispersing EU have the same probability of being colonized.

The results in Figure 3 were obtained with trees comprising 100 EUs and 50, 100, or 150 locations. With 150 locations, the posterior median of λ is generally close to its true value. When competition is very strong though

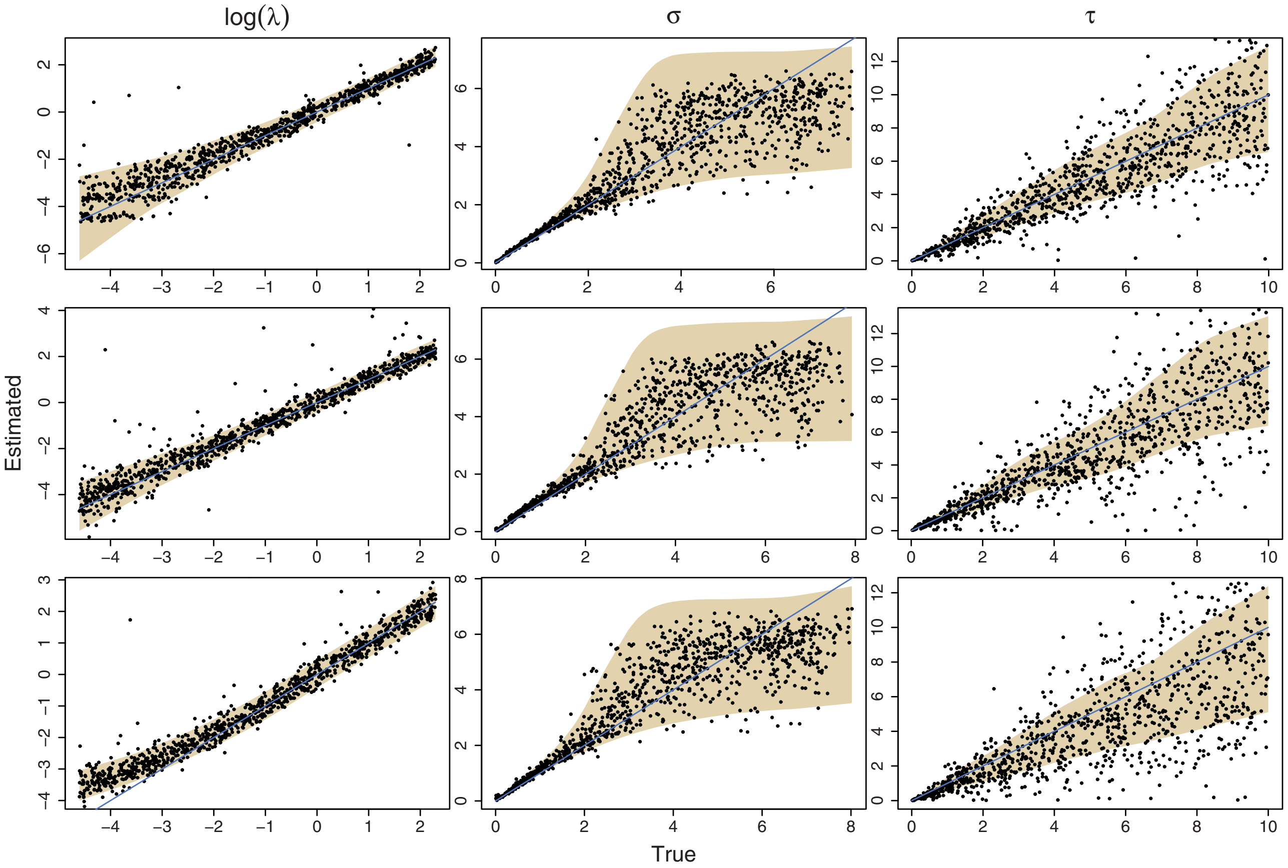

FIGURE 3. Posterior estimates of competition (λ) and dispersal (σ and τ) parameters—100 taxa, 150 locations (top row), 100 locations (middle), and 50 locations (bottom). The logarithm of the true values of λ are uniformly distributed in [log(0.01), log(10)]. The true values of σ are uniformly distributed in [0.001, 2σ*], where σ* is a threshold for this parameter such that dispersal is uniform when σ ≥ σ* (see Supplementary Appendix 1). The true values of τ are uniformly distributed in [0.01, 10]. Estimated parameter values correspond to posterior medians. The shaded (brown) areas delimit the 95% credibility interval for the corresponding parameter (Supplementary Appendix 3).

($\log(\lambda) \leq -2$) this parameter is slightly overestimated. The bias is more obvious when considering smaller numbers of EUs (Supplementary Appendix 3). Also, for small values of λ, the variance of the estimates across simulations is higher than that observed with larger values of λ. When the number of locations greatly exceeds that of the EUs, a dispersing EU will generally migrate to an empty location anyway, no matter how intense CE is. Opportunities for competition to occur are therefore scarce, making the corresponding parameter difficult to estimate, thus explaining the increased variance of our estimates. In a landscape where the number of taxa greatly exceeds the number of locations (Figure 3, bottom-left corner), ancestral EUs are bound to migrate to occupied locations at some stage during evolution as all the available locations have already been colonized. The choice of a new location for a dispersing EU is then solely determined by its distance to the current location and competition plays no part. This could explain the difficulty in accurately estimating λ when competition is very strong. Note also that with 50 locations and 100 EUs, the overall rate at which dispersal occurs, τ, is often underestimated (bottom-right corner). Here again, when every location is occupied, all the dispersal rates (R_{ij}) are a function of the product τλ. Since λ tends to be overestimated as already explained, τ is consequently underestimated. For similar reasons, large estimated values of λ (≥2) are generally compensated with small estimated values of τ (see Supplementary Appendix 4). This observation is particularly true for small datasets, although the correlation between these two parameters is limited overall.

The number of locations does not impact on the estimation of the dispersal parameter σ. When dispersal is biased toward short distances, corresponding to small values of σ, this parameter is estimated with high precision. However, beyond a certain threshold, the estimates, while still accurate, lack precision. Dispersal events are no longer biased toward short distances for values of σ greater than ∼5. In fact, for such values of this parameter, the normal kernel used to model dispersals is virtually identical to a uniform distribution. Hence, in this range of values of σ, the likelihood function is essentially flat, which explains the large variance of the estimates.

In the simulations presented here, we considered only census sampling, that is a tree was generated according to our model and all lineages in that tree were sampled.

TABLE 1. Sensitivity and specificity for the detection of competition and biased dispersal

H0 →		"no competition"		"no dispersal bias"			
# EUs	# loc.	Sens.	Spec.	Sens.	(Mantel)	Spec.	(Mantel)
100	150	0.960	0.981	0.860	(0.458; 0.526)	0.887	(0.948)
100	100	0.968	0.966	0.859	(0.512; 0.550)	0.882	(0.914)
100	50	0.970	0.969	0.807	(0.468; 0.470)	0.920	(0.926)
50	100	0.928	0.951	0.796	(0.459; 0.481)	0.881	(0.894)
50	50	0.964	0.951	0.787	(0.428; 0.548)	0.853	(0.928)
50	20	0.934	0.963	0.756	(0.425; 0.429)	0.883	(0.891)
20	50	0.918	0.928	0.747	(0.327; 0.475)	0.802	(0.919)
20	20	0.908	0.920	0.732	(0.368; 0.524)	0.820	(0.926)
20	10	0.899	0.891	0.697	(0.379; 0.525)	0.814	(0.934)

# EUs: number of evolutionary units. # loc.: number of locations. Two values are given for the sensitivity of the Mantel test for each number of EUs and locations. The first corresponds to the proportion of rejected null hypothesis of no dispersal bias using a nominal 5% rejection threshold. The second corresponds to the proportion of rejected null hypothesis with the rejection threshold adjusted such that the specificity of the Mantel test matches that of the model-based approach.

In practice, however, only a subsample of lineages are available. To assess the effect of noncensus sampling, we performed additional simulations in which randomly chosen subsets of lineages of different sizes were used for the inference. The results presented in Supplementary Figure S7 show that the estimates of the competition and dispersal parameters are still reasonably accurate even in cases where only 20% of the tips are sampled. The overall dispersal/speciation rate parameter τ is the only one that shows evidence of bias. This result is not surprising since, according to our model, τ directly determines the expected number of taxa. Overall, the results obtained suggest that the proposed model is likely to detect dispersal bias and competition in realistic experimental conditions where only a subset of all existing EUs were collected (see Supplementary Appendix 7).

Sensitivity and Specificity

Using the same simulation settings, we next aimed to assess the sensitivity and specificity of our approach for detecting CE and bias in dispersal. For each simulated dataset, the null hypothesis of no competition was rejected if the posterior median estimate of λ was smaller than 1.0. For dispersal, we calculated the threshold value for σ beyond which 95% of the area under the normal density used to model dispersal overlapped with that of a uniform distribution (Supplementary Appendix 1). The null hypothesis of no dispersal bias was rejected when the posterior median estimate of σ was smaller than this threshold.

The specificity and sensitivity for various numbers of EUs and locations are presented in Table 1. These results indicate that our approach successfully detects competition when it does occur, that is its sensitivity is high, even for small datasets with 20 EUs and 10 locations. The same conclusion applies to specificity: our model correctly fails to reject the null hypothesis of no competition in more than ~90% of the simulated datasets generated without competition. Moreover, the estimates of λ are only slightly affected by the topological distance between the true phylogeny and that used for the inference (Supplementary Appendix 5). By extension, the accuracy with which the topology is estimated will not impact on the sensitivity and specificity of the test of the null hypothesis of no competition, provided this tree is still reasonably close to the true one.

The performance of our model for testing the null hypothesis of no dispersal bias is also satisfactory, both in terms of specificity and sensitivity. For the same specificity, the power of our test for no dispersal bias is on average 1.56 times that of the traditional Mantel test.

Case Study

The Hawaiian archipelago is formed by volcanic activity, each island arising one after the other as the Pacific plate moves across a geological "hot spot." It constitutes a well-documented natural laboratory to test hypotheses about the evolution of model species. Colonization from major land masses is limited and dispersal almost exclusively takes place across islands of the archipelago. Strong competition can therefore potentially take place when new islands become available for colonization.

The phylogeny of the Hawaiian endemic species of the *Banza* genus (Hawaiian katydids) shows evidence for island radiation with different species inhabiting different islands (Shapiro et al. 2006). This tree exhibits a striking "progression rule" pattern whereby each island appears to have been colonized only once. No obvious environmental feature varying across islands explains the observed geographical distribution of these organisms (Roderick and Gillespie 1998). Adaptation is therefore unlikely to be the main evolutionary force explaining the observed spatial distribution. Instead, rapid divergence in courtship or sexual behavior following dispersal events can increase the rate of speciation as illustrated in another Hawaiian cricket species (Mendelson and Shaw 2005). Incompatible mating behavior could therefore prevent secondary colonizers from mating with the local population, leading to reproductive isolation in *Banza* (Shapiro et al.

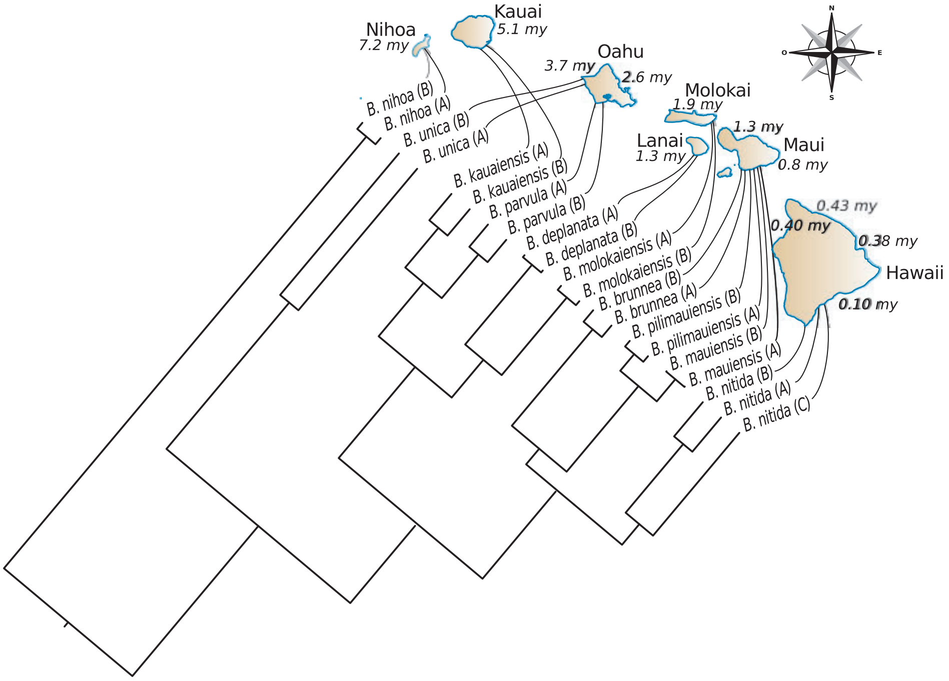

FIGURE 4. Phylogenetic tree and geographical locations of subspecies from the *Banza* genus. The node heights in the tree are posterior averages obtained through the Bayesian analysis of an alignment of 21 homologous nucleotide sequences, 2003-bp long.

2006). Such a process would be considered as CE in our model.

The competition-dispersal model was fitted to a rooted phylogenetic tree with node heights expressed in calendar time units (Fig. 4 and Supplementary Appendix 6). The posterior distributions of σ and λ are displayed at the top of Figure 5. The estimated posterior distribution of λ suggests that CE indeed takes place, with the majority of sampled values of that parameter smaller than 1 (the posterior probability for the event $\lambda < 1$ is equal to 0.61). Also, the inferred posterior density of σ indicates a clear bias toward short-range dispersals. The vast majority of the values sampled by the MCMC algorithm fall below the threshold corresponding to uniform dispersal (see the dotted vertical line on Fig. 5, top-left corner). The graph at the bottom-left corner of Figure 5 displays the distribution of the dispersal distances under a model with no dispersal bias (i.e., the uniform distribution) against the normal density with variance estimated from the data. Our model thus shows that a large proportion of jumps correspond to dispersal events within a small radius while large jumps are expected to be more frequent if dispersals were not biased. These two results are not particularly surprising as the phylogeny of these species itself strongly suggests that colonization took place via short jumps between neighboring islands. The fact that each of these events only occurred once suggest that competition is indeed playing a central role here.

The traces shown in Figure 6d–f do not indicate any particular issue with the mixing of the MCMC run. However, the scatterplot of the sampled values of λ and τ reveals potential identifiability problems for these two parameters (Fig. 6a), with a strong linear correlation between the logarithm of the sampled values ($r = -0.94$). As mentioned previously, when all locations are occupied, τ and λ occur as a product in the likelihood function, thus generating nonidentifiability in the lower part of the phylogeny. Nonetheless, these two parameters are separate in the upper part of the tree, suggesting each of them could in principle be estimated. Depending on the ratio of the number of EUs and the number of locations, nonidentifiability could be a serious problem (many EUs and few locations) or only a minor issue (few EUs compared to the number of locations).

In order to assess the degree of nonidentifiability with the *Banza* dataset, we incorporated a dummy variable in the likelihood function that occurs as a multiplicative factor of τ at all stages during the likelihood calculation. τ and this dummy variable are therefore completely confounded. A comparison of the scatterplot for the dummy variable *vs.* τ and λ *vs.* τ shows that a much wider range of sampled values for each of these two parameters is obtained when they are completely confounded (Supplementary Appendix 6, Fig. S6). Thus, even though τ and λ appear to be partially confounded here, a clear signal for each of these two parameters can still be extracted from the data.

DISCUSSION

The present study introduces a model that detects and quantifies the effects of competition and biased dispersal on populations of closely related species during the course of their evolution. We postulate that

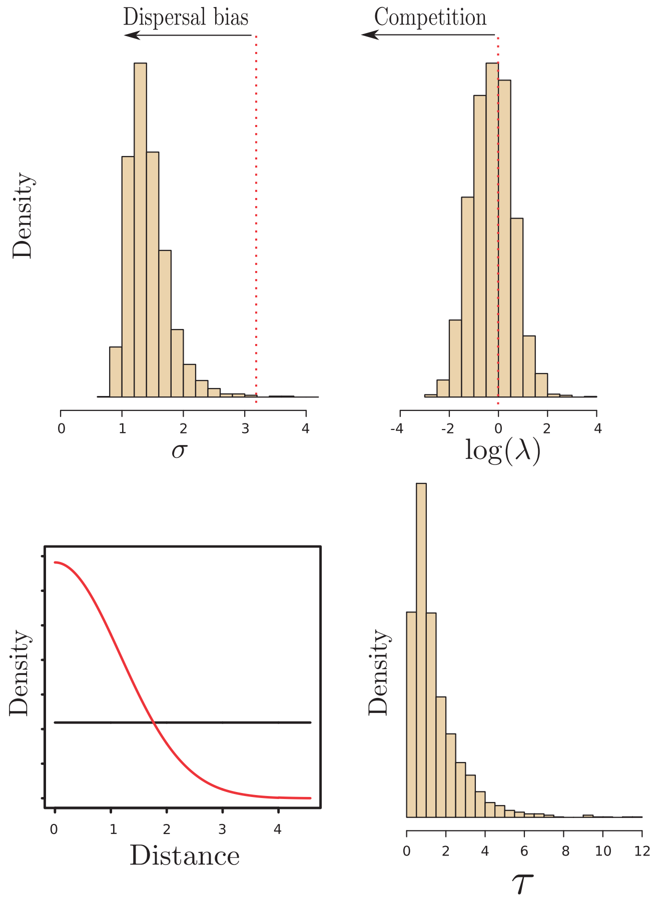

FIGURE 5. Posterior distribution of parameter estimates for the *Banza* data set. The two histograms at the top correspond to the posterior distributions of the parameters λ and σ, measuring competition intensity and dispersal bias, respectively. The dotted lines indicate thresholds below which competition occurs and dispersal is biased toward short distances. The histogram at the bottom gives the posterior distribution of τ, the overall dispersal rate parameter. The densities on the bottom-left corner are that of a uniform (in black) in [0,4.56], where 4.56 is the mean distance between islands (the unit is not relevant) and that of a normal density (in red) with mean 0 and variance equal to the posterior median of σ, truncated to [0,4.56].

the geographic distribution of the relevant evolutionary units results from a series of dispersal events that can be mapped onto their phylogeny/genealogy. Crucially, the rate of dispersal depends on the state of occupancy of the locations to be colonized.

Our simulations indicate that bias in the dispersal toward short distances can be detected in a broad range of conditions, including small to relatively large numbers of lineages or locations. Most importantly, competitive exclusion leaves a clear signature in the data that our approach helps to reveal. Nonetheless, as with any statistical model, the assumptions and approximations underlying our competition-dispersal model require careful scrutiny.

First and foremost, our model assumes that, at any point in the past, a given location is either free or occupied by an ancestor of one of the sampled organisms. However, a free location could in fact have been occupied by an ancestor which did not leave any descendant in the sample or by an ancestor that did not survive to the present. If competition indeed impacts on the colonization process, our model will then overestimate dispersal distances. Note however that the impact of such competing “ghost” ancestors will only be problematic if their density varies across locations. A uniform density throughout geographic locations would affect every dispersal event to the same extent and would therefore not hamper the inference of the competitive exclusion parameter characterizing the lineages sampled.

Moreover, occupied locations may be more likely to be recorded in the data compared to unoccupied ones in practice. “Presence-only” data are indeed commonplace in ecology (Pearce and Boyce 2006). Our simulations show that one can estimate competition and dispersal parameters in cases where every location is occupied. Nonetheless, a more satisfactory approach would incorporate the probability for any lineage to occupy a given location, rather than the “present or absent” approach implemented in this study.

We also assume that locations are defined in an obvious manner, typically corresponding to isolated islands. In its current implementation, our model does not apply to continuous landscapes where nonoverlapping locations are not readily defined. It would however be possible to extend it so that the strength of competition decreases continuously with spatial distance between populations. Such a modification would be the only requirement for our model to be able to deal with the continuous case.

Also, the competition-dispersal model constrains dispersal events to occur exclusively at the internal nodes of the gene genealogy. This assumption does no longer hold if dispersal is not followed by reproductive isolation, as it is often the case when focusing on intraspecies or intrapopulation data. A vast literature that takes its root in Wright's F_{ST} focuses on that problem (see, e.g., Felsenstein 2013). Relatively recent methodological advances, such as the structured coalescent (see Nordborg 2008), could help us relaxing that constraint. However, according to this last model, dispersal events are independent of one another, which contradicts the principle of competitive exclusion. Although adapting the structured coalescent to account for competition appears to be a challenging task, earlier work on a similar model (Wakeley 2001) could pave the way for further progress in this direction.

We elected to use a normal kernel for modelling dispersal distances in this study. The validity of the assumption of a Brownian motion governing the movements of individuals when considering long-time ranges is difficult to assess. Moreover, it is likely that distinct species will display different dispersal characteristics, suggesting that a single dispersal kernel may not fit all datasets. Fortunately, the structure of our model allows the simple incorporation of any continuous distribution, such as the exponential (Clark et al. 1999; Skarpaas et al. 2004), as dispersal kernel.

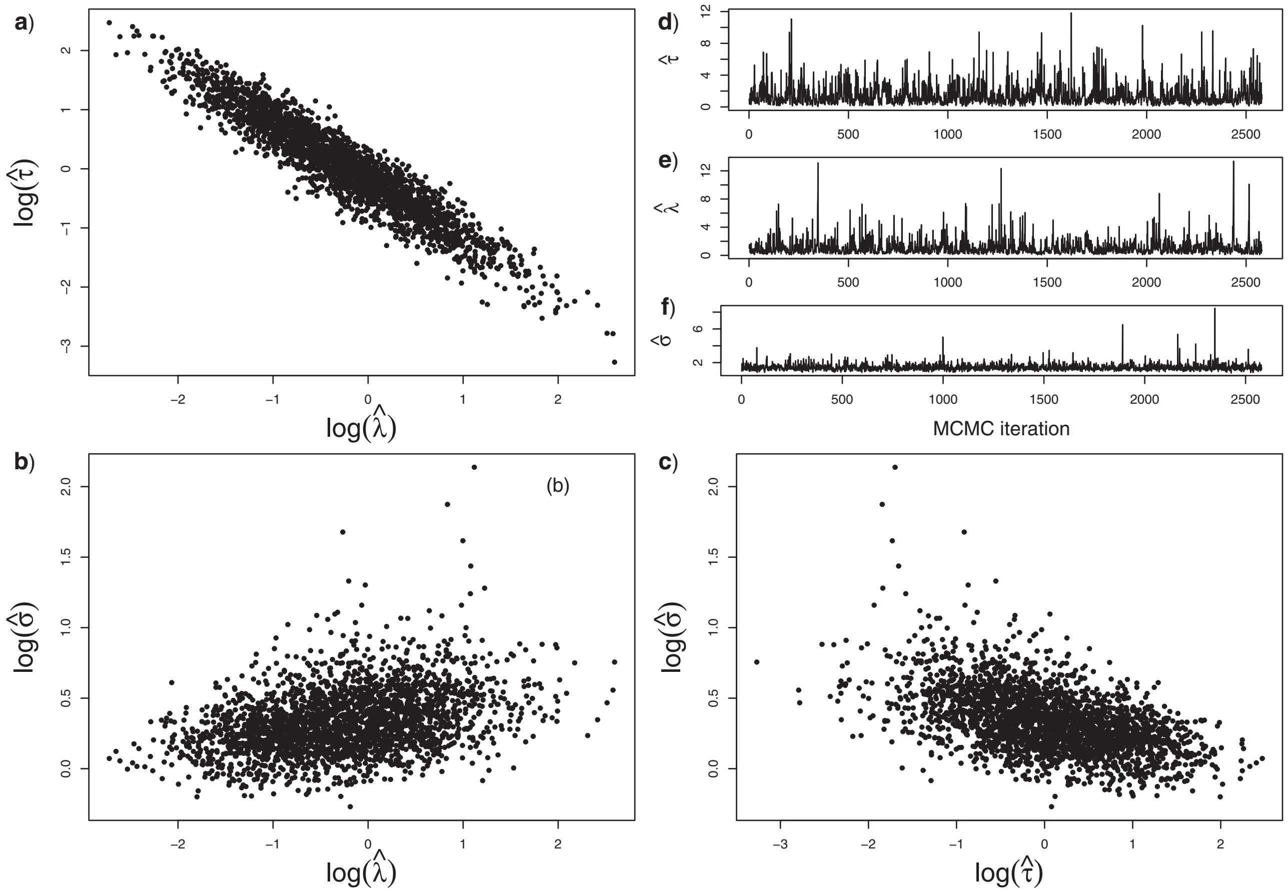

FIGURE 6. Scatterplots and traces of the MCMC analysis for the three model parameter estimates. a–c scatterplots of parameter values sampled during the MCMC for each pair of parameters. d–f traces of the MCMC analysis for each parameter.

Inference of dispersal and competition parameters requires a strong phylogenetic signal. Our model will therefore perform poorly if applied to intrapopulation data where genetic divergence is weak and does not permit accurate estimation of phylogenies/genealogies. At that scale, modeling dispersal as a continuous flux between demes is probably more realistic than the approach proposed here. Methods such as those described in Beerli and Felsenstein (1999) and Ewing et al. (2004) are therefore more relevant in that context, even though they do not account for competition. Our model is not suited to the analysis of distantly related species either. First, the landscape is unlikely to remain unchanged over the time scales involved. Also, complex processes such as the extinction of lineages and the increased impact of adaptation over deep evolutionary scales will most likely hamper the inference of competition or dispersal.

Finally, we fitted our model to phylogenies/genealogies with fixed tree topology and node heights. Although our results suggest that the three parameters of the dispersal model are estimated accurately even when the model is fitted to a tree distinct from the true one (see Supplementary Appendix 5), uncertainty around node heights and tree topology should be accounted for in practice. Importantly, only the relative heights of the nodes matter to the estimation of the competition and dispersal bias parameters (λ and σ respectively). Moreover, estimates of the overall dispersal parameter (τ) are only meaningful when node heights are expressed in calendar units. Nonetheless, provided the rates of nucleotide or amino acid substitution do not vary drastically across lineages (as is expected when analyzing closely related species), molecular sequence data, without fossil data, should in theory provide enough information for accurate estimation of the most relevant parameters, that is λ and σ.

Despite obvious limitations and simplifications, the competition-dispersal model introduced in this study is relevant to the analysis of a wide range of relatively closely related lineages. Its parameters are straightforward to interpret and can be estimated accurately from the phylogenetic analysis of homologous genetic sequence. This approach should therefore provide crucial insights into the importance of competition and dispersal for shaping the observed spatial patterns of genetic diversity in a variety of experimental conditions.

Software

The model presented here is implemented in PhyML (http://code.google.com/p/phyml or email s.guindon@auckland.ac.nz) as well as the `phyloland` package written in the `R` programming language (http://CRAN.R-project.org/package=phyloland or email l.ranjard@auckland.ac.nz). The package `phyloland` contains the *Banza* data set used for the case study.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.9pq70.

Funding

FRDF grant awarded by the University of Auckland to [L.R.].

Acknowledgments

The authors wish to acknowledge the Centre for eResearch at the University of Auckland and NeSI high-performance computing facilities.

References

Avise J. 2009. Phylogeography: retrospect and prospect. J. Biogeography 36:3–15.

Avise J., Arnold J., Ball R., Bermingham E., Lamb T., Neigel J., Reeb C., Saunders N. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann. Rev. Ecol., Evol. Sys. 18:489–522.

Ayala F.J. 1971. Competition between species: frequency dependence. Science 171:820–824.

Beerli P. and Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773.

Begon M., Townsend C.R., Harper J.L. 2006. Ecology: from individuals to ecosystems. Hoboken: Wiley-Blackwell.

Brown J., Brown T., Lomolino M. 1998. Biogeography. Sunderland: Sinauer Associates.

Chan L.M., Brown J.L., Yoder A.D. 2011. Integrating statistical genetic and geospatial methods brings new power to phylogeography. Mol. Phylogenet. Evol. 59:523–537.

Clark J.S., Silman M., Kern R., Macklin E., HilleRisLambers J. 1999. Seed dispersal near and far: patterns across temperate and tropical forests. Ecology 80:1475–1494.

Clobert J. 2001. Dispersal. Oxford: Oxford University Press.

Cooper N., Rodríguez J., Purvis A. 2008. A common tendency for phylogenetic overdispersion in mammalian assemblages. Proc. R. Soc. B Biol. Sci. 275:2031–2037.

Crisp M., Trewick S., Cook L. 2010. Hypothesis testing in biogeography. Trends Ecol. Evol. 26:66–72.

Ewing G., Nicholls G., Rodrigo A. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations (MEPs). Genetics 168:2407–2420.

Excoffier L. and Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. Trends Ecol. Evol. 23:347–351.

Felsenstein J. 2013. Theoretical evolutionary genetics. Distributed by the author.

Gause G. 1932. Experimental studies on the struggle for existence: 1. Mixed population of two species of yeast. J. Exp. Biol. 9: 389–402.

Hallatschek O., Hersen P., Ramanathan S., Nelson D.R. 2007. Genetic drift at expanding frontiers promotes gene segregation. Proc. Nat. Acad. Sci. 104:19926–19930.

Hardin G. 1960. The competitive exclusion principle. Science 131:1292–1297.

Hewitt G.M. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. Biol. J. Linnean Soc. 58: 247–276.

Hickerson M., Carstens B., Cavender-Bares J., Crandall K., Graham C., Johnson J., Rissler L., Victoriano P., Yoder A. 2010. Phylogeographys past, present, and future: 10 years after. Mol. Phylogenet. Evol. 54:291–301.

Hutchinson G.E. 1959. Homage to Santa Rosalia or why are there so many kinds of animals? Am. Nat. 93:145–159.

Knowles L.L. 2009. Statistical phylogeography. Ann. Rev. Ecol. Evol. Syst. 40:593–612.

Lemey P., Rambaut A., Welch J.J., Suchard M.A. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27:1877–1885.

MacArthur R.H. and Wilson E.O. 1967. The theory of island biogeography. Princeton, NJ: Princeton University Press.

MacDonald G. 2003. Biogeography: introduction to space, time, and life. illustrated ed. Hoboken: Wiley.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27:209–220.

Manel S., Schwartz M., Luikart G., Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol. Evol. 18:189–197.

Mendelson T.C. and Shaw K.L. 2005. Sexual behaviour: rapid speciation in an arthropod. Nature 433:375–376.

Nordborg M. 2008. Coalescent theory. Hoboken: John Wiley & Sons Ltd. pp. 843–877

Ōkubo A. and Levin S. 2001. Diffusion and ecological problems: modern perspectives, vol. 14. Berlin: Springer.

Pearce J.L. and Boyce M.S. 2006. Modelling distribution and abundance with presence-only data. J. Appl. Ecol. 43:405–412.

Riddle B. 2009. What is modern biogeography without phylogeography? J. Biogeogr. 36:1–2.

Roderick G. and Gillespie R. 1998. Speciation and phylogeography of Hawaiian terrestrial arthropods. Mol. Ecol. 7:519–531.

Ronquist F. and Sanmartín I. 2011. Phylogenetic methods in biogeography. Ann. Rev. Ecol. Evol. Syst. 42.

Shapiro L., Strazanac J., Roderick G. 2006. Molecular phylogeny of *Banza* (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago. Mol. Phylogenet. Evol. 41:53–63.

Skarpaas O., Stabbetorp O., Rønning I., Svennungsen T. 2004. How far can a hawk's beard fly? Measuring and modelling the dispersal of *Crepis praemorsa*. J. Ecol. 92:747–757.

Sokal R.R. and Rohlf F.J. 1995. Biometry. New York: W. H. Freeman and Co.

Tilman D. 2011. Diversification, biotic interchange, and the universal trade-off hypothesis. Am. Nat. 178:355–371.

Wakeley J. 2001. The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. 59:133–144.

Walker R. and Ribeiro L. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. Proc. R. Soc. B: Biol. Sci. 278:2562–2567.

Wang I. 2010. Recognizing the temporal distinctions between landscape genetics and phylogeography. Mol. Ecol. 19:2605–2608.

Waters J.M. 2011. Competitive exclusion: phylogeography's 'elephant in the room'? Mol. Ecol. 20:4388–4394.

Waters J.M., Fraser C.I., Hewitt G.M. 2013. Founder takes all: density-dependent processes structure biodiversity. Trends Ecol. Evol. 28:78–85.