What fruits can we get from this tree?

Laudanno, Giovanni

DOI:
10.33612/diss.155031292

Chapter 6

Synthesis

Synthesis

Guybrush: At least I've learnt something from all of this.
Elaine: What's that?
Guybrush: Never pay more than 20 bucks for a computer game.
– The Secret of Monkey Island

6.1 Modelling biology

Biology, from a broad perspective, is the study of life. Unsurprisingly, the description of what life is entails a high degree of complexity. Our best chance to overcome such complexity is to resort to the scientific method, which is the best tool we have developed so far to collect knowledge about the world. According to its dictates, hypotheses are tested with experiments to add new bricks to an ordered structure of knowledge (or to remove some, if data suggests so). One ordered way we can read biology is by means of nested compartments (Simon, 1991). A bottom-up description would start from very small components such as carbon-based molecules like DNA and RNA. Genetic macro-molecules contain heritable characters that compose the instruction manual of life. Through those the basic blocks of organisms can be built: the cells. Each cell not only contains genetic material but can also perform a variety of tasks, including response to stress, communication with other cells, or production of energy. Ultimately, the complex functioning of each cell aims to maintain homeostasis and reproduce. In complex organisms cells can specialize to perform different tasks. In animals, for example, bone cells are completely different from skin cells that, in turn, are completely different from lung cells et cetera. Some of these cells are specialized for reproduction.

Through reproduction an individual can pass on its genetic material even after its death. The odds of this occurring depend on a variety of factors, like how the

organism interacts with the environment, as well as individuals of its own and other species. This is the essence of natural selection.

In a nutshell, this allows one to see how different compartments of biology can create a description of life on different scales: from genetics, to cell biology, to ecology, all the way to the evolutionary perspective. There are, of course, more details contained across this continuum, but they are not essential for what we want to present here. The epistemological approaches to each of these compartments focus on different elements and adopt different strategies. The study of each compartment has to take into account phenomena occurring at the other levels. So, for studying how the cell works it is essential to understand how genes work in the nuclei of the cells, what proteins they code for, how these proteins affect the cell and so forth. Likewise the physiology of an organism is dictated by the behaviour of cells, which in turn take their cues from environmental signals– individual cells sense stimuli from the environment and propagate signals to other cells to form a collective behavior.

Despite the fact that the description at each level depends on the other levels, the details of other levels can often be simplified as to reduce the burden of unnecessary complexity.

The very first sentence in this thesis is "Why don't you use an individual-based model?". The answer falls somewhat in the need of avoiding unnecessary complexity for the level we are working at. When studying the behaviour of a gas we know that the gas itself is composed by molecules and that each of those follows the laws of dynamics. However no physicist would ever solve a system of 6.022×10^{23} differential equations. The pragmatical approach, instead, is to define macroscopic variables, such as temperature and pressure, to describe the system as a whole. The study of these quantities, the thermodynamics, must take into account the underlying particle dynamics but follows a different paradigm: it is a matter of scale. Likewise, the field of macro-evolution, i.e., the study of how organisms evolve beyond the single species level, cannot overlook important concepts coming from ecology. At the same time, however, it needs to be formalized in a different way. The very concept of speciation rate is quite blurred. It is a major simplification that assumes, for example, that all speciation events occurring in a phylogeny could be considered as belonging to a homogeneous category. We know that the factors that drive each speciation event could, in principle, be completely different from other ones in the same phylogeny. However, the extremely long times spanned by a macro-evolutionary process and the high number of individuals involved in the process across the many generations demand us to approach the problem from a simplified perspective. This allows us

to obtain information that would be otherwise inaccessible, such as an estimate of how many species went extinct during the diversification process or how traits could influence the process.

6.2 Different paradigms

The focal point of the paradigm discussed so far is to assume that speciation and extinction events could be seen as the birth and death events of a simplified birth-death (BD) process. There is a great benefit obtained in doing so, as this formalism is well established and has been already largely and successfully employed in many other fields of science such as physics, chemistry, economy et cetera. What we pay in terms of oversimplification in the description of phenomena is regained in terms of a greater depth of understanding.

One of the sentences that a theoretician hears more often is "All models are wrong, but some are useful". This is because building a model always entails some degree of compromise. In the words of Levins (1966) there are three main characteristics that one should seek while building a model: generality, realism and precision. Having them all would be the best scenario but it is usually impossible, hence the necessity of a compromise. Levins, 1966 identifies three classes of biological models, according to what characteristics one chooses to sacrifice. When realism is sacrificed we incur in models usually based on set of differential equations, as in the famous case of the Lotka-Volterra model or in the case of the models in this thesis. If a model is tailored on a certain phenomenon (e.g., a specific species or a specific set of genes) and constrained in a specific setting (e.g., limited time frame, specific habitats) its predictions can become more reliable but limited in their space of applicability. The last option is to focus on realism and generality sacrificing precision. The idea behind is to create models able to detect macro-trends (e.g., identifying monotonic trends of some functions or presence/absence of some signal).

The idea is that a single model alone is typically unable to describe the complexity of biological phenomena but the combination of many models can help overcome the shortcomings of each of them when taken separately.

6.3 Inference limitations in Birth Death models

For most of this thesis we rely on the framework originally introduced by Nee, May, and Harvey (1994) to assign a likelihood to a phylogenetic tree based

on the BD paradigm. The fundamental idea of the BD description is to describe the process through the dynamics of the distribution for the number of species in the phylogeny. This can be obtained by solving a system of ODEs (the P-equation system) and its solution is a geometrical distribution with a modified zero term (Kendall, 1948b). The fundamental idea by Nee, May, and Harvey (1994) is to use this solution to assign a probability to each part of the phylogeny (we do so by "breaking the tree"). The joint probability distribution is the likelihood of the entire phylogeny.

Recently, Louca and Pennell (2020) mathematically proved that for BD models, time-calibrated phylogenies do not contain enough information to uniquely determine speciation and extinction rates in case of constant or time-dependent rates. They prove that all such models can be grouped in congruence classes characterized by the same pulled diversification rate $r_p(t) = \lambda(t) - \mu(t) + \frac{1}{\lambda(t)}\frac{d\lambda(t)}{dt}$. Inference techniques are unable, unless provided by additional information (e.g., additional fossil data) to discriminate models in the same congruence class. However, such limitations seem, at the current stage, not to affect trait-based models (as in the SSE approach, e.g., Maddison, Midford, and Otto, 2007; FitzJohn, 2010; FitzJohn, 2012; Goldberg and Igić, 2012; Goldberg, Lancaster, and Ree, 2011; Herrera-Alsina, Els, and Etienne, 2019), diversity-dependent models (as the one presented in chapter 3) nor rate shift models (as the one presented in chapter 4). In general, the history of science holds many examples in which two or more models performed equivalently at explaining the data. The possibility of discriminating between models inevitably depends on the quality of the available data. There have been cases in which the controversies have been resolved as new and more accurate data became available. One well-known example is the case of development of the theory of general relativity. General relativity, in fact, has been developed from sheer theoretical foundations by Albert Einstein in 1915. Until the observation of the precession of Mercury's orbit in 1919, Newton's and Einstein's theories of gravity were equally capable of explaining the available data. Einstein's theory was in a better agreement than Newton's with the electromagnetic theory, but this alone cannot constitute a definitive proof without experimental evidence. If this criterion was enough, in fact, we should also conclude that either general relativity or quantum mechanics must be wrong, being the two theories not compatible with each other. Despite the fact that BD-based models can still prove to be a great resource to infer information from phylogenies, the result from Louca and Pennell (2020) highlights that many models can deliver the same result and an arbitrary choice of a model to use can produce results that are not actually strictly derived from the data. Empiricists should, therefore, interpret results com-

ing from these models cautiously keeping in mind that they cannot overcome the problem of limited information present in the data.

6.3.1 Specific limitations of the Q-framework

The Q-framework variant of the BD models proved itself to be quite powerful. In particular, it can be used to describe processes for which the standard P-framework cannot be used. One notable example is the MBD model of chapter 3, which entails very complex interactions between species. First of all, unlike the P-approach, the ODE set is usually impossible to integrate analytically. For large systems the task might become impossible due to exceedingly long computation times. This was the case, for example, for the calculation of the conditional probability of the MBD model. The `mbd` R package, in fact, features functions to calculate such probabilities by integrating the Q-system. However we ended up not using them for calculating the likelihood because of their relative slowness when compared to the other methods. They were, however, very useful for testing the consistency of the results obtained with the other methods. We performed similar consistency tests also for the likelihood presented in chapter 4. Another limitation is that the dependence on phylogenetic diversity (i.e., the dependence on the phylogenetic branch length) cannot be modelled through the Q-approach. This is due to the fact that phylogenetic diversity (Faith, 1992a; Faith, 1992b), defined according to all the pairwise distances between species across the phylogeny, is intrinsically topology-dependent and the Q-framework always assumes topologies to be equally probable (see chapter 2). Hence a complete description is impossible (at least at the current stage).

Another possible application of the Q-framework is for a model with multiple locations and diversity-dependence within each. Unfortunately, as already mentioned earlier, calculations can quickly become too cumbersome (as in the case of Xu and Etienne, 2018, where the authors could resort only to a simulated approach). In fact, already for a model taking into account 2 locations (say, A and B with AB meaning contemporary presence in both), the variable $Q^{k_A,k_B,k_{AB}}_{m_A,m_B,m_{AB}}$ would require a three-dimensional ODE system, which can be computationally demanding. Furthermore, for any combination of the (m_A, m_B, m_{AB}) variables not only standard (sympatric) speciation and extinction events should be considered, but also contributions related to contractions (state changes such as $(m_A, m_B, m_{AB}) \to (m_A + 1, m_B, m_{AB} - 1)$ or $(m_A, m_B, m_{AB}) \to (m_A, m_B + 1, m_{AB} - 1)$), allopatric speciation events ($(m_A, m_B, m_{AB}) \to (m_A + 1, m_B + 1, m_{AB} - 1)$) or migration events ($(m_A, m_B, m_{AB}) \to (m_A - 1, m_B, m_{AB} + 1)$). The implementation of more locations

would make the model even more intractable (e.g., a system with 3 locations would already scale the system dimensionality up to 7). We cannot exclude that it could possible to find a clever way of adapting the framework to a many locations system, but it would certainly demand some particular attention to overcome the computational challenges.

Another interesting question to ask, that we did not explore in chapter 2, is whether the Q-framework could be expanded to also include trait-based diversification. Again, a brute force approach would probably increase the number of required equations to an intractable level. We tried to explore it to some extent (using the formalism presented in Eq. 6.4.4), showing some interesting analogies between the Q-framework and the formalism usually used in trait models (SSE). Again, we do not exclude the possibility that this could be done and it certainly is something that we would like to study more in the future.

6.4 Future prospects

The BD approach is the backbone of chapters 2, 3 and 4. The same chapters mention/use also the Q-framework, which is another way to adopt the birth-death paradigm and it is useful to describe diversification process where rates depend on the number of species present. Chapter 5, rather than presenting a new model, presents a tool to estimate whether the introduction of a new model is needed, in terms of the necessity of the implementation of its likelihood formula in a broader Bayesian framework.

As the goal of each chapter is to introduce new models/tools, rather than their applications, I will discuss the potential applications of such models/tools and what benefits they can bring.

6.4.1 Applications of chapter 2

In this chapter we proved that the Q-framework yields results in agreement with other models in the literature. This applies to any time-dependent rate model as well as any implementation of $\rho-$sampling (i.e., where the observed extant species are assumed to be a fraction ρ of the total number of species) and $n-$sampling schemes (where n additional species are not reported in the phylogeny). This provides strong support for the correctness of the framework. The consequence is that the set of Q-framework's ODEs can be used to create new models. This allows building models where keeping track of the number of unseen species (e.g., species going extinct before the present) along the process is crucial, as in

(but not only) the diversity-dependent diversification model. One example of this is the MBD model presented in chapter 3. However, many other applications are, in principle, possible. This refers to all the possible cases in which the breaking-the-tree hypothesis (as in Nee, May, and Harvey, 1994) does not hold. In other words, this can be helpful to build models in which branches in the phylogeny are interacting with each other in some way, of which one notable example is DAISIE (Valente, Phillimore, and Etienne, 2015).

In chapter 2, to find the analytical solution, a pivotal step has been to identify the factor $c(z,t) = (\mu(t) - z\lambda(t))(1-z)$. This factor appears in the equation 2.5.5 for the generating function $G(z,s,t) = \sum_n z^n P_n(s,t)$ for the P-framework (a function that summarizes the distribution $P_n(s,t)$ of number of species in a process starting at time s and ending at time t)

$$\frac{\partial G(z,s,t)}{\partial t} = c(z,t)\frac{\partial G(z,s,t)}{\partial z}. \tag{6.4.1}$$

The factor $c(z,t)$ also occurs in the equation 2.5.2 for the generating function $F_k(z,t) = \sum_m z^m Q_m^k(t)$ of the Q-framework (using constant rates)

$$\frac{\partial F_k(z,t)}{\partial t} = c(z,t)\frac{\partial F_k(z,t)}{\partial z} + k\frac{\partial c(z,t)}{\partial z}F_k(z,t). \tag{6.4.2}$$

Interestingly enough, this factor also appears in the core equations for the SSE models (see, for example, equations (10) and (11) in Rabosky, 2014)

$$\begin{aligned}\frac{dE(t)}{dt} &= c(E,t)\\ \frac{dD(t)}{dt} &= \frac{\partial c(E,t)}{\partial E}D.\end{aligned} \tag{6.4.3}$$

We can combine these equations to create another differential equation for the variable $\psi_k(t) = E(t)D^k(t)$

$$\begin{aligned}\frac{d\psi_k}{dt} &= \frac{dE}{dt}D^k + kD^{k-1}\frac{dD}{dt}E\\ &= c(E,t)D^k + kD^{k-1}\frac{\partial c(E,t)}{\partial E}DE\\ &= c(E,t)\frac{\partial \psi_k}{\partial E} + k\frac{\partial c(E,t)}{\partial E}\psi_k,\end{aligned} \tag{6.4.4}$$

which looks very similar in shape to 6.4.2. This suggests a formal link between the Q-framework and the SSE-framework and it may, possibly, pave a way towards a

trait-driven (as in SSE) diversity-dependent model. Unfortunately I did not have the time to explore more in this direction. However, it is indeed interesting to see how the two approaches, built from two totally different starting points, are consistent to each other. In particular, it is worth noting that the interpretation for the factor of 2 in $\partial_z c(z,t) = 2z\lambda - \lambda - \mu$ originally given by Maddison, Midford, and Otto (2007) (see panels *c* and *d* of Fig. 2 in the original article) provides an explanation for the same factor appearing in the Q-equation from Etienne et al. (2012) (see Eq.1.3.6).

We also tried to analytically solve the system of differential equations for the model in the case of diversity-dependent rates. Despite the fact that this might look like a small change with respect to the system with time-dependent rates, this actually makes things much harder to deal with. The issue becomes evident when transforming the infinite ODE system into a single partial differential equation (PDE) for the generating function. In fact, when dealing with constant or time-dependent rates this features only first order derivatives. Unfortunately this is no longer true in the case of diversity-dependent rates $\lambda_n = an + b$ and $\mu_n = \mu$, for which the transformed equation is

$$\frac{\partial F_k}{\partial t} = \alpha(z)\frac{\partial^2 F_k}{\partial z^2} + \beta_k(z)\frac{\partial F_k}{\partial z} + \gamma_k(z)F_k \tag{6.4.5}$$

where

$$\begin{aligned}
\alpha(z) &= az^2(z-1) \\
\beta_k(z) &= [(3ak+a+b)z^2 - (2ak+a+b+\mu)z + \mu] \\
\gamma_k(z) &= k[2(ak+b) - (ak+b+\mu)].
\end{aligned} \tag{6.4.6}$$

In the case of eq. 6.4.3 we exploited the method of characteristics to find the analytical solution for the system. For its diversity-dependent version, eq. 6.4.5, its application resulted to be not as simple.

6.4.2 Applications of chapter 3

This chapter presents the building of a novel model, namely the Multiple-Birth Death model (MBD). The MBD model presents one of the possible applications of the Q-framework formalized in chapter 2. The framework was indeed necessary because the model accounts for the possibility of an environmentally-driven large scale event whose effects depend on the current number of species. In fact such an event, when triggered, can induce a speciation on each of the lineages currently

present in the phylogeny. The rationale for its introduction is to build in the effects of a species pump mechanism (Jetz, Rahbek, and Colwell, 2004) into the model. The model is, in fact, specifically tailored to analyze phylogenies where the pace of evolution is so high that standard models struggle to describe it.

There are two possible alternative strategies to model diversification driven by a species pump. The first alternative strategy (*A*) is to use a specific model where rates are time-dependent. In such a scenario having rates with localized peaks in time could induce the effects of rapid bursts in speciation. A second strategy (*B*) is to implement two different sets of rates: one for the standard diversification regime and one to describe the intense bursts of speciation events. The second set of rates would, in this context, act only in very short time windows.

Both approaches have some disadvantages. Alternative model *A* cannot reproduce one main characteristic of the species pump mechanism that is instead captured by the MBD model: the impact of such an event should become more powerful as more and more species populate the phylogeny. As a consequence, diversification in the MBD model is intrinsically diversity-dependent, because the effects of the large-scale event are influenced by the current number of species in a non-linear way. The same effect cannot be mimicked by any time-dependent model, although one can come close (Pannetier et al., 2020). In traditional diversity-dependence models, the probability of speciation decreases as species accumulate, leading to a lineages-through-time profile that flattens around the carrying capacity value. In the MBD model this mechanism is reversed: as more species accumulate, the effects of the large-scale events become greater and greater.

To implement alternative model *B* it would be necessary to define the mechanisms that activate and deactivate the regime of enhanced diversification in the clade. The easiest way to do so is to define a system with two states: one regular and one ephemeral with enhanced speciation potential. As in BiSSE (Maddison, Midford, and Otto, 2007), this would require two rates for the state-change (such as the $q_{0,1}$ and $q_{1,0}$ rates in BiSSE): one for the initiation and one to return to the standard regime. To make sure that only rapid bursts of speciation are allowed, the latter must be much higher than the former. Unlike BiSSE, such rates would not affect only one lineage at the time, but the entire clade. This is due to the fact that the state change must reflect the impact of a changing landscape on all species.

mimic the effect of an external environmental condition. In addition to these two rates an enhanced speciation rate would be needed, leading to a model with three additional parameters, compared to the standard BD model. The MBD model instead only adds two additional rates (labelled as ν and q in chapter 3). As a general rule, whenever possible, a model with less complexity should always be

preferred to more complex alternatives. The rate $q_{0,1}$ bears some resemblance to rate ν in the MBD model, as both trigger speciation events. However, in alternative B the rate is a per-species rate and hence the speciation events induced by the transition from state 0 to 1 will all take place in the same subclade, whereas the speciation events in the MBD model take place across the entire phylogeny. Furthermore, speciation events will continue to occur in alternative B until the state changes back to 0, whereas in the MBD model only one burst of speciation events will take place.

Despite the fact that the MBD likelihood inference proved to be effective on simulated phylogenies (see Fig. 3.3), one main issue is that empirical phylogenies with aligned speciation events are not currently available. This is due to the fact that currently implemented tree priors in Bayesian phylogenetic tools, such as BEAST2 (Bouckaert et al., 2019), MrBayes (Huelsenbeck and Ronquist, 2001) or RevBayes (Höhna et al., 2016), do not feature multiple simultaneous speciation events. However, BEAST2 allows for the implementation of novel third party tree priors, so in principle it would be possible to develop and implement a new MBD prior. Understanding whether such work is needed is one of the main reasons that drove us to develop `pirouette` (see chapter 5). In chapter 3 instead we took another approach. We developed a metric aimed to detect the distinctive features of an MBD phylogeny. Because the greatest difference between MBD and BD phylogenies is the presence of aligned events, we needed a metric able to measure the extent of the clustering of branching times. To do so we introduced the Distance from the Nearest Branching Time (DNBT) metric. We successfully tested its effectiveness on datasets of simulated trees. Then, we used it to extract the signal from the empirical phylogeny of lake Tanganyika endemic cichlids, detecting a strong signal. At this point, the natural next step would be to develop a method to apply the MBD likelihood inference to these phylogenies. We started to develop an alternative approach that does not require any BEAST2 implementation. The idea of the approach is based on calculating the likelihood of the alignment given the MBD parameters $P(DNA|\theta_{MBD})$, obtained by marginalizing over the space of trees:

$$\begin{aligned} P(DNA|\theta_{MBD}) &= \sum_{T_i \in T} P(DNA|T_i) P(T_i|\theta_{MBD}) \\ &\sim \sum_{T_j \in T_{MBD}} P(DNA|T_j), \end{aligned} \tag{6.4.7}$$

where $P(DNA|T_i)$ is the tree likelihood calculated according to the Felsenstein's algorithm (Felsenstein, 1973) for a particular substitution model (e.g., JC69) and

$P(T_i|\theta_{MBD})$ is the MBD likelihood. A way to do it (as in the second line of eq. 6.4.7) is to sum over a representative sample of the MBD tree space generated by those parameters. This sample can be generated by simulating the MBD process, as described in chapter 3. With the likelihood $P(DNA|\theta_{MBD})$ it would be possible to infer the MBD parameters directly from the alignment data. We have not explored this possibility fully yet, because calculating this sum by Monte Carlo sampling is computationally demanding. One of the challenges is that producing phylogenies with exactly the observed number of species is not trivial. For this we suggest to use the combined backward- and forward approach used by Etienne et al. (2012) to compute the expected number of lineages conditional on the phylogeny. Another challenge is that the topology generated by a simulated MBD process is rarely compatible with the observed DNA sequence alignments. We can overcome this problem by taking topologies from a BEAST2 or RaxML (Kozlov et al., 2019) analysis, and combining these with the branching times of the MBD simulation, because all topologies are equally likely under the MBD process, and hence the topology of the phylogeny does not contain any information on the MBD process. Combining these ideas should be the next step of the project and it would probably make the entire model finally available to empiricists.

6.4.3 Applications of chapter 4

In this chapter we mathematically prove that some of the current models involving single-lineage rate shifts lead to an incorrect likelihood. We developed the correct likelihood formula for such cases. When a lineage undergoes a single-lineage rate shift, its diversification starts to occur at a different pace than other lineages in the phylogeny. This occurs when such species obtain a competitive advantage with respect to its competitors, due to the extinction of one or more antagonists, a new environment becoming available or for the development of a key innovation (Heard and Hauser, 1995; Etienne and Haegeman, 2012). The regime shift could potentially occur either on an observable lineage (i.e., surviving to the present) or an unobservable one (i.e., going extinct before the present). We provided likelihood formulas for both cases and proved how the combination of the two for a dummy shift (where new rates are equal to the old ones) yields the traditional Nee, May, and Harvey (1994) likelihood. Our formula works not only for constant rates, but also for time-dependent and for diversity-dependent rates. Furthermore we expanded the likelihood formula for the case in which multiple single-lineage shifts are present in the phylogeny.

The first consequence of our work is that we provided a solution to the debate

expressed by Moore et al. (2016) about the accuracy of estimations in BAMM (Rabosky, 2014). They identify two major problems: (1) the choice of a Coalescent Poisson Process prior distribution for diversification parameters makes the inference extremely sensitive to the prior and (2) the implemented likelihood is fundamentally incorrect. In our work we address the latter. In particular they show that the bias in the likelihood calculation is due an incorrect way of accounting for shifts on extinct lineages. They also propose a numerical solution to approximate the correct solution for the likelihood using Monte Carlo simulations. They use them to expose the issue with likelihood calculation but the method is not suitable for normal use in BAMM as it is too computationally intensive.

Our likelihood formula for unobserved rate shifts 4.2.22 provides an analytical solution to this problem, thus much faster than the one proposed by Moore and colleagues. Furthermore, with the likelihood formula being correct, now it is possible to perform hypothesis testing by comparing the marginal likelihoods (which was another critical aspect of BAMM exposed by Moore et al.).

Our likelihood formula can be implemented not only in BAMM, but also in MEDUSA (Alfaro et al., 2009) and other related multi-shift methods.

Another consequence is that the diversity-dependent model for key innovations (Etienne and Haegeman, 2012) now correctly calculates the likelihood. This has been already implemented in the R package DDD (Etienne and Haegeman, 2020).

As in chapter 2, the consequences of our work are not only limited to current models but will apply to any future model that involves lineage-specific rate shifts.

Quite a few papers have used the incorrect likelihood on empirical phylogenies, including papers of my co-authors (Etienne and Haegeman, 2012; Rabosky, 2014). The analyses in these papers should ideally be repeated to check whether the conclusions they draw are still valid.

6.4.4 Applications of chapter 5

The R package `pirouette` has been actually developed to provide a generalized tool to realize two other projects. These projects have not been finished yet. The goal of the first one was to establish whether the likelihood for the Protracted Birth-Death (PBD) model was needed to be implemented as tree prior in BEAST2 (Drummond and Rambaut, 2007; Bouckaert et al., 2019).

The PBD model (Rosindell et al., 2010; Etienne and Rosindell, 2012; Lambert, Morlon, and Etienne, 2015), as the name suggests, is a model where speciation is not instantaneous. This is an approximation used in many other models

but it does not reflect how the speciation process actually occurs. Within the PBD framework, there are two stages to go through before realizing a proper speciation: a first event, called speciation-initiation, produces an incipient species; later on, a second stochastic event, the speciation-completion, transforms an incipient species into an actual species. The model has been initially proposed to provide an explanation to the observed phenomenon of the pull-of-the-present, which is the pull that can be observed in the final part of a lineages-through-time plot.

The second project was to perform a `pirouette` analysis for the MBD model, which is extensively explained in chapter 3. In that chapter we use the DNBT statistics to provide a similar answer to the same question. Despite having obtained some results for the MBD model using the `pirouette` approach, I decided not to include them as a chapter of this thesis because I had concerns about the correctness of the implementation as well as doubts on the effective capacity of the standard metric in `pirouette` (the nLTT statistic, introduced by Janzen, Höhna, and Etienne (2015)) to detect the major MBD characteristics that cannot be captured by the BD model. We believe that the implementation of the DNBT statistics could be very promising in achieving this goal.

Apart from PBD and MBD, `pirouette` could be used to perform the same analysis for several other similar models. One straightforward application would be for models for which is relatively easy to write simulation routines but whose likelihood would be very hard to develop. The simplest way to picture that is by increasing the complexity of current models adding an additional element, e.g., by letting carrying capacities in diversity-dependent models depend on territory ontogeny (Valente, Etienne, and Phillimore, 2014). Another possibility could be to evaluate the performance of current BEAST2's tree priors on phylogenies simulated according to individual based models. One could use a phylogeny obtained starting from spatially explicit models, where speciation and extinction are defined as local events (which could be subject to local diversity-dependence in terms of local carrying capacities, as in Herrera-Alsina et al. (2018)). Alternatively it is possible to build phylogenies from individual based models accounting for complex interactions between individuals' genotypes, phenotypes and the environment (Aguilée et al., 2018; Rangel et al., 2018). Besides individual-based models, also combinations of current models can be explored. For example one could try to add to current models some dependency on traits, as in the SSE models. Likelihood functions for such models would be extremely difficult to develop but, in principle, a `pirouette` analysis could be used to prove that implementing tree priors for these models are not really needed (or, conversely, that they are).

In summary, I hope that I have contributed both new insights and new tools

that can further our understanding of patterns of macroevolutionary diversification and the mechanisms that drive them.

Bibliography

Aguilée, R, A Lambert, and D Claessen (2011). "Ecological speciation in dynamic landscapes". In: *Journal of evolutionary biology* 24.12, pp. 2663–2677.

Aguilée, Robin, David Claessen, and Amaury Lambert (2013). "Adaptive radiation driven by the interplay of eco-evolutionary and landscape dynamics". In: *Evolution* 67.5, pp. 1291–1306.

Aguilée, Robin et al. (2018). "Clade diversification dynamics and the biotic and abiotic controls of speciation and extinction rates". In: *Nature communications* 9.1, pp. 1–13.

Akaike, Hirotogu (1998). "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of hirotugu akaike*. Springer, pp. 199–213.

Alfaro, Michael E et al. (2009). "Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates". In: *Proceedings of the National Academy of Sciences* 106.32, pp. 13410–13414.

Alin, Simone R and Andrew S Cohen (2003). "Lake-level history of Lake Tanganyika, East Africa, for the past 2500 years based on ostracode-inferred water-depth reconstruction". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 199.1-2, pp. 31–49.

Allaire, JJ et al. (2017). *rmarkdown: Dynamic Documents for R*. R package version 1.8. URL: `https://CRAN.R-project.org/package=rmarkdown`.

Allender, Charlotte J et al. (2003). "Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration". In: *Proceedings of the National Academy of Sciences* 100.24, pp. 14074–14079.

Bache, Stefan Milton and Hadley Wickham (2014). *magrittr: A Forward-Pipe Operator for R*. R package version 1.5. URL: https://CRAN.R-project.org/package=magrittr.

Bailey, Norman TJ (1990). *The elements of stochastic processes with applications to the natural sciences*. Vol. 25. John Wiley & Sons.

Barido-Sottani, Joëlle, Timothy G Vaughan, and Tanja Stadler (2020). "A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates". In: *Systematic Biology*.

Beaulieu, Jeremy M and Brian C O'Meara (2016). "Detecting hidden diversification shifts in models of trait-dependent speciation and extinction". In: *Systematic biology* 65.4, pp. 583–601.

Bilderbeek, Richèl J.C. (2020). *mcbette: Model Comparison Using 'babette'*. R package version 1.8.3. URL: https://github.com/richelbilderbeek/mcbette.

Bilderbeek, Richèl JC and Rampal S Etienne (2018). "babette: BEAUti 2, BEAST 2 and Tracer for R". In: *Methods in Ecology and Evolution*.

Blount, Zachary D, Christina Z Borland, and Richard E Lenski (2008). "Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli". In: *Proceedings of the National Academy of Sciences* 105.23, pp. 7899–7906.

Bouckaert, Remco et al. (2012). "Mapping the origins and expansion of the Indo-European language family". In: *Science* 337.6097, pp. 957–960.

Bouckaert, Remco et al. (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 10.4, e1003537.

Bouckaert, Remco et al. (2019). "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 15.4, e1006650.

Bouckaert, Remco R, Claire Bowern, and Quentin D Atkinson (2018). "The origin and expansion of Pama–Nyungan languages across Australia". In: *Nature ecology & evolution* 2.4, pp. 741–749.

Caetano, Daniel S, Brian C O'Meara, and Jeremy M Beaulieu (2018). "Hidden state models improve state-dependent diversification approaches, including biogeographical models". In: *Evolution* 72.11, pp. 2308–2324.

Chaves, Jaime A, Jason T Weir, and Thomas B Smith (2011). "Diversification in Adelomyia hummingbirds follows Andean uplift". In: *Molecular Ecology* 20.21, pp. 4564–4576.

Cohen, Andrew S et al. (1997). "Lake level and paleoenvironmental history of Lake Tanganyika, Africa, as inferred from late Holocene and modern stromatolites". In: *Geological Society of America Bulletin* 109.4, pp. 444–460.

Condamine, Fabien L, Jonathan Rolland, and Helene Morlon (2013). "Macroevolutionary perspectives to environmental change". In: *Ecology letters* 16, pp. 72–85.

Cotton, Richard (2016). *assertive: Readable Check Functions to Ensure Code Integrity*. R package version 0.3-5. URL: `https://CRAN.R-project.org/package=assertive`.

Drummond, Alexei J and Remco R Bouckaert (2015). *Bayesian evolutionary analysis with BEAST*. Cambridge University Press.

Drummond, Alexei J and Andrew Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC evolutionary biology* 7.1, p. 214.

Drummond, Alexei J et al. (2005). "Bayesian coalescent inference of past population dynamics from molecular sequences". In: *Molecular biology and evolution* 22.5, pp. 1185–1192.

Drummond, Alexei J et al. (2006). "Relaxed phylogenetics and dating with confidence". In: *PLoS biology* 4.5, e88.

Duchêne, David A et al. (2015). "Evaluating the adequacy of molecular clock models using posterior predictive simulations". In: *Molecular Biology and Evolution* 32.11, pp. 2986–2995.

Duchene, Sebastian et al. (2018). "Phylodynamic model adequacy using posterior predictive simulations". In: *Systematic biology* 68.2, pp. 358–364.

Esquerré, Damien et al. (2019). "How mountains shape biodiversity: The role of the Andes in biogeography, diversification, and reproductive biology in South America's most species-rich lizard radiation (Squamata: Liolaemidae)". In: *Evolution* 73.2, pp. 214–230.

Etienne, Rampal S. (2017). "Corrigendum". In: *Evolution*. DOI: `10.1111/evo.13314`.

Etienne, Rampal S and Bart Haegeman (2012). "A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence". In: *The American Naturalist* 180.4, E75–E89.

Etienne, Rampal S. and Bart Haegeman (2020). *DDD: Diversity-Dependent Diversification*. R package version 4.2. URL: `https://CRAN.R-project.org/package=DDD`.

Etienne, Rampal S, Hélène Morlon, and Amaury Lambert (2014). "Estimating the duration of speciation from phylogenies". In: *Evolution* 68.8, pp. 2430–2440.

Etienne, Rampal S, Alex L Pigot, and Albert B Phillimore (2016). "How reliably can we infer diversity-dependent diversification from phylogenies?" In: *Methods in Ecology and Evolution* 7.9, pp. 1092–1099.

Etienne, Rampal S and James Rosindell (2012). "Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification". In: *Systematic Biology* 61.2, pp. 204–213.

Etienne, Rampal S et al. (2012). "Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1732, pp. 1300–1309.

Faith, Daniel P (1992a). "Conservation evaluation and phylogenetic diversity". In: *Biological conservation* 61.1, pp. 1–10.

— (1992b). "Systematics and conservation: on predicting the feature diversity of subsets of taxa". In: *Cladistics* 8.4, pp. 361–373.

Farris, James S (1970). "Methods for computing Wagner trees". In: *Systematic Biology* 19.1, pp. 83–92.

Felsenstein, Joseph (1973). "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters". In: *Systematic Biology* 22.3, pp. 240–249.

— (1978). "Cases in which parsimony or compatibility methods will be positively misleading". In: *Systematic zoology* 27.4, pp. 401–410.

— (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach". In: *Journal of molecular evolution* 17.6, pp. 368–376.

Fitch, Walter M (1971). "Toward defining the course of evolution: minimum change for a specific tree topology". In: *Systematic Biology* 20.4, pp. 406–416.

FitzJohn, Richard G (2010). "Quantitative traits and diversification". In: *Systematic biology* 59.6, pp. 619–633.

— (2012). "Diversitree: comparative phylogenetic analyses of diversification in R". In: *Methods in Ecology and Evolution* 3.6, pp. 1084–1092.

Gavryushkina, Alexandra et al. (2014). "Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration". In: *PLoS Comput Biol* 10.12, e1003919.

Gillespie, Daniel T (1976). "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of computational physics* 22.4, pp. 403–434.

— (1977). "Exact stochastic simulation of coupled chemical reactions". In: *The journal of physical chemistry* 81.25, pp. 2340–2361.

Glor, Richard E et al. (2004). "Partial island submergence and speciation in an adaptive radiation: a multilocus analysis of the Cuban green anoles". In: *Pro-*

ceedings of the Royal Society of London. Series B: Biological Sciences 271.1554, pp. 2257–2265.

Goldberg, Emma E and Boris Igić (2012). "Tempo and mode in plant breeding system evolution". In: *Evolution: International Journal of Organic Evolution* 66.12, pp. 3701–3709.

Goldberg, Emma E, Lesley T Lancaster, and Richard H Ree (2011). "Phylogenetic inference of reciprocal effects between geographic range evolution and diversification". In: *Systematic biology* 60.4, pp. 451–465.

Goldman, Nick (1993). "Statistical tests of models of DNA substitution". In: *Journal of molecular evolution* 36.2, pp. 182–198.

Haffer, Jürgen (1969). "Speciation in Amazonian forest birds". In: *Science* 165.3889, pp. 131–137.

Hagen, Oskar et al. (2018). "Estimating age-dependent extinction: contrasting evidence from fossils and phylogenies". In: *Systematic biology* 67.3, pp. 458–474.

Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of molecular evolution* 22.2, pp. 160–174.

Heard, Stephen B and David L Hauser (1995). "Key evolutionary innovations and their ecological mechanisms". In: *Historical Biology* 10.2, pp. 151–173.

Heled, Joseph and Alexei J Drummond (2015). "Calibrated birth–death phylogenetic time-tree priors for Bayesian inference". In: *Systematic Biology* 64.3, pp. 369–383.

Herrera-Alsina, Leonel, Paul van Els, and Rampal S Etienne (2019). "Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data". In: *Systematic biology* 68.2, pp. 317–328.

Herrera-Alsina, Leonel et al. (2018). "The influence of ecological and geographic limits on the evolution of species distributions and diversity". In: *Evolution* 72.10, pp. 1978–1991.

Hester, Jim (2016). *lintr: Static R Code Analysis*. R package version 1.0.0. URL: `http://CRAN.R-project.org/package=lintr`.

Higashi, M, G Takimoto, and N Yamamura (1999). "Sympatric speciation by sexual selection". In: *Nature* 402.6761, pp. 523–526.

Höhna, Sebastian (2013). "Fast simulation of reconstructed phylogenies under global time-dependent birth–death processes". In: *Bioinformatics* 29.11, pp. 1367–1374.

Höhna, Sebastian, Michael R May, and Brian R Moore (2016). "TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates". In: *Bioinformatics* 32.5, pp. 789–791.

Höhna, Sebastian et al. (2016). "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language". In: *Systematic biology* 65.4, pp. 726–736.

Höhna, Sebastian et al. (2019). "A Bayesian Approach for Estimating Branch-Specific Speciation and Extinction Rates". In: *bioRxiv*, p. 555805.

Hua, Xia and Lindell Bromham (2017). "Darwinism for the genomic age: connecting mutation to diversification". In: *Frontiers in genetics* 8, p. 12.

Huelsenbeck, John P and Fredrik Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees". In: *Bioinformatics* 17.8, pp. 754–755.

Janzen, Thijs and Richel Bilderbeek (2020). *nLTT: Calculate the NLTT Statistic*. R package version 1.4.3. URL: `https://CRAN.R-project.org/package=nLTT`.

Janzen, Thijs and Rampal Etienne (2016). "Inferring the role of habitat dynamics in driving diversification: evidence for a species pump in Lake Tanganyika cichlids". In: *bioRxiv*, p. 085431.

Janzen, Thijs, Sebastian Höhna, and Rampal S Etienne (2015). "Approximate Bayesian Computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT". In: *Methods in Ecology and Evolution* 6.5, pp. 566–575.

Janzen, Thijs et al. (2017). "Community assembly in Lake Tanganyika cichlid fish: quantifying the contributions of both niche-based and neutral processes". In: *Ecology and evolution* 7.4, pp. 1057–1067.

Jetz, Walter, Carsten Rahbek, and Robert K Colwell (2004). "The coincidence of rarity and richness and the potential signature of history in centres of endemism". In: *Ecology Letters* 7.12, pp. 1180–1191.

Jukes, Thomas H, Charles R Cantor, et al. (1969). "Evolution of protein molecules". In: *Mammalian protein metabolism* 3.21, p. 132.

Kendall, David G (1948a). "On some modes of population growth leading to RA Fisher's logarithmic series distribution". In: *Biometrika* 35.1/2, pp. 6–15.

— (1948b). "On the generalized" birth-and-death" process". In: *The annals of mathematical statistics*, pp. 1–15.

Kozlov, Alexey M et al. (2019). "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference". In: *Bioinformatics* 35.21, pp. 4453–4455.

Kühnert, Denise et al. (2014). "Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model". In: *Journal of the Royal Society Interface* 11.94, p. 20131106.

Lambert, Amaury, Hélène Morlon, and Rampal S Etienne (2015). "The reconstructed tree in the lineage-based model of protracted speciation". In: *Journal of mathematical biology* 70.1-2, pp. 367–397.

Lambert, Amaury and Tanja Stadler (2013). "Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies". In: *Theoretical population biology* 90, pp. 113–128.

Laudanno, Giovanni (2020a). `https://github.com/Giappo/mbd`.

— (2020b). `https://github.com/Giappo/sls`.

Laudanno, Giovanni, Bart Haegeman, and Rampal S Etienne (2020). "Additional analytical support for a new method to compute the likelihood of diversification models". In: *Bulletin of Mathematical Biology* 693176.

Laudanno, Giovanni et al. (2020). "Detecting Lineage-Specific Shifts in Diversification: A Proper Likelihood Approach". In: *Systematic Biology*. DOI: `10.1093/sysbio/syaa048`.

Lemey, Philippe, Marco Salemi, and Anne-Mieke Vandamme (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

Levins, Richard (1966). "The strategy of model building in population biology". In: *American scientist* 54.4, pp. 421–431.

Liem, Karel F (1973). "Evolutionary strategies and morphological innovations: cichlid pharyngeal jaws". In: *Systematic Zoology* 22.4, pp. 425–441.

Louca, Stilianos and Matthew W Pennell (2020). "Extant timetrees are consistent with a myriad of diversification histories". In: *Nature* 580.7804, pp. 502–505.

Maddison, Wayne P, Peter E Midford, and Sarah P Otto (2007). "Estimating a binary character's effect on speciation and extinction". In: *Systematic biology* 56.5, pp. 701–710.

Maechler, Martin (2019). *Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable*. R package version 0.7-2. URL: `https://CRAN.R-project.org/package=Rmpfr`.

Mahler, D Luke et al. (2010). "Ecological opportunity and the rate of morphological evolution in the diversification of Greater Antillean anoles". In: *Evolution* 64.9, pp. 2731–2745.

Maliet, Odile, Florian Hartig, and Hélène Morlon (2019). "A model with many small shifts for estimating species-specific diversification rates". In: *Nature ecology & evolution* 3.7, pp. 1086–1092.

Manceau, Marc et al. (2019). “The ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data”. In: *BioRxiv*, p. 755561.

May, Michael R and Brian R Moore (2016). “How well can we detect lineage-specific diversification-rate shifts? A simulation study of sequential AIC methods”. In: *Systematic Biology* 65.6, pp. 1076–1084.

Mitter, Charles, Brian Farrell, and Brian Wiegmann (1988). “The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification?” In: *The American Naturalist* 132.1, pp. 107–128.

Moore, Brian R et al. (2016). “Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures”. In: *Proceedings of the National Academy of Sciences* 113.34, pp. 9569–9574.

Moore, William S (1995). “Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees”. In: *Evolution* 49.4, pp. 718–726.

Muellner-Riehl, Alexandra N et al. (2019). “Origins of global mountain plant biodiversity: Testing the "mountain-geobiodiversity hypothesis"”. In: *Journal of Biogeography* 46.12, pp. 2826–2838.

Nagoshi, Makoto (1983). “Distribution, abundance and parental care of the genus Lamprologus (Cichlidae) in Lake Tanganyika”. In: *African Study Monographs* 3, pp. 39–47.

Nee, Sean, Robert Mccredie May, and Paul H Harvey (1994). “The reconstructed evolutionary process”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1309, pp. 305–311.

Pannetier, Théo et al. (2020). “Branching patterns in phylogenies cannot distinguish diversity-dependent diversification from time-dependent diversification”. In: *Evolution*.

Papadopoulou, Anna and L Lacey Knowles (2015). “Genomic tests of the species-pump hypothesis: recent island connectivity cycles drive population divergence but not speciation in Caribbean crickets across the Virgin Islands”. In: *Evolution* 69.6, pp. 1501–1517.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer (2004). “APE: analyses of phylogenetics and evolution in R language”. In: *Bioinformatics* 20.2, pp. 289–290.

Pybus, Oliver G and Paul H Harvey (2000). “Testing macro–evolutionary models using incomplete molecular phylogenies”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267.1459, pp. 2267–2272.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rabosky, Daniel L (2014). "Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees". In: *PloS one* 9.2, e89543.

Rabosky, Daniel L and Irby J Lovette (2008). "Explosive evolutionary radiations: decreasing speciation or increasing extinction through time?" In: *Evolution: International Journal of Organic Evolution* 62.8, pp. 1866–1875.

Rabosky, Daniel L., Jonathan S. Mitchell, and Jonathan Chang (2017). "Is BAMM Flawed? Theoretical and Practical Concerns in the Analysis of Multi-Rate Diversification Models". In: *Systematic Biology* 66.4, pp. 477–498. DOI: 10.1093/sysbio/syx037.

Rangel, Thiago F et al. (2018). "Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves". In: *Science* 361.6399.

Ratnakumar, Sridhar, Trent Mick, and Trevor Davis (2016). *rappdirs: Application Directories: Determine Where to Save Data, Caches, and Logs*. R package version 0.3.1. URL: https://CRAN.R-project.org/package=rappdirs.

Revell, Liam J (2012). "phytools: an R package for phylogenetic comparative biology (and other things)". In: *Methods in ecology and evolution* 3.2, pp. 217–223.

Ritchie, Andrew M, Nathan Lo, and Simon YW Ho (2017). "The impact of the tree prior on molecular dating of data sets containing a mixture of inter-and intraspecies sampling". In: *Systematic Biology* 66.3, pp. 413–425.

Ronco, Fabrizia et al. (2019). "The taxonomic diversity of the cichlid fish fauna of ancient Lake Tanganyika, East Africa". In: *Journal of Great Lakes Research*.

Ronquist, Fredrik and John P Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". In: *Bioinformatics* 19.12, pp. 1572–1574.

Rosindell, James et al. (2010). "Protracted speciation revitalizes the neutral theory of biodiversity". In: *Ecology Letters* 13.6, pp. 716–727.

Russel, Patricio Maturana et al. (2019). "Model selection and parameter inference in phylogenetics using nested sampling". In: *Systematic biology* 68.2, pp. 219–233.

Sarver, Brice AJ et al. (2019). "The choice of tree prior and molecular clock does not substantially affect phylogenetic inferences of diversification rates". In: *PeerJ* 7, e6334.

Schliep, Klaus Peter (2011). "phangorn: phylogenetic analysis in R". In: *Bioinformatics* 27.4, pp. 592–593.

Schluter, Dolph (2000). *The Ecology of Adaptive Radiation*. Oxford University Press.

Sedano, Raul E and Kevin J Burns (2010). "Are the Northern Andes a species pump for Neotropical birds? Phylogenetics and biogeography of a clade of Neotropical tanagers (Aves: Thraupini)". In: *Journal of Biogeography* 37.2, pp. 325–343.

Simon, Herbert A (1991). "The architecture of complexity". In: *Facets of systems science*. Springer, pp. 457–476.

Simpson, George Gaylord (1944). *Tempo and mode in evolution*. 15. Columbia University Press.

— (1955). *Major Features of Evolution*. Columbia University Press.

Stadler, Tanja (2009). "On incomplete sampling under birth–death models and connections to the sampling-based coalescent". In: *Journal of theoretical biology* 261.1, pp. 58–66.

— (2011). "Mammalian phylogeny reveals recent diversification rate shifts". In: *Proceedings of the National Academy of Sciences* 108.15, pp. 6187–6192.

— (2012). "How can we improve accuracy of macroevolutionary rate estimates?" In: *Systematic Biology* 62.2, pp. 321–329.

Stadler, Tanja et al. (2012). "Estimating the basic reproductive number from viral sequence data". In: *Molecular biology and evolution* 29.1, pp. 347–357.

Stadler, Tanja et al. (2013). "Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)". In: *Proceedings of the National Academy of Sciences* 110.1, pp. 228–233.

Sturmbauer, Christian et al. (2001). "Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes". In: *Molecular Biology and Evolution* 18.2, pp. 144–154.

Sturmbauer, Christian et al. (2010). "Evolutionary history of the Lake Tanganyika cichlid tribe Lamprologini (Teleostei: Perciformes) derived from mitochondrial and nuclear DNA data". In: *Molecular Phylogenetics and Evolution* 57.1, pp. 266–284.

Tamura, Koichiro and Masatoshi Nei (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees." In: *Molecular biology and evolution* 10.3, pp. 512–526.

Tavaré, Simon (1986). "Some probabilistic and statistical problems in the analysis of DNA sequences". In: *Lectures on mathematics in the life sciences* 17.2, pp. 57–86.

Title, Pascal O and Daniel L Rabosky (2019). "Tip rates, phylogenies and diversification: what are we estimating, and how good are the estimates?" In: *Methods in Ecology and Evolution* 10.6, pp. 821–834.

Turner, George F et al. (2001). "How many species of cichlid fishes are there in African lakes?" In: *Molecular Ecology* 10.3, pp. 793–806.

Valente, Luis M, Rampal S Etienne, and Albert B Phillimore (2014). "The effects of island ontogeny on species diversity and phylogeny". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1784, p. 20133227.

Valente, Luis M, Albert B Phillimore, and Rampal S Etienne (2015). "Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands". In: *Ecology letters* 18.8, pp. 844–852.

Verheyen, Erik et al. (1996). "Mitochondrial phylogeography of rock-dwelling cichlid fishes reveals evolutionary influence of historical lake level fluctuations of Lake Tanganyika, Africa". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 351.1341, pp. 797–805.

Wagner, Catherine E, Luke J Harmon, and Ole Seehausen (2012). "Ecological opportunity and sexual selection together predict adaptive radiation". In: *Nature* 487.7407, pp. 366–369.

Weir, J. T. et al. (2016). "Explosive ice age diversification of kiwi". In: *Proceedings of the National Academy of Sciences* 113.38, E5580–E5587.

Wellborn, Gary A and R Brian Langerhans (2015). "Ecological opportunity and the adaptive diversification of lineages". In: *Ecology and Evolution* 5.1, pp. 176–195.

Wickham, Hadley (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN: 978-0-387-98140-6. URL: `http://had.co.nz/ggplot2/book`.

— (2011). *testthat: Get Started with Testing*, pp. 5–10. URL: `http://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf`.

— (2015). *R packages: organize, test, document, and share your code*. O'Reilly Media, Inc.

— (2017). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.2.0. URL: `https://CRAN.R-project.org/package=stringr`.

Wickham, Hadley and Winston Chang (2016). *devtools: Tools to Make Developing R Packages Easier*. R package version 1.12.0.9000. URL: `http://CRAN.R-project.org/package=devtools`.

Wickham, Hadley and Lionel Henry (2019). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.3. URL: https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley et al. (2011). "The split-apply-combine strategy for data analysis". In: *Journal of Statistical Software* 40.1, pp. 1–29.

Wickham, Hadley et al. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5. URL: https://CRAN.R-project.org/package=dplyr.

Wu, Guohong Albert et al. (2018). "Genomics of the origin and evolution of Citrus". In: *Nature* 554.7692, pp. 311–316.

Xie, Yihui (2014). *testit: A Simple Package for Testing R Packages*. R package version 0.4, http://CRAN.R-project.org/package=testit. URL: http://CRAN.R-project.org/package=testit.

— (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17. URL: https://yihui.name/knitr/.

Xu, Liang and Rampal S Etienne (2018). "Detecting local diversity-dependence in diversification". In: *Evolution* 72.6, pp. 1294–1305.

Yoder, JB et al. (2010). "Ecological opportunity and the origin of adaptive radiations". In: *Journal of Evolutionary Biology* 23.8, pp. 1581–1596.

Yule, George Udny (1925). "II.—A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S". In: *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character* 213.402-410, pp. 21–87.

Zuckerkandl, Emile and Linus Pauling (1965). "Molecules as documents of evolutionary history". In: *Journal of theoretical biology* 8.2, pp. 357–366.