Modelling competitive exclusion and limited dispersal in a statistical phylogeographic framework

Louis Ranjard, David Welch, Stéphane Guindon

February 19, 2013

Abstract

Competitive exclusion embodies the idea of the first mover advantage where species or populations arriving first at a suitable location prevent secondary colonisation of the already occupied site. While 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. In particular, competition could explain observed patterns of low genetic and phenotypic diversity following rapid colonisation events in Europe as well as the "progression" pattern observed in the phylogeny of species living on various islands along the Hawaiian archipelago. Competitive exclusion has been absent from past quantitative analyses because of the difficulty in designing adequate methods for assessing its impact. We present here a new statistical framework that integrates competition along with limited dispersal into a phylogenetic model of migration. Using simulations, we assess the power and sensitivity of our approach and demonstrate its ability to detect competition from the comparative analysis of homologous genetic sequences using geographic information.

1 Introduction

Identifying the processes responsible for the spatial distribution of species and populations is the fundamental problem in biogeograpy [13]. Candidate processes include biotic and abiotic processes [48] such as vicariance, long distance and restricted dispersal, population fragmentation and contiguous range expansion. Among these, dispersal and competition are widely recognised as important forces shaping spatial species distributions [28, 21, 29, 42]. Dispersal is the movement away from a birthplace to a new site [5]. Whether organisms and their propagules move 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. A widely used dispersal kernel is thus a normal distribution centred at the location of the last ancestor, where the variance of the kernel is proportional to the dispersal range [34]. Long-distance dispersal events, called jump dispersals, are sometimes observed, particularly in seed dispersal, and require explicit modelling by heavier-tailed distributions [9, 6]. At long time scales, it is commonly assumed that limited dispersal kernels are do not limit an organism's distribution because of the accumulation of many short-distance events and rare long-distance events [7, 11].

Nonetheless, if dispersal is limited and if the mutation rate is sufficiently high, isolation by distance can lead to correlations between genetic distance and geographic distance [10, 4]. Thus the analysis of genetic patterns in space can shed light on underlying disposal processes. To accurately understand these processes, we need to model dispersal in a flexible manner by accounting for uncertainty about the dispersal kernel [32, 24].

Competition is another force shaping the geographic distribution of species and populations. It is widely held that species that compete for the same resource are less likely to occupy the same ecological niche [17, 19, 3]. It has been proposed [46] that such competition may prevent secondary colonisation of space at the population level, in a process known broadly as competitive exclusion. Competitive exclusion prevents interbreeding between closely related populations, thereby promoting speciation by increasing genetic isolation in cases of incomplete parapatric or peripatric speciation.

Evidence for competitive exclusion has been seen in microbial communities where one population of microorganisms are eliminated from a common habitat [16] and in barnacle species where competition coupled with environmental factors can explain species geographical range limits [47]. On a larger scale, it has been argued that competition for ecological niches can limit range expansion preventing diversification and speciation in birds [36].

Some examples, however, have been described where even in the presence of competition, secondary colonisation and co-occurence of similar species does occur [48]. Several studies suggest that competition is not a major driver of geographic distribution among species [29]. In plants, Crawford and Whitney (2010) have shown that density does not affect colonisation success but that high genetic diversity of the dispersers increases the chance of successful colonisation.

Pigot et al. (2010) and Wiens (2003) flagged the need for explicit statistical models and inference of competitive exclusion, in particular at the intraspecific level [46], to accurately assess its impact on the history of landscape colonisation.

We propose that these questions can best be answered by the use of genetics data. Biogeography has greatly benefited from the large amount of genetic data generated by recent advances in molecular techniques and increased computer power [32, 23, 41]. In particular, the fields of phylogeography and landscape genetics have emerged as major disciplines over the last 15 years.

Landscape genetic methods focus on relatively short time scales and rely on techniques from population genetics [30, 45]. They aim to explain the spatial distribution of genetic diversity which is typically estimated through the distribution of haplotypes and the reduction of this data to pairwise genetic distances. Features of the landscape that correlate with the observed genetic patterns can be identified, for example, using a Mantel test [43]. These methods are somewhat limited by their treatment of the data as pairwise distances rather than using the more detailed full sequence.

Phylogeography uses phylogenies estimated from sequences to explain species geographic distribution over greater time scales than landscape genetics [1, 39]. It is an integrative approach combining inputs from a broad range of disciplines such as genetics, climatology, geology, evolutionary genomics and ecology [2, 20, 8, 41].

The most widely-used program in this area is DIVA [40]. DIVA uses an event-based approach to reconstruct biogeographic histories with ancestral locations at the internal nodes of a phylogeny. The reconstructed succession of events — including vicariance, dispersal, extinction or sympatric speciation — is the one that most parsimoniously explains the observed data.

More recently, coalescent-based methods have been developed to identify contemporary factors affecting genetic variation [15, 27]. Such approaches take into account uncertainty in genealogies and rely on measurably evolving markers such as viral sequences [26] or human languages [18, 44]. These methods differ from 'classical' approaches such as DIVA in that they employ parametric models and so the main focus is on estimating parameters and uncertainty rather than mapping particular events onto a genealogy. For instance in Lemey et al. (2010) , dispersal is modelled as a spatial diffusion process that is running along the genealogy and the posterior distribution of the diffusion parameter is estimated.

None of the methods discussed so far explicitly model dispersal, competition and their interaction [24, 31, 46, 38, 25, 13].

In this paper, we present a new phylogeographic approach that explicitly models dispersal and competitive exclusion in a unified statistical framework. Our model can be used to test whether these two factors have had a significant impact on the process of landscape colonisation by the evolutionary units of interest (i.e., individuals, populations or species).

We consider a migration model where the landscape is defined as a set of vacant locations that are colonised through a sequence of dated migration events. Internal nodes in the estimated tree correspond to speciations and migration/dispersal events. As a consequence, patterns of migration define the shape and node heights of the underlying genealogy. Geographic distances between the different locations are considered as fixed during the history of migration, restricting our model to short time scales that are typical of population data.

The model has two parameters in addition to those governing the genealogical processes. The first parameter controls the dispersal of individuals, independent of competition. The second parameter controls competition, modelling the relationship between the occupancy of a location and the probability of migrating to that location.

We will show that we can accurately estimate the parameters of this colonisation model within a Bayesian framework. We will also show, using simulated data, that we can estimate dispersal paths in the landscape (biogeographic corridors) using just sequence and location data, even in the presence of competitive exclusion

The remainder of the paper is organised as follows. In Section 2.1, we formally introduce the model and, in Section 3, introduce the Bayesian Markov chain Monte Carlo (MCMC) algorithm used to estimate the model parameters. In Section 4, we present results from numerical simulations used to assess the accuracy of parameter estimates. We finish in Section 5 by discussing our results and proposing various limitations of and extensions to the model.

2 Model

The proposed model relies on using the information conveyed by genetic sequences to decipher the timing of migration and speciation events. When incorporating geographical information, it then becomes possible to quantify the strength of competition between taxa. Without entering yet into the specifics of our model, Figure 1 presents a proof of concept, demonstrating that molecular sequences indeed convey information about competition. The trees on the left and on the right handsides display nodes with distributions of heights matching that expected under strong and weak competition respectively. We assume here that the colonized landscape is very similar in these two examples. This landscape is made of discrete locations (e.g., islands) that can potentially be colonized by any of the taxa under study. We also consider that internal nodes in these trees correspond to events where speciation (or reproduction) and migration occur concomitantly. Their heights represent the times at which these events occured. For the tree on the right, corresponding to weak competition, the rate (per lineage) at which speciation/migration event occur is constant throughout the time period considered. In fact, since competition is weak, the expected time to the next speciation/migration event does not depend on the occupation state of the territory. For the tree on the left, the distribution of node heights is biased toward large (i.e., old) values. Such bias is the consequence of an increase of the amount of time to the next succesful speciation/migration event as the number of unoccupied land patches decreases with time. In such situation, competition prevents the migration to already occupied locations, hence delaying the next succesful migration event. In short, when calibration information is available, it is possible to estimate node heights in terms of calendar time units. Our model then exploits the distribution of these heights in order to extract signal about the migration process.

2.1 The migration process

The evolutionary unit of interest (EU) in the model may be individuals of the same species or populations of related species. When the EUs are single individuals, a migration corresponds to a reproductive event where the off-spring attempts to migrate to a new post-dispersal location while the parent remains in the original location. When the EUs are populations, a migration event has a part of a source population establishing a new

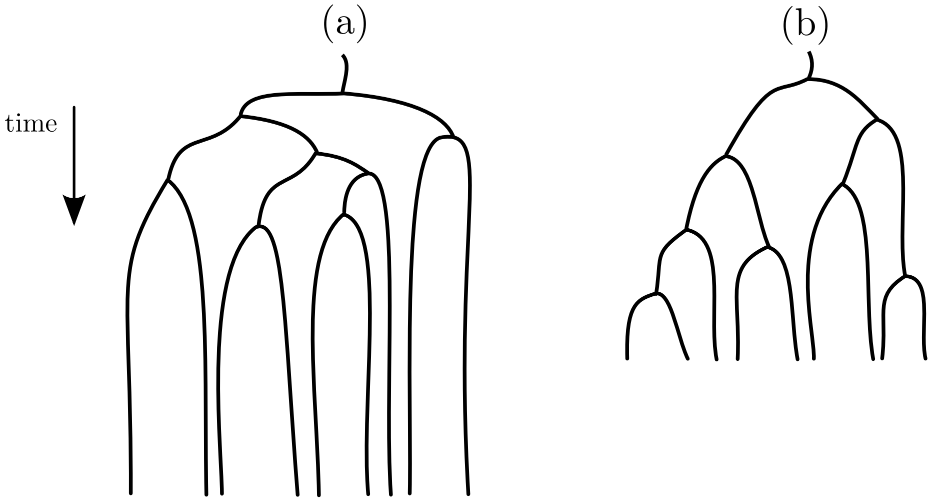


Figure 1: **Competition between taxa is strong for the tree on the left and weak for the tree on the right.** Node heights in these two trees correspond to the times at which speciation/migration occur (see text).

population at a post-dispersal location while the source population remains in the original location. In the following, we assume that the EU corresponds to populations.

According to the migration model, multiple populations occupy a fixed number of locations. Each population produces migrants at some constant rate which attempt to colonise a new location (which may be the same as the original location). The success of the attempt will depend on the distance between the two locations and whether or not the new location is occupied. The migration process from each population can be viewed as a thinned Poisson process where the level of thinning changes as the occupancy of the locations changes. The complete process can be viewed as an inhomogeneous branching process where each population is represented by a lineage that branches whenever there is a successful migration originating from that population.

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 ith location defined in c dimensions. c is typically two, though the inclusion of non-spatial dimensions 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 migration attempts according to a Poisson process with constant rate, τ. If dispersal range is unlimited and competitive exclusion plays no part, every migration attempt would be successful and the new colony would establish at location i with probability $1/m$, for all i. This would give a migration 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 competitive exclusion may have some effect. The dispersal range of a migrant is given by the dispersal kernel, which we choose to be a normal distribution centred 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 (unnormalised) density function of a multivariate normal distribution centred on y with diagonal covariance matrix Σ. 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 colonising 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) migration 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 competitive exclusion while $\lambda > 1$ means already colonised locations are easier to colonise 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^2 and λ. We call these parameters the migration rate, dispersal parameter and exclusion parameter, respectively.

2.2 Viewing the process as a genealogy

A realisation of the migration process can be represented as a genealogy which is just a binary tree with times associated with the nodes. Each population is represented by a lineage and a migration originating from that population is represented in the tree as a node with two child lineages. One child lineage represents the originating population that remains, while the other represents the newly establish population at the new location.

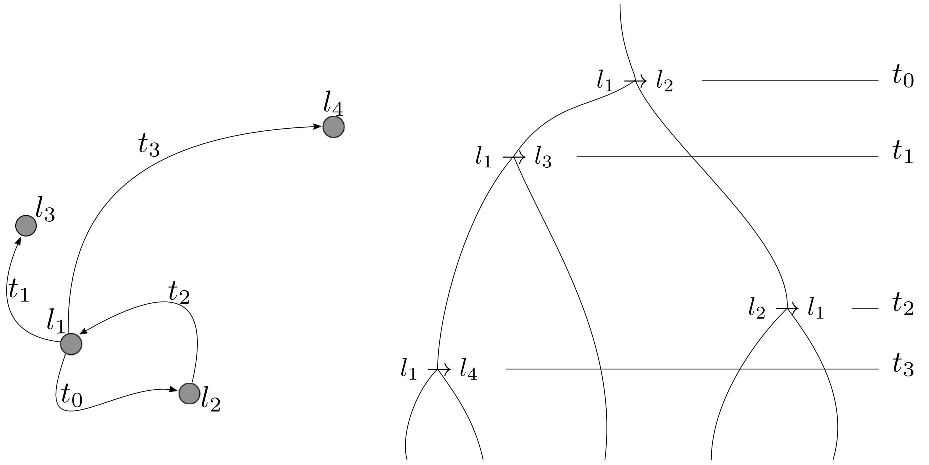


Figure 2: **Two visualisations of the same realisation of the migration process on four locations.** The initial population is found at l_1. At time t_0, a successful migration originating at l_1 has a colony established at l_2 and so on. By convention, the left child of a node represents the population remaining in the ancestral location. We stop the process shortly after the t_3 when there are five populations (lineages) on the four locations.

The time associated with the node is precisely the time of the migration event. A realisation of the migration process and the associated genealogy is given in Figure 2. Where convenient, we employ the tree metaphor to describe the migration process.

2.3 Simulating from the model

For clarity, we provide an algorithm for simulating from the model. This algorithm takes as input a set of coordinates and returns a rooted phylogeny with N taxa.

```
Data: $l_i$ for $i \in [1, m]$
Add $v_0$ to $V$ ;                    /* $V$ is the set of vertices. $v_0$ is above the root */
Add $e_1$ to $E$ ;                    /* $E$ is the set of edges that can branch */
$c_1(v_0) \leftarrow e_1$ ;           /* Make $e_1$ child of $v_0$ */
$t_0 \leftarrow 0$ ;                  /* Set the time for $v_0$ */
$k \leftarrow 1$ ;                    /* $k$ is the number of lineages */
$i \sim \text{Uniform}[1, m]$ ;       /* Choose the first location */
$l(e_1) \leftarrow l_i$ ;             /* Assign it to $l(e_1)$ */
$u_j \leftarrow 0, \forall j \in [1, m]; u_i = 1;$   /* Update vector of occupancy */
while $k \leq N$ do
    $t \sim \text{Exp}(R)$, where $R = \sum_i^m \sum_j^m R_{ij} u_i;$
    $t_k \leftarrow t_{k-1} + t;$     /* Time of $k$th vertex */
    if $k < N$ then
        $p_x \leftarrow \sum_i^m R_{l(e_x)i}/R, \forall e_x \in E;$
        Sample $e_x$ with probability $p_x$;   /* Select edge to branch */
        Add $v_k$ to $V$ ;            /* Make node */
        $p(v_k) \leftarrow e_x$, $c_1(v_k) \leftarrow e_{k+1}$ and $c_2(v_k) \leftarrow e_{k+2}$ ;   /* Make daughter edges */
        Remove $e_x$ from $E$. Add $e_{k+1}$ and $e_{k+2}$ to $E$;   /* $e_x$ can no longer branch */
        $q_i \leftarrow R_{l(e_x)i}/\sum_j^m R_{l(e_x)j};$   /* $q_i$: probability of migrating to $i$ from $l(e_x)$ */
        Sample $i$ with probability $q_i$;
        $l(e_{k+1}) \leftarrow l(e_x)$ and $l(e_{k+2}) \leftarrow l_i;$
        $u_i \leftarrow u_i + 1;$
        $k \leftarrow k + 1;$
    end
end
$t_k \leftarrow |t_k - t_N|, \forall k \in [0, N]$ ;   /* Reverse time scale */
for $e \in E$ do
    $k \leftarrow k + 1$ ;            /* Make leaf vertex for each remaining lineage */
    Make vertex $v_k$ with $t(v_k) = 0$ and $p(v_k) = e;$
end
```

2.4 Likelihood

Here, we provide an expression for the likelihood $f(g|\tau, \sigma, \lambda)$, of a given 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 non-leaf vertex in the tree represents a migration 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 migration event and vertices $n, \ldots, 2n-1$ being the leaves. Let t_k be the time of the kth vertex. We assume that 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 Section 2.1 so that, for a given time t, $u(t) = (u_1(t), \ldots, u_m(t))$ is the occupancy status of locations at t, and $R(t)$ is the total rate of migration in the system at t. Note that $\sum_{i=1}^m u_i(t) = |E(t)|$ is the number of lineages in the tree at time t and that

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

The likelihood is thus

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

where t^- denotes the time immediately prior to t and $w : V \to \mathbb{R}$ is the function defined by

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

Note that the contribution to the likelihood (1) of concurrent leaf vertices is 1 since, if v_n and v_{n+1} are leaves, $t_{v_{n+1}} - t_{v_n} = 0$ and $w(v_n) = w(v_{n+1}) = 1$.

3 Parameter estimation

A Bayesian Markov chain Monte Carlo (MCMC) algorithm was used to sample the dispersal parameter σ, the overall migration rate τ and the competition parameter λ from their joint posterior distribution. The MCMC algorithm is also used to integrate over the unknown geographical locations of ancestral lineages. More specifically, the set of edges g is made of external branches g_e, for which the geographical locations are known, and internal edges, g_i, for which the geographical locations are not always known without uncertainty. We then sample alternatively from the distributions of the following random variables:

$$\lambda, \tau, \sigma | l(g_e), l(g_i),$$

and

$$l(g_i) | \lambda, \tau, \sigma, l(g_e),$$

where $l(g_e)$ and $l(g_i)$ are the sets of geographical locations on external and internal edges respectively. $l(g_i)$ is therefore considered as a parameter rather than data, which posterior distribution, along that of λ, σ and τ, is estimated using Gibbs sampling. The densities of the two variables above are proportional to the following product of densities:

$$f(l(g_i), l(g_e) | \lambda, \tau, \sigma) f(\lambda, \tau, \sigma),$$

where $f(l(g_i), l(g_e)|\lambda, \tau, \sigma) = f(g|\lambda, \tau, \sigma)$ is the likelihood of the augmented data (see Section 2.4) and $f(\lambda, \tau, \sigma)$ is the joint prior density of the migration model parameters. In the present study, the prior density is flat. Hence, in practice, sampling from the distribution of the two variables of interest relies on a Metroplis-Hastings step, with the ratio of target densities corresponding to the ratio of likelihoods.

Standard proposal densities were used to update the values of λ, σ and τ in the corresponding Metropolis-Hastings steps. For more information, we refer the reader to the many relevant textbooks on that topic. The data augmentation step, i.e., sampling from the distribution of $l(g_i)|\lambda, \tau, \sigma, l(g_e)$ is less conventional and warrants more explanations.

In order to update the value of $l(g_i)$, an internal node was first chosen uniformly at random. Let v be the selected node. The geographical locations of the edges below v were then updated as follows: if $l(c_1(v)) = l(p(v))$, select a new value for $l(c_2(v))$ by choosing uniformly amongst all the geographical locations found below $c_2(v)$. Otherwise, select a new value for $l(c_1(v))$ by choosing uniformly amongst all the geographical locations found below $c_1(v)$. v is then replaced by $c_2(v)$ in the first case or by $c_1(v)$ in the second case. The same updating scheme applies again and the algorithm stops when v is a tip. The Hastings ratio for this proposal is equal to 1 and the newly proposed ancestral geographical location are accepted or rejected using the standard Metropolis-Hastings rejection rule.

4 Results

A total of 1,000 simulations was performed, followed by model parameters estimation using MCMC. For each simulation, the number of geographic locations was uniformly drawn at random between 5 and 100, the value of τ between 0.1 and 10, the value of σ between 0.1 and 2.53 and the value of λ between 0.01 and 0.99.

4.1 Model parameters estimations

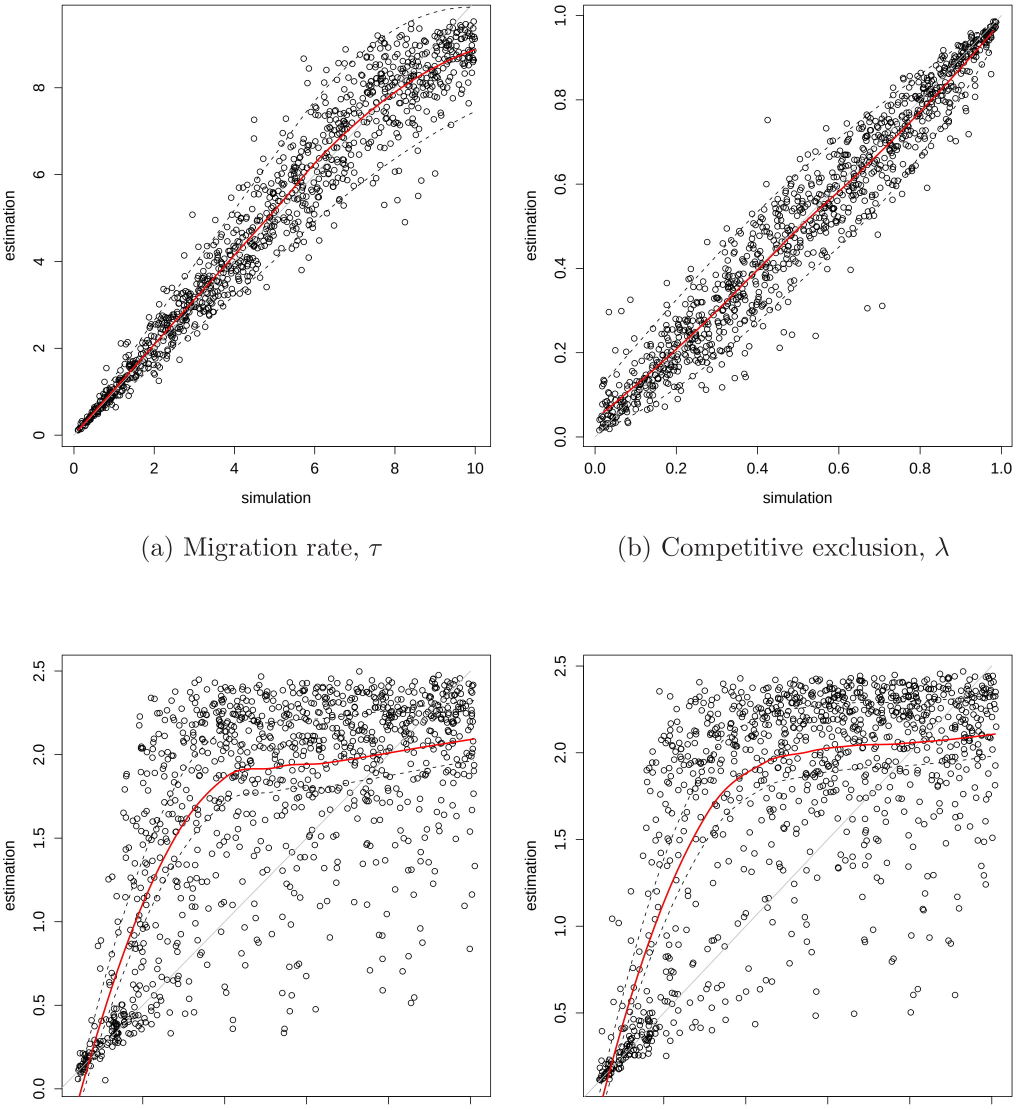
Both τ and λ parameters were correctly estimated (Figure 3a and 3b) with strong correlation between the estimated and true values of these parameters and relatively narrow credibility intervals. For small values of σ (i.e., $\sigma < 0.5$) in the simulated data, the estimates of this parameter show satisfactory accuracy. However, for larger values ($\sigma > 0.5$), estimation were poorer (Figure 3). Note that for $\sigma = 0.52$, the density of the truncated normal distribution is very similar to unifsvn statusorm $\mathcal{U}(0, 1)$ (80% of the area under the uniform curve overlaps with the truncated normal). Therefore, in such circumstances, the bias toward short dispersal distances is relatively weak, making estimation more difficult.

4.2 Testing limited dispersal

We next focused on using the proposed model to test for dispersal bias towards short distances. The posterior proportion of cases where σ belongs to the class of the mixture model for which dispersal is biased (see Methods) can be used to test the null hypothesis of no bias.

For $\sigma = 1.12$ 95% of the area below the truncated normal density curve overlaps with the uniform distribution (see Figure 4). We therefore first consider that data simulated with $\sigma > 1.12$ conform to the null hypothesis of no dispersal bias. After estimating the model parameters for each simulated data set, the hypothesis of no dispersal bias is rejected if the estimated proportion of MCMC samples for which $\sigma = 2.53$ is smaller than 0.5. The contingency table on the right of Figure 4 gives the proportion of true positives (TP), true negatives (TN), false positive (FP) and false negatives (FN). The specificity (i.e., TN / (TN + FP)) is here equal to 58% while the sensitivity (TP / (TP + FN)) is equal to 0.85.

For $\sigma = 0.52$, 80% of the area below the truncated normal density curve overlaps with



(a) Migration rate, τ

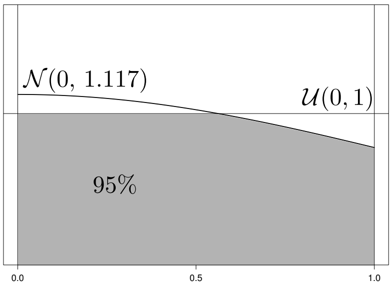
(b) Competitive exclusion, λ

(c) Latitudinal dispersal parameter, σ_l

(d) Longitudinal dispersal parameter, σ_L

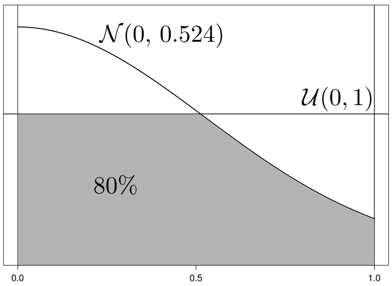
Figure 3: Model parameter values used for simulation plotted against estimated values obtained by MCMC. A polynomial regression fitting of the means of the posterior distributions is shown (red) on each plot as well as a regression fitting of the 5% and 95% quantiles (dotted lines).

the uniform distribution (see Figure 4). Using this threshold for defining the lack (i.e. $\sigma \geq 0.52$) or presence (i.e. $\sigma < 0.52$) of dispersal bias, the sensitivity and specificity reach 92% and 79% respectively.



Simulated	Estimated dispersal	
dispersal	*Limited*	*Uniform*
Limited	0.23	0.17
Uniform	0.09	0.51

Figure 4: Identification of limited dispersal from σ estimations. Non-uniform dispersal is considered to occur in the simulations when $\sigma \leq 1.117$. Non-uniform dispersal is estimated when more than 50% of the posterior values of σ are lesser than 2.53. The density of the dispersal kernel $\mathcal{N}(0,\ 1.117)$ is 80% similar to $\mathcal{U}(0,1)$ between 0 and 1.
Specificity=0.58 and sensitivity=0.85 (see text).



Simulated	Estimated dispersal	
dispersal	*Limited*	*Uniform*
Limited	0.15	0.01
Uniform	0.17	0.66

Figure 5: Identification of limited dispersal from σ estimations. Non-uniform dispersal is considered to occur in the simulations when $\sigma \leq 0.524$. Non-uniform dispersal is estimated when more than 50% of the posterior values of σ are lesser than 2.53. The density of the dispersal kernel $\mathcal{N}(0,\ 0.524)$ is 80% similar to $\mathcal{U}(0,1)$ between 0 and 1.
Specificity=0.92 and sensitivity=0.79 (see text).

4.3 Internal nodes geographic location estimation

The most probable locations at each internal node of the genealogy were recorded in each simulation and the Euclidean distance to the actual geographic location used in the simulation calculated (Figure 6). The internal nodes were grouped according to their heights in the tree and two null distributions of distances were constructed. The first null distribution was constructed by selecting uniformly at random the locations for the internal nodes of the tree among the whole set of locations. For the second null distribution, the locations at internal nodes were selected uniformly at random among compatible locations according to the locations of the tips in the clade rooted by the internal nodes of

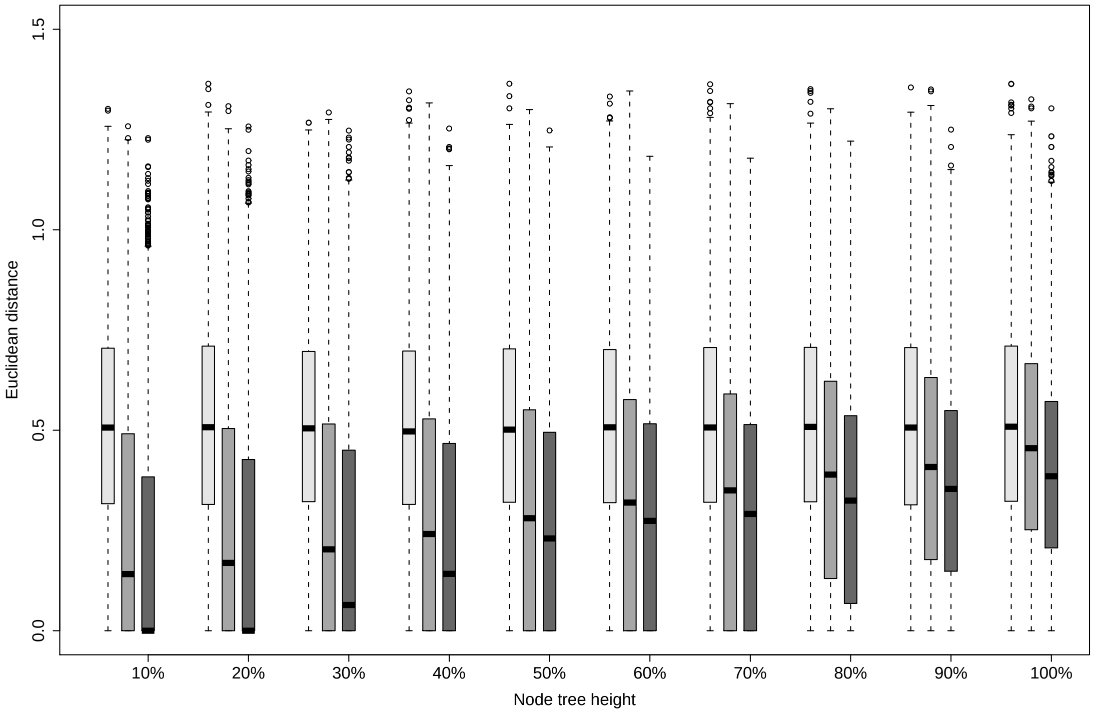


Figure 6: Euclidean distances between estimated node geographic location and true locations as defined in the simulations (dark grey). The distances were binned according to the heights of the nodes in the trees, expressed as the percentage of total height (x-axis). Additionally, two null distributions of distances are shown (light and middle grey). Boxes spread from the 25% to the 75% quantiles with the median represented as an horizontal thick segment. The vertical line represent the spread of the data, i.e. 1.5 times the inner quartile range, defined as the difference between the 75% quantile and the 25% quantile, added to the 75% quantile and subtracted from the 25% quantile with outliers shown as circles. As the node height of the tree increases the estimation of the node locations becomes more difficult.

interest. The results show that the accuracy of the estimation of internal node geographic locations, estimated by the geographic distance between estimated and simulated locations, decreases with the height of the internal nodes. However, these estimations remain more accurate than random assignment of the node locations. T-tests between null distributions of distances and the distances calculated from the estimated locations were all significant, showing that estimated locations are closer from the true node locations than randomly chosen locations.

4.4 Migrations

The proposed model can be use to retrace the succession of migrations that gave rise to the current geographic distribution of the taxa of interest. Each simulation corresponds to a sequence of migrations from which m distinct migrations can be identified. These migration were compared to the n most probable events from the posterior distribution of migration occurence. The n events were selected so that their cumulative posterior probabilty is greater than 0.95. The sensitivity, i.e., the estimated probability that a migration amongst the n most probable was also within the m 'true' events, reaches 88%. The specificity, i.e., the estimated probability that migration not identified as one of the n most probable was indeed amongst the m simulated events, reaches 79% (see Table 1).

Table 1: Identification of the migration events. The number of migration events are expressed as percentage of the total number of events over all simulations. The majority of events that are compatible with a given topology were not sampled. Specificity=0.79 and sensitivity=0.88.

	Estimations	
	Present	*Absent*
Simulated migrations	0.033	0.005
Non-simulated migrations	0.203	0.759

5 Discussion

This study introduces a phylogeographic model that accounts for the effects of dispersal and competitive exclusion. According to our approach, the geographic distribution of evolutionary units results from a sequence of migrations that can be mapped onto their phylogeny/genealogy. The rate at which migrations occur depends on the state of occupancy of the location to be colonised, making the migration events non-independent.

Numerical simulations on a two-dimensional landscape shows that the overall migration rate, the dispersal parameter and the competitive exclusion parameter can be accurately estimated using a Bayesian MCMC approach. Limited dispersal is succesfuly detected when the dispersal kernel departs from the uniform distribution, i.e., when migrations to nearby locations are generally preferred over more distant ones. While our results indicate that estimating the geographical locations of 'old' ancestral taxa is difficult, the accuracy of the estimates improve drastically as one considers 'younger' ancestral taxa. Moreover, our simulation results show that it is possible to correctly estimate the sequence of migration events that occured during the colonisation process of the landscape with high specificity and sensitivity.

The most original feature of our model lies in its ability to account for competitive exclusion. Our simulation results are encouraging as they demonstrate that it is possible to recover useful information about this ecological force from the analysis of geographical data in a phylogenetic framework. However, our approach relies on several assumptions and approximations which require careful scrutiny.

First and foremost, our model assumes that a given location is either free or occupied by an ancestor of one the sampled taxa. However, a location considered as free by the model could in fact be occupied by an ancestral taxa which did not leave any descendant in the sample. As a consequence, the proposed model will overestimate dispersal distances if competitive exclusion impacts on the colonization process. The consequence of such approximation on the estimation of the competitive exclusion parameter will depend on the proportion of locations that were considered as free by the model but were in fact occupied by an ancestral taxa competing with the ancestors of the sampled individuals.

Note however that the impact of competing ancestors of non-sampled taxa will only be problematic if their density varies across locations. A uniform density of these 'hidden' ancestors throughout geographic locations would affect every migration event to the same extent and would therefore not hamper the estimation the competitive exclusion parameter characterizing the sampled taxa.

Another approximation that impacts on measuring competitive exclusion lies the way the occupancy state of a location is defined. The proposed model only considers a on-off model where a given location is occupied or not. A more realistic treatment would model the probability of migrating to a given location as a function of the number of individuals on this location.

Our model also assumes that migrations occur exclusively at the internal nodes of the gene genealogy. While this is relevant to organisms that disperse from their parents' ecological range to avoid competiting for the same resources (e.g., seed dispersal in plants, acquisition of a new territory in birds), relaxing this assumption would allow dispersal to take place throughout the life of each organism. Relaxing this constraint is particularly relevant in cases where lineages correspond to populations or species as there is no obvious biological phenomenon that would prevent a population (or a species) to migrate at any point along the corresponding edge.

The structured coalescent [33] could provide an adequate framework to implement such model of migration. Indeed, according to this process, migrations can occur anywhere along the tree. However, the structured coalescent relies on the hypothesis that migrations events are independant from one another, and is therefore not compatible with the idea of competitive exclusion. Preliminary investigation on this topic (S.G. and D.W., not published) suggests that the structured coalescent could be modified to account for non-independant migrations. Calculating likelihoods under this new model would be more computationally demanding than for the plain vanilla structured coalescent. Further work needs to be done in order to assess whether such approach could be useful in practice.

In its current implementation, distances between geographic locations and their corresponding occupancy states both determine the locations found on internal edges as well as the heights of the internal nodes in the genealogy. A slightly different model could rely on using a more conventional model to calculate the joint density of node heights. Kingman's coalescent [22] would probably be the most obvious choice here. Such approach is utilised in Lemey et al. (2010) while not taking into account competition. According to this model, while migration/dispersal and genealogy are not independent, the genealogy is estimated from the sequences only and migration events do not affect the topology nor the node heights. Comparing the two models would then provide a statistical tool to assess whether geographic locations has had a significant impact on the timing of migration/coalescent events.

The statistical framework described in this study could easily be adapted to account for sources of data other than phylogeny and taxon geographic locations. It is for instance straightfoward to combine the present model with abiotic factors in a similar manner to that used in niche modelling [14]. Variables such as temperature or humidity for instance could indeed evolve along the tree according to relevant stochastic processes [49, 35] and be accounted for in the likelihood function.

6 Acknowledgements

The authors wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research (http://www.nesi.org.nz). New Zealand's national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSI's collaborator institutions and through the Ministry of Science and Innovation's Research Infrastructure programme.

References

[1] J. Avise. Phylogeography: retrospect and prospect. *Journal of Biogeography*, 36(1):3–15, 2009.

[2] J. Avise, J. Arnold, R. Ball, E. Bermingham, T. Lamb, J. Neigel, C. Reeb, and N. Saunders. Intraspecific phylogeography: the mitochondrial dna bridge between population genetics and systematics. *Annual review of ecology and systematics*, 18:489–522, 1987.

[3] F. J. Ayala. Competition between Species: Frequency Dependence. *Science*, 171(3973):820–824, 1971.

[4] M. Begon, C. R. Townsend, and J. L. Harper. *Ecology: From Individuals to Ecosystems*. Wiley-Blackwell, 2006.

[5] J. Brown and M. Lomolino. *Biogeography*. Sinauer Associates, 1998.

[6] J. Bullock and R. Clarke. Long distance seed dispersal by wind: measuring and modelling the tail of the curve. *Oecologia*, 124(4):506–521, 2000.

[7] S. Carlquist. Chance dispersal: Long-distance dispersal of organisms, widely accepted as a major cause of distribution patterns, poses challenging problems of analysis. *American Scientist*, 69(5):509–516, 1981.

[8] L. M. Chan, J. L. Brown, and A. D. Yoder. Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Molecular Phylogenetics and Evolution*, 59(2):523–537, 2011.

[9] J. Clark. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, 152(2):204–224, 1998.

[10] J. Clobert. *Dispersal*. Oxford University Press, 2001.

[11] L. Cook and M. Crisp. Directional asymmetry of long-distance dispersal and colonization could mislead reconstructions of biogeography. *Journal of Biogeography*, 32(5):741–754, 2005.

[12] K. Crawford and K. Whitney. Population genetic diversity influences colonization success. *Molecular Ecology*, 19(6):1253–1263, 2010.

[13] M. Crisp, S. Trewick, and L. Cook. Hypothesis testing in biogeography. *Trends in Ecology & Evolution*, 2010.

[14] J. Elith and J. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.

[15] G. Ewing, G. Nicholls, and A. Rodrigo. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations (MEPs). *Genetics*, 168(4):2407–2420, Dec 2004.

[16] A. G. Fredrickson and G. Stephanopoulos. Microbial competition. *Science*, 213(4511):972, 1981.

[17] G. Gause. Experimental studies on the struggle for existence: 1. Mixed population of two species of yeast. *Journal of Experimental Biology*, 9(4):389–402, 1932.

[18] R. D. Gray, A. J. Drummond, and S. J. Greenhill. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science*, 323(5913):479–483, 2009.

[19] G. Hardin. The Competitive Exclusion Principle. *Science*, 131(3409):1292–1297, 1960.

[20] M. Hickerson, B. Carstens, J. Cavender-Bares, K. Crandall, C. Graham, J. Johnson, L. Rissler, P. Victoriano, and A. Yoder. Phylogeographys past, present, and future: 10 years after. *Molecular Phylogenetics and Evolution*, 54(1):291–301, 2010.

[21] S. Hubbell. *The unified neutral theory of biodiversity and biogeography*, volume 32. Princeton Univ Dept of Art &, 2001.

[22] J. F. C. Kingman. The Coalescent. *Stochastic Processes and their Applications*, 1982.

[23] L. L. Knowles. Statistical Phylogeography. *Annual Review of Ecology, Evolution, and Systematics*, 40:593–612, 2009.

[24] U. Kodandaramaiah. Use of dispersal–vicariance analysis in biogeography–a critique. *Journal of Biogeography*, 37(1):3–11, 2010.

[25] K. Lamm and B. Redelings. Reconstructing ancestral ranges in historical biogeography: properties and prospects. *Journal of Systematics and Evolution*, 47(5):369–382, 2009.

[26] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard. Bayesian Phylogeography Finds Its Roots. *PLoS Computational Biology*, 5(9):e1000520, 2009.

[27] P. Lemey, A. Rambaut, J. J. Welch, and M. A. Suchard. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, 2010.

[28] R. H. MacArthur and E. O. Wilson. *The Theory of Island Biogeography*. Princeton University Press, Princeton, 1967.

[29] G. MacDonald. *Biogeography: Introduction to Space, Time, and Life*. Wiley, illustrated edition, 2003.

[30] S. Manel, M. Schwartz, G. Luikart, and P. Taberlet. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4):189–197, 2003.

[31] R. McDowall. A conceptual basis for biogeography. *New Zealand Freshwater Fishes*, pages 87–103, 2010.

[32] R. Nielsen and M. Beaumont. Statistical inferences in phylogeography. *Molecular Ecology*, 18(6):1034–1047, 2009.

[33] M. Nordborg. *Coalescent Theory*, pages 843–877. John Wiley & Sons Ltd, 2008.

[34] A. Ōkubo and S. Levin. *Diffusion and ecological problems: modern perspectives*, volume 14. Springer Verlag, 2001.

[35] P. Pearman, A. Guisan, O. Broennimann, and C. Randin. Niche dynamics in space and time. *Trends in Ecology & Evolution*, 23(3):149–158, 2008.

[36] A. Phillimore and T. Price. Density-dependent cladogenesis in birds. *PLoS Biology*, 6(3):e71, 2008.

[37] A. Pigot, A. Phillimore, I. Owens, and C. Orme. The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Systematic biology*, 59(6):660, 2010.

[38] R. Ree, B. Moore, C. Webb, and M. Donoghue. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*, 59(11):2299–2311, 2005.

[39] B. Riddle. What is modern biogeography without phylogeography? *Journal of biogeography*, 36(1):1–2, 2009.

[40] F. Ronquist. Reconstructing the history of host-parasite associations using generalised parsimony. *Cladistics*, 11(1):73–89, 1996.

[41] F. Ronquist and I. Sanmartín. Phylogenetic methods in biogeography. *Annual Review of Ecology, Evolution, and Systematics*, (0), 2011.

[42] F. Smith, S. Lyons, S. Ernest, and J. Brown. Macroecology: more than the division of food and space among species on continents. *Progress in Physical Geography*, 32(2):115–138, 2008.

[43] R. R. Sokal and F. J. Rohlf. *Biometry.* W. H. Freeman and Co., New York, 1995.

[44] R. Walker and L. Ribeiro. Bayesian phylogeography of the arawak expansion in lowland south america. *Proceedings of the Royal Society B: Biological Sciences*, 278(1718):2562–2567, 2011.

[45] I. Wang. Recognizing the temporal distinctions between landscape genetics and phylogeography. *Molecular Ecology*, 19(13):2605–2608, 2010.

[46] J. M. Waters. Competitive exclusion: phylogeography's 'elephant in the room'? *Molecular Ecology*, 20(21):4388–4394, 2011.

[47] D. Wethey. Biogeography, competition, and microclimate: the barnacle chthamalus fragilis in new england. *Integrative and Comparative Biology*, 42(4):872–880, 2002.

[48] J. J. Wiens. The niche, biogeography and species interactions. *Philosophical Transactions of the Royal Society B*, 366:2336–2350, 2011.

[49] C. Yesson and A. Culham. Phyloclimatic modeling: combining phylogenetics and bioclimatic modeling. *Systematic Biology*, 55(5):785–802, 2006.