ECOLOGY LETTERS

Ecology Letters, (2015) doi: 10.1111/ele.12415

LETTER

Phylogenies support out-of-equilibrium models of biodiversity

Marc Manceau,[1,2] Amaury Lambert,[2,3] and Hélène Morlon,[1,2]

Abstract

There is a long tradition in ecology of studying models of biodiversity at equilibrium. These models, including the influential Neutral Theory of Biodiversity, have been successful at predicting major macroecological patterns, such as species abundance distributions. But they have failed to predict macroevolutionary patterns, such as those captured in phylogenetic trees. Here, we develop a model of biodiversity in which all individuals have identical demographic rates, metacommunity size is allowed to vary stochastically according to population dynamics, and speciation arises naturally from the accumulation of point mutations. We show that this model generates phylogenies matching those observed in nature if the metacommunity is out of equilibrium. We develop a likelihood inference framework that allows fitting our model to empirical phylogenies, and apply this framework to various mammalian families. Our results corroborate the hypothesis that biodiversity dynamics are out of equilibrium.

Keywords

Neutral Biodiversity Theory, Macroevolution, Speciation by Genetic Differentiation, Birth-death models, Metacommunity dynamics, Diversification, Speciation, Extinction.

Ecology Letters (2015)

INTRODUCTION

Ever since MacArthur and Wilson proposed their equilibrium theory of island biogeography (MacArthur & Wilson 1967), equilibrium models have played a major role in ecology. Of particular influence has been the Neutral Theory of Biodiversity (NTB) (Hubbell 2001) that has allowed to analytically derive major macroecological patterns at equilibrium, including the species abundance distribution (Etienne & Alonso 2005), the species–area relationship (O'Dwyer & Green 2010), and the distance–decay relationship (Chave & Leigh 2002; O'Dwyer & Green 2010). The NTB has been relatively successful at predicting realistic macroecological patterns, making this model a central model in ecology (but see e.g. McGill *et al*. 2006 for a debate on the empirical support of NTB). The theory, however, has been much less successful at predicting realistic macroevolutionary patterns, in particular phylogenetic tree shapes (Davies *et al*. 2011). At a time when ecologists are increasingly interested in the role of history on present-day patterns of biodiversity (Webb *et al*. 2002; Wiens *et al*. 2010), in understanding phylogenetic patterns of diversity (Graham & Fine 2008; Morlon *et al*. 2011b), and in preserving evolutionary history (Nee & May 1997; Lambert & Steel 2013), designing a model of biodiversity predicting realistic phylogenetic trees is critically needed.

There are macroevolutionary models capable of predicting realistic phylogenies (see Morlon 2014 for a recent review). However, most of these models are based on so-called 'birth-death models of cladogenesis', which were historically designed to estimate rates of speciation and extinction in groups where fossil data are scarce (Nee *et al*. 1992). Since they were first introduced, lineage-based models have been further developed to account for diversity-dependent effects, as well as heterogeneities in diversification rates across time and species groups (Rabosky & Lovette 2008; Alfaro *et al*. 2009; Morlon *et al*. 2010; Etienne *et al*. 2011; Morlon *et al*. 2011a; Stadler 2011; Lambert & Stadler 2013; Rabosky 2014). The simplest models, which assume time-constant speciation and extinction rates, produce trees that are more balanced and 'tippy' than empirical trees (Blum & François 2006; Mooers *et al*. 2007). Realistic balance can be obtained by allowing diversification rates to vary across lineages, while realistic branching times (sensu Morlon 2014) can be obtained by allowing diversification rates to vary through time.

Birth–death models of cladogenesis have tremendous applications for understanding biodiversity patterns (Morlon 2014). They, however, have serious limitations. In particular, birth–death models consider the 'birth' (speciation) and 'death' (extinction) events of lineages, or species, while ignoring the numbers of individuals constituting these species. By not incorporating population dynamics, these models implicitly assume that speciation and extinction events are independent from species' population sizes. However, several lines of evidence suggest that species' abundances and the extant of their geographic range influence probabilities of speciation and extinction (Rosenzweig 1995). Larger areas likely offer greater opportunities for geographical isolation due to a higher incidence of dispersal barriers, greater habitat heterogeneity, and the limits to gene flow (Pigot *et al*. 2010). Large

[1] *École Normale Supérieure, Institut de Biologie, CNRS UMR 8197, 46 rue d'Ulm, 75005 Paris, France*

[2] *Collège de France, Center for Interdisciplinary Research in Biology, CNRS UMR 7241, 11 place Marcelin-Berthelot, 75005 Paris, France*

[3] *UPMC Univ Paris 06, Laboratoire de Probabilités et Modèles Aléatoires, CNRS UMR 7599, 4 place Jussieu, 75005 Paris, France*

Correspondence: E-mail: marc.manceau@ens.fr

ranges also provide a buffer against stochastic or environmentally driven fluctuations in size that may lead to extinction (McKinney 1997). In addition, it would seem more natural to model extinction by a process in which all individuals die, rather than a process independent of population sizes.

There exist very few evolutionary models that explicitly incorporate population or range sizes and yield predictions for phylogenetic trees (Hubbell 2001; McPeek 2008; Pigot *et al.* 2010). Contrary to traditional lineage-based birth–death models for which likelihood expressions allow parameter inference and model comparison, the phylogenetic trees arising from these models have mainly been investigated with simulations. Parameter inference approaches have been developed only for Hubbell's NTB, using approximate Bayesian computation and data on local species abundance and phylogenetic relatedness (Jabot & Chave 2009). Inference methods for McPeek's model of ecological differentiation (McPeek 2008) and Pigot's model of geographic speciation (Pigot *et al.* 2010), which produce trees with realistic branching times, have yet to be developed.

In this study, we develop a new individual-based neutral model inspired by Hubbell's neutral model. One of the big contributions of Hubbell's model has been to provide a 'unified' theory of biodiversity accounting for both the short time-scale processes of individuals' birth and death, and the long time-scale processes of speciation and extinction. As a result, the theory generates predictions for both macroecological patterns, such as species abundance distributions, and macroevolutionary patterns, such as phylogenies. Our model keeps this same 'unifying' particularity, thus also generating both types of patterns. Hubbell's original NTB model relies on three main assumptions. The first one, known as the hypothesis of neutrality, is that individuals behave similarly whichever species they belong to. In the continuity of this hypothesis, we stick to the assumption that individual demographic rates are independent of species identity. The second assumption, known as the zero-sum assumption, is that the metacommunity size is constant. Each death event is assumed to occur simultaneously with a birth event, as in the Moran process of population genetics. In our model, we instead allow metacommunity size to vary according to the stochastic birth and death of individuals. Metacommunity size is not bounded; for example, if the birth and death rates remain constant through time, the metacommunity grows exponentially. Unlike what happens in a metacommunity of constant size where diversity necessarily reaches an equilibrium limit, diversity may not be bounded in our model. Given that previous analyses found little support for equilibrium diversity models in terms of phylogenetic branching times (Morlon *et al.* 2010), we hypothesised that relaxing the zero-sum assumption could lead to more realistic branching times. The third assumption of NTB is linked to the speciation process. Here, we design a mode of speciation based on gradual genetic differentiation that presents several advantages compared to previously considered speciation modes. We analyse phylogenies arising from this model, provide related likelihood formulas and apply the model to mammalian trees. Finally, we discuss the implication of the results for our understanding of biodiversity dynamics.

THE MODEL OF SPECIATION BY GENETIC DIFFERENTIATION

We consider a model of biodiversity incorporating population dynamics, mutations and speciation events, hereafter referred to as the model of Speciation by Genetic Differentiation (SGD). This model and the resulting phylogenies are illustrated in Fig. 1 and are summarised in Box 1. Population dynamics are given by a stochastic birth–death process in which individuals give birth and die with rates $b(t)$ and $d(t)$ that are identical across individuals and can potentially vary with time t (see Fig. 1a). Genetic mutations arise at per-individual rate $v(t)$. We derive all our analytical results in the most general case, with the three rates varying through time. In our empirical applications of the model, however, we consider the case of constant rates, denoted b, d and v.

Similar to the infinite-allele model in population genetics, each mutation gives rise to an entirely new genetic type. These mutations are assumed to be neutral, meaning that they do not affect the demography of individuals. We define species as being the smallest monophyletic groups of extant individuals such that any two individuals of same genetic type always belong to the same group. Hence, speciation occurs when two sister populations no longer contain individuals of the same genetic type. This typically happens as follows: A first birth event in an ancestral individual generates two descents. At least one individual in either descent undergoes a mutation. Genetic drift makes the two descents fully differentiated (e.g. if there is one mutation, this mutation invades the population by drift) leading to speciation. In the context of sexually reproducing species, mutations can be seen as barriers to hybridisation, either pre-zygotic (e.g. in the form of mechanical, behavioural or habitat isolation), or post-zygotic (hybrid inviability or sterility). Species are then the smallest monophyletic groups of individuals such that any two individuals that are interfertile always belong to the same group (see Fig. 1 and Box 1). Finally, SGD naturally includes extinction events, which occur when all individuals of a species die without leaving any descendant. The constant-rate SGD model is thus entirely characterised by only three parameters (b, d and v).

As long as the ancestral type is alive in two descents from one ancestral individual, these descents form a single species. Hence, the speciation rate is negatively correlated with the time it takes for the ancestral type to disappear in at least one of the two descents and one expects to see different behaviors of the model as a function of a trade-off between population growth rate ($b - d$) and the mutation rate (v) (see Fig. S1 for an illustration). As mutations are drawn following a Poisson process of parameter v on the reconstructed genealogy, it adds a death rate of parameter v for clonal lines. A clonal population therefore follows a birth–death process with parameters b and $d + v$. When v is larger than $b - d$, clonal populations have a negative net growth rate ($b - d - v$) and thus cannot survive indefinitely (Champagnat & Lambert 2013). When v is smaller than $b - d$, however, clones can coexist indefinitely. Hence, there should be a phase transition with long-lived lineages when $b - d > v$ and fast lineage turnover when $b - d < v$.

The mode of speciation resulting from SGD is more biologically realistic than the point mutation mode of speciation

Box 1 From the genealogy of individuals to a species-level phylogeny in the model of speciation by genetic differentiation

The process of speciation by genetic differentiation (SGD) and the resulting phylogeny are illustrated in Fig. 1. Here, we clarify the terms used throughout and how phylogenies are obtained from the underlying genealogies of individuals.

GENEALOGY OF INDIVIDUALS

The *genealogy* is the tree representing ancestor–descendant relationships for all individuals arising from the individual-based birth–death process. A *line* is a path on the genealogy joining an ancestral individual at the base to its successive descendants. The *reconstructed genealogy* is the genealogy in which extinct lines have been removed. The *descent* of an individual is the genealogical subtree of all individuals descending from this ancestral individual (on the genealogy or the reconstructed genealogy). The *metacommunity* is the set of all individuals alive at a given time.

MUTATIONS

A mutation event on the genealogy is an event that changes the *genetic type* of the mutant individual and all its descent. The *clonal descent* of an individual is the set of all individuals of same genetic type descending from this ancestral individual. In Fig. 1a and b, each clonal descent of a mutant is represented with a different color. We need to consider *divergent nodes* on the reconstructed genealogy, defined as nodes such that any two pairs of individuals picked at random, one in each of the two extent descents from the node, are of different genetic types.

SPECIES

A *species* is the smallest monophyletic group of extant individuals such that any two individuals of same genetic type always belong to the same group. This ensures consistency between genealogical and phylogenetic relationships, in agreement with the genealogical concordance species concept (Avise & Ball 1990).

PHYLOGENIES

The reconstructed phylogeny (simply called *phylogeny* throughout) is the tree representing the evolutionary relationships between extant species. A *phylogenetic node* is a node that appears on the reconstructed phylogeny. Given our monophyletic definition of species, we can take phylogenetic nodes as subsets of genealogical nodes. Phylogenetic nodes are obtained recursively as follows. The oldest node in the reconstructed genealogy is phylogenetic if it is divergent (otherwise the phylogeny is made of a single extant species). Each other genealogical node is phylogenetic if its parent node is phylogenetic and if it is divergent. *Branching times* are the times when there is a phylogenetic node. A *lineage* is a path on the phylogeny (in contrast to a *line*, which is a path on the genealogy). A *tip lineage* is a lineage joining an extant species to its most recent ancestral node. It can correspond to one or several genealogical lines. All other lineages are *internal lineages* and correspond to a single line in the genealogy.

LINEAGE TYPES

A *type 0 lineage* is a lineage starting from an ancestral individual whose clonal descent survived to the present. A *type 1 lineage* is a lineage starting from an ancestral individual whose clonal descent did not survive to the present. A *'frozen' lineage* is a lineage that cannot experience further speciation or extinction events. This happens when the ancestral individual at the base of the lineage gives birth to two individuals with both clonal descents surviving up to the present.

under which most of NTB's previous analytical results have been derived. Contrary to the point mutation model in which speciation happens instantaneously, mutations in the SGD model give rise to new genetic types rather than new species. Speciation thus takes time to complete, as a result of a gradual accumulation of genetic differentiation. In this respect, our model holds some analogies with the protracted mode of speciation introduced by Rosindell and Etienne (Rosindell *et al.* 2010; Etienne & Rosindell 2011; Etienne *et al.* 2014; Lambert *et al.* 2015). In the protracted speciation model, speciation is modeled as a gradual rather than instantaneous process, such that a population of a new type gives rise to a new species only after a fixed or random time span. In our model, the time span is not an input of the model, but rather arises naturally from the accumulation of mutations. In addition, our model generates monophyletic species which are not clonal (i.e. there is genetic diversity within species).

Our primary interest lies in the shape of phylogenies arising from the SGD model. These phylogenies are obtained by a three-step process: (1) population dynamics generate a stochastic genealogy of individuals (Fig. 1a), (2) mutations arise on the genealogy according to a Poisson process (Fig. 1a and

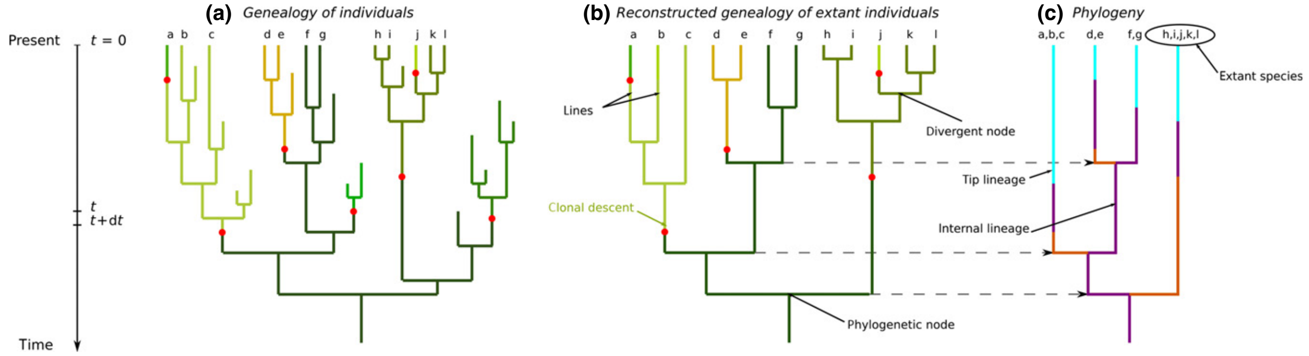

Fig. 1. Phylogeny arising from the model of speciation by genetic differentiation (see Box 1 for details). (a) Genealogy arising from the stochastic birth and death of individuals. Red dots denote mutations. Each mutation gives rise to a new genetic type (represented with a new color) characterising the mutated individual and all its descent until the next mutation. (b) Resulting reconstructed genealogy of extant individuals, obtained by removing all dead lines from the genealogy. (c) Resulting phylogeny, obtained as explained in Box 1. Our derivations and simulations involve defining lineages of different types. Purple lineages are type 0 lineages (extant genetic types that are not frozen), orange lineages are type 1 lineages (extinct genetic types) and the blue lineages are 'frozen' lineages (lineages that cannot experience further speciation or extinction events). Letters at the tips of the genealogy and phylogeny represent individuals. The set of all extant individuals forms the present-day metacommunity.

b), (3) the phylogeny is a subtree of the reconstructed genealogy obtained according to our species definition, as detailed in Fig. 1 and Box 1.

THEORETICAL RESULTS

Key formulas

We aim to analyse phylogenies arising from the SGD model and to develop tools for fitting the model to empirical phylogenies. We measure time from the present to the past, such that $t = 0$ denotes the present and t increases into the past (Fig. 1). As we show below, we can simulate phylogenies under SGD efficiently (i.e. without simulating the whole individual-based process) and compute associated likelihood formulas using the analytical expressions of two key probabilities. The phylogeny strongly relies on the reconstructed genealogy of individuals, and the two probabilities relate to events happening on this genealogy. The probability of observing a branching event in the reconstructed genealogy between t and $t + dt$ is denoted $g(t)dt$. It corresponds to the probability that an ancestral individual gives birth to two individuals whose descents do not go extinct before the present. The probability that an ancestral individual living at time t has at least one descendant at present carrying its genetic type (i.e. there is a descendant line without mutation), conditioned on the survival of at least one descendant, is denoted $m(t)$.

Using results from Kendall (1948), we show (Supporting information) that $g(t)$ and $m(t)$ are given by:

$$g(t) = \frac{b(t)e^{\int_0^t b(z)-d(z)dz}}{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}ds} \tag{1}$$

$$m(t) = \frac{e^{\int_0^t b(z)-d(z)-\nu(z)dz}}{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)-\nu(z)dz}ds}\,\frac{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}}{e^{\int_0^t b(z)-d(z)dz}} \tag{2}$$

These probabilities relate to events happening on the genealogies of individuals and therefore do not depend on population sizes. Intuitively, $m(t)$ is an inverse measure of genetic drift and depends on a trade-off between population growth and mutation events.

Simulating phylogenies arising from the model

We show (Supporting information) that phylogenies under SGD can be generated by a multitype branching process with the three following types (see Fig. 1c).

(1) a lineage of type 0 is an extant genetic type. It corresponds to a line from the underlying genealogy that has at least one descendant of the same genetic type at present.
(2) a lineage of type 1 is an extinct genetic type. It corresponds to a line from the underlying genealogy that has no descendant of same genetic type at present.
(3) a lineage of type 0 'freezes' when it cannot experience further splitting or extinction events up to the present. This occurs if there exists at least two individuals of the same genetic type, one in each of the two descents from the incident node in the underlying genealogy. In this case, the whole descent of this node is collapsed into a single species.

We derive the rates of the following events, at any given time t (Supporting information):
A lineage of type 1 becomes of type 0:

$$\rho_{1\to 0}(t) = \frac{\nu(t)m(t)}{1-m(t)}$$

A lineage of type 1 branches and gives rise to two lineages of type 1:

$$\rho_{1\to +1}(t) = g(t)(1-m(t))$$

A lineage of type 0 branches and gives rise to one lineage of type 0 and one lineage of type 1:

$$\rho_{0\to +1}(t) = 2g(t)(1-m(t))$$

A lineage of type 0 freezes, giving rise to a tip lineage in the phylogeny:

$$\rho_{0\to\varnothing}(t) = g(t)m(t)$$

To simulate a phylogeny for a total time duration T, we start with a single lineage whose type is 0 with probability $m(T)$ and 1 with probability $1 - m(T)$. We then simulate the above events with the corresponding rates, until time $t = 0$ is reached. This provides a very efficient way of simulating phylogenies arising from SGD.

The detailed protocol for these simulations is provided as Supporting information.

Computing the likelihood of phylogenies arising from the model

We assume that a clade has evolved according to the SGD model. We allow for the possibility that some extant species are missing from the phylogeny of this clade by assuming that each extant species was sampled with probability f. We derive differential equations governing the probability of observing any ultrametric tree given our model. We obtain a set of coupled differential equations involving the two key functions g and m, which can be computed analytically in the case $f = 1$ and integrated numerically in the case $f < 1$. This allows us to follow a natural 'peeling algorithm' (Felsenstein 1981), which consists in computing recursively the likelihood of a tree, by decomposing it into subtrees until finding tip lineages. Likelihoods satisfy ordinary differential equations that we solve using numerical integration. The differential equations, analytical solutions and details of the algorithm are given as Supporting information.

Estimating the parameters of the model

Given a phylogeny, the parameters of the model can be estimated by maximum likelihood. To test the ability of the approach to recover the true parameters, we simulated phylogenies under a wide range of parameter values and applied our maximum likelihood inference algorithm. We found that the approach performs well to recover the net growth rate $b - d$ and the mutation rate v (Fig. 2). Estimates of b alone are biased, due to the fact that this parameter has a weak influence on the likelihood surface.

Codes for the simulations, likelihood computations and parameter estimations are available in Python from the authors upon request. They are also implemented in the R package RPANDA (Morlon *et al.* 2014).

EMPIRICAL RESULTS

Phylogenies arising from the model have realistic balance and branching times

To test whether our model produces realistic trees, we analysed the branching times and balance of both simulated and empirical trees. We use the γ statistic (Pybus & Harvey 2000) to measure branching times. This statistic reflects the relative position of nodes in a phylogeny: stemmy phylogenies (i.e. phylogenies with many nodes close to the root) are characterised by negative γ values, while tippy ones are characterised by positive γ values. We use the β statistic (Blum & François 2006) to measure phylogenetic balance, computed by maximum likelihood using the R package *apTreeshape*.

We begin by evaluating how each of the three parameters of the time-constant SGD model influences phylogenetic trees. To do this, we vary each parameter while constraining the others (Fig. 3). These analyses confirm that tree shape is principally constrained by a balance between the population growth rate $b - d$ and the mutation rate v. The higher the mutation rate v at $b - d$ constant, the higher γ and β, meaning trees tend to be tippy and balanced. On the contrary, higher $b - d$ values at v constant lead to lower γ and β, that is, stemmy and unbalanced trees. The parameter b alone has little if any impact on both β and γ, thus explaining why there is little signal in phylogenies to infer this parameter.

We compare the γ and β values of simulated phylogenies to the γ and β values of the 84 empirical binary trees from McPeek's repository with more than 10 species (McPeek 2008). We considered only trees with more than 10 species because the variance in β values increases very rapidly for trees of small size, and thus estimates of β can be inaccurate

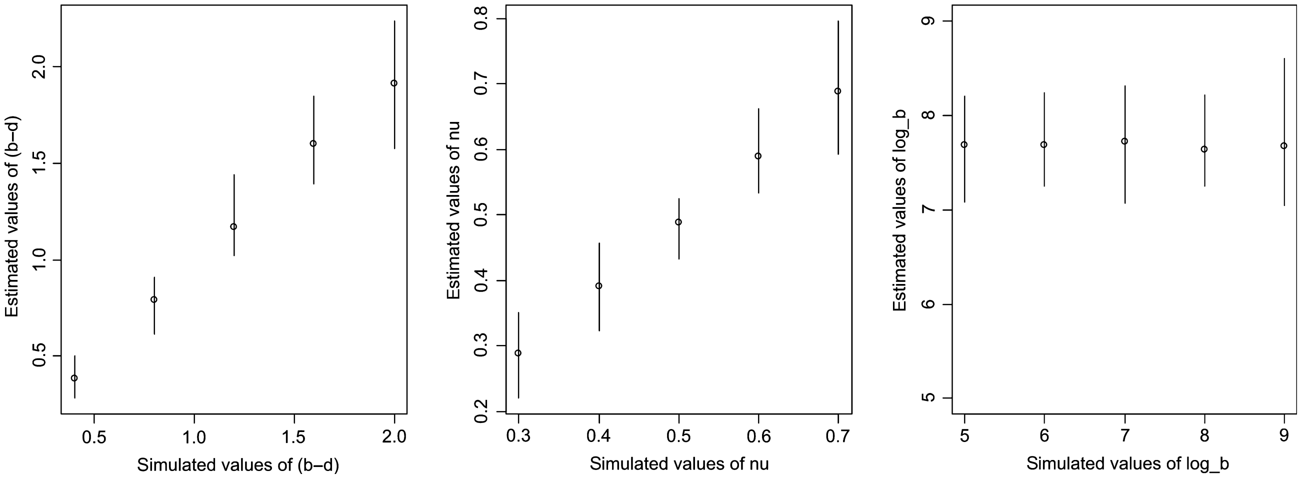

Fig. 2. Growth rates and mutation rates can be robustly inferred from molecular phylogenies, but not birth rates. The figure shows maximum likelihood parameter estimates for phylogenies simulated under different parameter sets. The true, simulated parameters are indicated on the x axis, while inferences are indicated on the y axis (expressed in number of events per time unit). Points and error bars indicate the median and 95% quantile range of the maximum likelihood parameter estimates. Left panel: estimates of $b - d$ are unbiased ($b = 10^6$ and $v = 0.5$). Middle panel: estimates of v are unbiased ($b = 10^6$ and $b - d = 0.8$). Right panel: estimates of b are biased ($b - d = 0.8$ and $v = 0.5$). Units are expressed in Myr^{-1}.

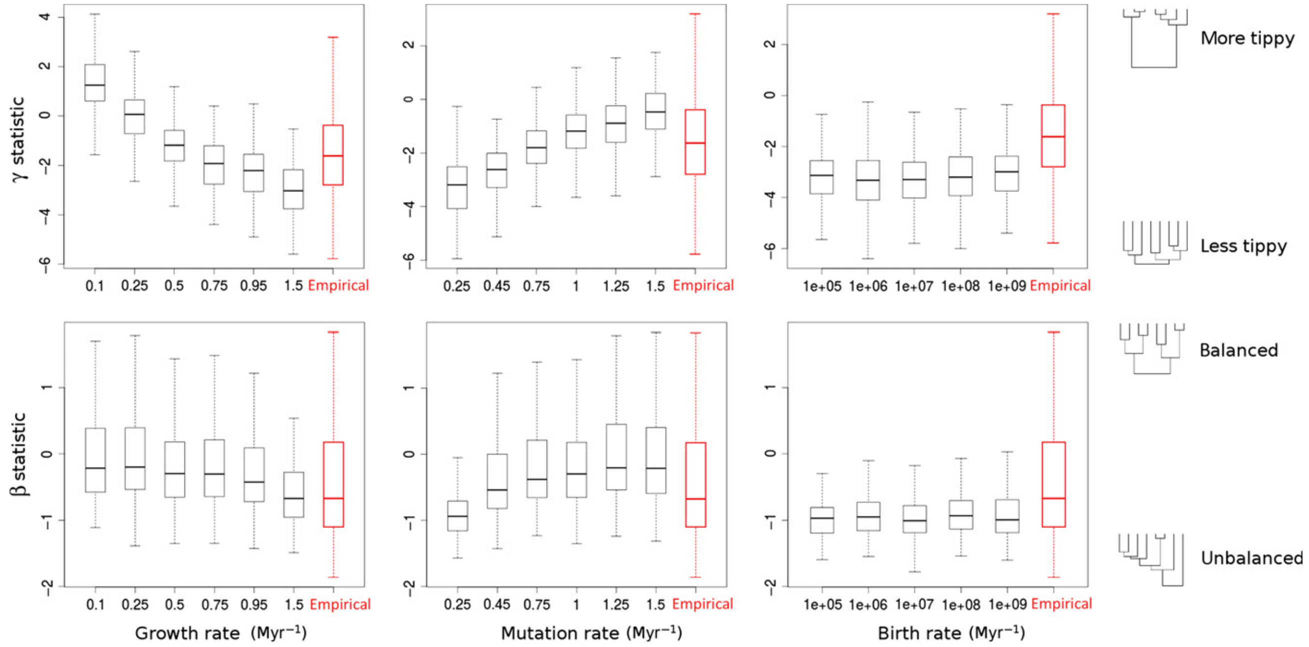

Fig. 3. Branching times and balance under the model of speciation by genetic differentiation. First column : high growth rates $b - d$ at constant birth and mutation rates lead to phylogenies that are stemmy and unbalanced ($b = 10^6$, $v = 1$). Second column: high mutation rate v at constant birth and growth rates ($b = 10^6$, $b - d = 0.5$) lead to phylogenies that are tippy and balanced. Third column : the birth rate b has little effect on phylogenies at constant growth and mutation rates ($b - d = 1$, $v = 0.5$). Units are expressed in Myr^{-1}. Each box-plot summarises results for 200 simulated phylogenies. Empirical box-plot corresponds to the 84 binary phylogenies in the McPeek repository comprising more than 10 species.

for small trees (Blum & François 2006). We find that the SGD model can generate trees with a wide range of β and γ values, including those of empirical trees (Fig. 3). The model produces trees with levels of imbalance and branching times similar to those observed in nature, but only when the growth rate is sufficiently large (of the same order of magnitude as v), meaning in out-of-equilibrium dynamics.

Fit to mammalian phylogenies

We infer the parameters of our model for 14 mammalian phylogenies used by Pigot *et al.* (2012). For each fit, the value of f is fixed, computed by dividing the number of species in the tree by the total number of known species in the clade. Figure 4 shows the likelihood surface corresponding to four of these phylogenies, which are typical of likelihood surfaces obtained for the data. These likelihood surfaces confirm that there is no ambiguity in finding the maximum likelihood parameters, with a single, well-defined peak, and no local optima. We do not report estimates of b, which showed an order of magnitude of difference across clades, confirming that phylogenetic data are not useful for estimating this parameter.

Interestingly, we find parameter estimates for $b - d$ and v that are rather consistent across mammalian groups (Table 1). The mutation parameter is constrained to a narrow range (from 0.16 to 0.39 Myr^{-1}) for 11 of 14 phylogenies; only three outliers (Calomys, Microtus and Macaca) have higher values (up to 1.72 for Calomys). The range is slightly broader for the growth rate (from 0.05 to 0.53 Myr^{-1}), with only one outlier (1.84 for Microtus). In agreement with results presented above, we find that the estimated growth rate is of the same order as the mutation rate (slightly higher for exactly half of the phylogenies and slightly lower for the other half). Hence, the growth rate is far from being null, suggesting that the metacommunity is not at equilibrium.

DISCUSSION

We developed a neutral, out-of-equilibrium model of biodiversity that produces realistic phylogenetic trees. We developed a fast simulation algorithm for this model, as well as a method of inference that allows fitting the model to empirical data efficiently. We illustrated our method using 14 mammalian phylogenies. Our results corroborate the hypothesis that phylogenies are better explained by out-of-equilibrium models of biodiversity.

Our model can be seen as an extension to the NTB. The first main difference lies in the speciation process. Similar to the point mutation model, speciation arises as a result of mutation events. However, contrary to the original point mutation model, a single mutation is typically not enough to induce speciation. The SGD process leads to the split of an ancestral species' population into two daughter species. In this respect, it can be seen as providing a mutational basis to the random fission mode of speciation. The SGD model also offers a good microscopic basis to the hypothesis of 'protracted speciation' by which there is a time lag between the initiation of population divergence and the time when gene flow completely stops and distinct species are recognised (Coyne & Orr 2004; Rosindell *et al.* 2010; Etienne & Rosindell 2011). Indeed, divergence starts in SGD with a mutation event, but a new species is formed and recognised only if (and after) enough mutations have accumulated.

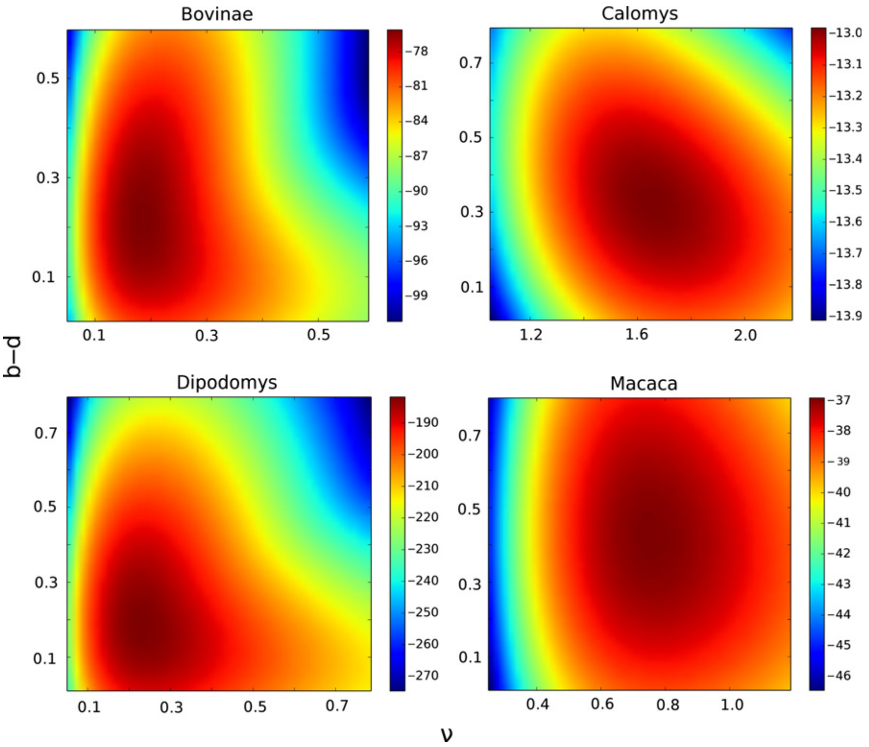

Fig. 4. Growth rates and mutation rates under SGD estimated for four mammalian phylogenies. Colors correspond to likelihood values. The likelihood landscapes have a single peak, demonstrating the ability to infer the parameters of SGD from phylogenies.

Table 1. Parameters of the SGD model inferred for various clades of mammals

Clade	*f*	Clade age (Myr)	*n*	*β*	*γ*	*b* − *d*	*v*	*b* − *d* − *v*
Bovinae	1	19.58	25	−1.28	−1.2	0.19	0.16	0.03
Calomys	0.85	2.99	13	−1.77	0.58	0.45	1.72	−1.27
Caprinae	0.89	9.94	38	−1.06	−1.38	0.40	0.39	0.01
Dasyuridae	0.92	29.5	72	−0.52	−5.22	0.20	0.23	−0.03
Dipodomys	0.95	20.66	65	−0.33	−2.41	0.05	0.30	−0.25
Duikers	0.83	17.16	58	−1.03	−0.79	0.38	0.30	0.08
Viverrinae	0.88	14.92	25	−1.55	0.64	0.36	0.31	0.05
Hylobatidae	1	8.81	14	−1.57	−1.47	0.53	0.32	0.21
Alouatta	0.91	16.46	14	−0.7	−0.42	0.20	0.39	−0.19
Macaca	0.95	5.8	22	0.9	−1.45	0.42	0.76	−0.34
Microtus	0.69	4.08	154	−0.37	−5.33	1.84	1.05	0.79
Mustelidae	0.85	22.53	59	−1.28	−1.65	0.38	0.22	0.16
Ochotona	0.92	13.69	39	−0.5	−0.89	0.25	0.35	−0.10
Talpa	0.77	26.85	35	−0.99	−1.58	0.16	0.21	−0.05

f, sampling fraction; clade age, crown age; *n*, number of extant species in the clade (not all species are sampled); the inferred parameters $b - d$ and v are expressed in Myr^{-1}.

The second main difference between our model and the classical NTB is that we relax the constant metacommunity size hypothesis. We find that when the growth rate $b - d$ decreases, meaning that the metacommunity is close to equilibrium, phylogenies become unrealistic in terms of branching times. This confirms results from the classical equilibrium NTB model showing that realistic branching times are hardly ever obtained (Davies *et al*. 2011).

Here, non-equilibrium refers to a growing metacommunity size, as opposed to a constant (equilibrium) metacommunity size. Data on metacommunity sizes over evolutionary time scales are generally missing. However, a growing metacommunity size seems more realistic than a constant one at time scales spanning the history of entire clades. Indeed, the number of individuals in radiating clades has to have increased lastingly during diversification, during expansion phases corresponding, for example, to the colonisation of new territories. Still, the exponential growth model considered here is simplistic with regard to the complex history of clades. More complex scenarios with time-inhomogeneous rates could be analysed with our framework. We are hopeful that such developments could allow us to detect broad trends in the way metacommunity size varies over long time scales.

Our results refer to equilibrium in terms of metacommunity size, not diversity. Still, it is more likely that diversity reaches equilibrium when metacommunity size reaches equilibrium than when metacommunity size is expanding. This confirms earlier results stemming from lineage-based models that have found a better support for non-equilibrium models compared to stationary ones (Hey 1992; Morlon *et al.* 2010), and suggest that non-equilibrium models should be preferred over equilibrium models to predict the loss of evolutionary history due to species loss (Nee & May 1997), as well as the way phylogenetic diversity scales spatially (Morlon *et al.* 2011b).

Applying our model to mammalian phylogenies, we found rather consistent parameter estimates across groups. As expected, demographic parameter estimates ($b - d$) were more heterogeneous than mutational parameter estimates (v). In our results, the two clades showing the highest values of inferred v were Microtus and Calomys. These clades also had high values of $b - d$ compared to other clades, such as Bovinae, Dipodomys and Talpa. Microtus and Calomys are small rodent species, indeed known to have a high reproduction rate (Golley *et al.* 1975), and small generation times potentially leading to high mutation rates. We do not see any other obvious history traits that could explain differences across groups in terms of growth and mutation rates. It would be interesting to compare our growth estimates to effective population size curves obtained from genetic data, although such data for entire clades are not yet available. We could also look at these growth estimates in light of the age and current global population sizes of clades. Estimates for $b - d$ may seem low (in the order of one event per Myr) in comparison with the usual instantaneous growth rate of population dynamics. At the time scales considered here, the growth rate $b - d$ reflects the long-term growth of the entire metacommunity, that is, an average trend rather than the fast oscillating dynamics of populations. Similarly, estimates for the mutation rate v may seem low in comparison with the usual genomic mutation rate, and high in comparison with estimates of the point-mutation speciation rate in NTB (Condit *et al.* 2002). However, mutation rate in the SGD model refers to mutations that have an effect on speciation. They represent only a small fraction of the mutations arising on a DNA sequence, leading to values much lower than genomic mutation rate. As a high number of these mutations do not directly lead to speciation, SGD's mutation rate is indeed expected to be larger than NTB's mutation rate. Finally, the mutation rate we infer is orders of magnitude lower than reasonable birth rates, which shows that the mutations playing a role in speciation in SGD arise on a much slower timescale than those in population dynamics. Typical values of b (e.g., 10^5 Myr^{-1}) yield a ratio of birth to mutation in the order of a population size, which is in line with the traditional assumption in population genetics.

Our study provides an alternative to previous interpretations of patterns observed in empirical trees, in terms of both branching times and balance. Negative γ values have traditionally been interpreted as the effect of adaptive diversification (Phillimore & Price 2008; Rabosky & Lovette 2008), biogeographical processes (Pigot *et al.* 2010) or protracted speciation (Rosindell *et al.* 2010; Etienne & Rosindell 2011) (see Moen & Morlon 2014 for a review). Here, we show that a non-adaptive, non-spatial model can explain the branching times of real trees. In our model, stemmy phylogenies arise both from the chosen mode of speciation, which naturally accounts for protractedness, and as a result of an expanding metacommunity. Phylogenetic imbalance suggests that some groups of organisms are more species rich than others. This variation in species richness across taxonomic groups has traditionally been interpreted as evidence that non-neutral, ecological differences among lineages drive differences in speciation and extinction rates (Alfaro *et al.* 2009). In agreement with previous studies (Jabot & Chave 2009; Pigot *et al.* 2010; Davies *et al.* 2011), our analyses demonstrate that the levels of phylogenetic imbalance observed in nature can arise from purely neutral processes. In the NTB with point mutation, phylogenetic imbalance arises as a by-product of stochastic differences in population sizes (per-lineage speciation rate is a linear function of abundance). Similarly, in Pigot's biogeographic model (2010), which is also neutral, phylogenetic imbalance arises from stochastically driven differences in range sizes (species with wider ranges are more likely to experience vicariance events). In our model, however, the link between speciation rates and abundance is not as straightforward. On the one hand, abundant species 'see' more mutations, which could promote speciation. On the other hand, the ancestral type survives longer in rapidly expanding populations, such that speciation may become more difficult. Another explanation for imbalance in phylogenetic trees is differences in diversification linked to a heritable trait (Heard 1996). In general, this has been interpreted as different abilities for species with different ecological characteristics to speciate and/or go extinct. Here, the model is neutral, meaning individuals across species all have the same birth, death and mutation rates. However, the process of speciation generates differences across species in terms of the interconnection of individuals through potential hybridisation. Some species are big hubs (those where the ancestral type has not disappeared) that do not easily speciate, while others (those where the ancestral type has disappeared) speciate more easily. This hidden trait is heritable, generating imbalance without invoking ecological differences between species.

While our study provides an alternative to previous interpretations of patterns observed in empirical trees, assessing the goodness of fit of our models compared to other models is not yet possible. Such comparisons cannot currently be performed, as we are lacking a robust approach for fitting models – such as the adaptive (McPeek 2008) and the biogeographic (Pigot *et al.* 2010) models – to phylogenetic trees. However, we could use the framework presented here to compare the fit of models with constant birth and death rates, leading to an exponentially growing metacommunity, to that of models with time-varying growth rates. We could consider models with population-level density dependence that could lead to clade-wide diversity dependence (Phillimore & Price 2008; Rabosky & Lovette 2008; Etienne *et al.* 2011). This would provide a (non-adaptive) diversity-dependent model certainly worth exploring. We could also consider models in which the metacommunity net growth rate switches from positive (expanding metacommunity) to negative (shrinking metacommunity) along history, which could result in periods of diversity expansion followed by diversity decline

(Morlon *et al.* 2011a; Quental & Marshall 2013; Morlon 2014). It would also be particularly interesting to consider a spatial version of the model accounting for dispersal limitation (MacArthur & Wilson 1967; Etienne & Alonso 2005; Jabot & Chave 2009; Rosindell & Phillimore 2011), which would allow us to fit the model to a much broader array of data sets at the community scale.

An interesting aspect of our model is that it not only produces predictions for macroevolutionary patterns (phylogenies) but also for macroecological patterns (species abundance distributions). We have not yet fully explored the shape of species abundance distributions arising from SGD, but preliminary results suggest that the model can produce shapes covering the classical log-series and log-normal shapes depending on the choice of the parameters.

CONCLUSION

Our study is one of the first attempts at proving analytical solutions for phylogenies arising from an individual-based model. Further work in this direction will be clearly needed for a better integration of macroevolution into macroecology and community ecology. Importantly, we showed that considering out-of-equilibrium models will be crucial to this integration. In macroevolution, out-of-equilibrium models are the norm, but they had not been previously linked to non-equilibrium metacommunity sizes. Our framework provides perspectives for better understanding how diversity dynamics relate to metacommunity dynamics. In macroecology and community ecology, our results call for a major shift from our current focus on steady-state predictions to a focus on transient dynamics.

ACKNOWLEDGEMENTS

We thank Alexandre Pigot for sharing his data. We also thank Franck Jabot, Eric Lewitus, as well as former and current members of the BioDiv team at the ENS for discussions and comments on the manuscript. HM acknowledges the CNRS and grants ECOEVOBIO-CHEX2011 from the French National Research Agency (ANR) and PANDA from the European Research Council (ERC). AL thanks the Center for Interdisciplinary Research in Biology (Collège de France) for funding.

STATEMENT OF AUTHORSHIP

MM, AL & HM designed research, MM, AL & HM performed research, MM & AL contributed analytical tools, MM analysed the data and MM, AL & HM wrote the manuscript.

REFERENCES

Alfaro, M.E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L. *et al.* (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl. Acad. Sci. USA*, 106, 13410–13414.

Avise, J.C. & Ball, R. (1990). Principles of genealogical concordance in species concepts and biological taxonomy. *Oxford Surv. Evol. Bio.*, 7, 45–67.

Blum, M.G.B. & François, O. (2006). Which random processes describe the tree of life? A large scale study of phylogenetic tree imbalance. *Syst. Biol.*, 55, 685–691.

Bortolussi, N., Durand, E., Blum, M. & François, O. (2006). apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*. Oxford University Press (OUP): Policy B - Oxford Open Option B, 22(3), 363–364.

Champagnat, N. & Lambert, A. (2013). Splitting trees with neutral Poissonian mutations II: largest and oldest families. *Stoch. Proc. Appl.*, 123, 1368–1414.

Chave, J. & Leigh, E. (2002). A spatially explicit neutral model of beta-diversity in tropical forests. *Theor. Popul. Biol.*, 62, 153–168.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B. *et al.* (2002). Beta-diversity in tropical forest trees. *Science*, 295, 666–669.

Coyne, J. & Orr, H. (2004). *Speciation*. vol. 37. Sinauer Associates, Sunderland, MA.

Davies, T.J., Allen, A.P., Borda-de Água, L., Regetz, J. & Melián, C.J. (2011). Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. *Evolution*, 65, 1841–1850.

Etienne, R.S. & Alonso, D. (2005). A dispersal-limited sampling theory for species and alleles. *Ecol. Lett.*, 8, 1147–1156.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. *et al.* (2011). Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc. Biol. Sci.*, 279, 1300–1309.

Etienne, R.S., Morlon, H. & Lambert, A. (2014). Estimating the duration of speciation from phylogenies. *Evolution*, 68, 2430–2440.

Etienne, R.S. & Rosindell, J. (2011). Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Syst. Biol.*, 61, 204–213.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17, 368–376.

Golley, F.B., Petrusewicz, K. & Ryszkowski, L. (1975). *Small Mammals: Their Productivity and Population Dynamics*. vol. 5. Cambridge University Press, Cambridge.

Graham, C.H. & Fine, P.V.A. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecol. Lett.*, 11, 1265–1277.

Heard, S.B. (1996). Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, 50, 2141–2148.

Hey, J. (1992). Using phylogenetic trees to study speciation and extinction. *Evolution*, 46, 627–640.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.

Jabot, F. & Chave, J. (2009). Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecol. Lett.*, 12, 239–248.

Kendall, D.G. (1948). On the generalized "birth-and-death" process. *Ann. Math. Stat.*, 19, 1–15.

Lambert, A., Morlon, H. & Etienne, R.S. (2014). The reconstructed tree in the lineage-based model of protracted speciation. *J. Math. Biol.*, 70, 367–397.

Lambert, A. & Stadler, T. (2013). Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. *Theor. Popul. Biol.*, 90, 113–128.

Lambert, A. & Steel, M. (2013). Predicting the loss of phylogenetic diversity under non-stationary diversification models. *J. Theor. Biol.*, 337, 111–124.

MacArthur, R.H. & Wilson, E.O. (1967). *The Theory of Island Biogeography*. Princeton University Press, Princeton.

McGill, B.J., Maurer, B.A. & Weiser, M.D. (2006). Empirical evaluation of neutral theory. *Ecology*, 87, 1411–1423.

McKinney, M.L. (1997). Extinction vulnerability and selectivity: combining ecological and paleontological views. *Annu. Rev. Ecol. Syst.*, 28, 495–516.

McPeek, M.A. (2008). The ecological dynamics of clade diversification and community assembly. *Am. Nat.*, 172, 270–284.

Moen, D. & Morlon, H. (2014). Why does diversification slow down? *Trends Ecol. Evol.*, 29, 190–197.

Mooers, A.O., Harmon, L.J., Blum, M.G.B., Wong, D.H.J. & Heard, S.B. (2007). Some models of phylogenetic tree shape. In: *Reconstructing Evolution: New Mathematical and Computational Advances* (eds. Gascuel, O. & Steel, M.). Oxford University Press, Oxford, pp. 149–170.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecol. Lett.*, 17, 508–525.

Morlon, H., Condamine, F. & Manceau, M. (2014). RPANDA: Phylogenetic ANalyses of Diversification in R. R package version 1.0.

Morlon, H., Parsons, T.L. & Plotkin, J.B. (2011a). Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. USA*, 108, 16327–16332.

Morlon, H., Potts, M.D. & Plotkin, J.B. (2010). Inferring the dynamics of diversification: a coalescent approach. *PLoS Biol.*, 8, 1–13.

Morlon, H., Schwilk, D.W., Bryant, J.A., Marquet, P.A., Rebelo, A.G., Tauss, C. *et al.* (2011b). Spatial patterns of phylogenetic diversity. *Ecol. Lett.*, 14, 141–149.

Nee, S., Harvey, P. & Mooers, A. (1992). Tempo and mode of evolution revealed from molecular phylogenies. *Proc. Natl. Acad. Sci. USA*, 89, 8322–8326.

Nee, S. & May, R.M. (1997). Extinction and the loss of evolutionary history. *Science*, 278, 692–694.

O'Dwyer, J.P. & Green, J.L. (2010). Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity. *Ecol. Lett.*, 13, 87–95.

Phillimore, A.B. & Price, T.D. (2008). Density-dependent cladogenesis in birds. *PLoS Biol.*, 6, 483–489.

Pigot, A.L., Owens, I.P.F. & Orme, C.D.L. (2012). Speciation and extinction drive the appearance of directional range size evolution in phylogenies and the fossil record. *PLoS Biol.*, 10, 1–9.

Pigot, A.L., Phillimore, A.B., Owens, I.P.F. & Orme, C.D.L. (2010). The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Syst. Biol.*, 59, 660–673.

Pybus, O.G. & Harvey, P.H. (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc. Biol. Sci.*, 267, 2267–2272.

Quental, T.B. & Marshall, C.R. (2013). How the Red Queen drives terrestrial mammals to extinction. *Science*, 341, 290–292.

Rabosky, D.L. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE*, 9, 1–15.

Rabosky, D.L. & Lovette, I.J. (2008). Density-dependent diversification in North American wood warblers. *Proc. Biol. Sci.*, 275, 2363–2371.

Rosenzweig, M.L. (1995). *Species Diversity in Space and Time*. Cambridge university press, Cambridge.

Rosindell, J., Cornell, S.J., Hubbell, S.P. & Etienne, R.S. (2010). Protracted speciation revitalizes the neutral theory of biodiversity. *Ecol. Lett.*, 13, 716–727.

Rosindell, J. & Phillimore, A.B. (2011). A unified model of island biogeography sheds light on the zone of radiation. *Ecol. Lett.*, 14, 552–560.

Stadler, T. (2011). Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl. Acad. Sci. USA*, 108, 6187–6192.

Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002). Phylogenies and community ecology. *Annu. Rev. Ecol. Syst.*, 33, 475–505.

Wiens, J.J., Ackerly, D.D., Allen, A.P., Anacker, B.L., Buckley, L.B., Cornell, H.V. *et al.* (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecol. Lett.*, 13, 1310–1324.

SUPPORTING INFORMATION

Additional Supporting Information may be downloaded via the online version of this article at Wiley Online Library (www.ecologyletters.com).

Editor, Arne Mooers
Manuscript received 10 September 2014
First decision made 22 October 2014
Second decision made 16 December 2014
Manuscript accepted 6 January 2015

Supplementary Material

-

Phylogenies support out-of-equilibrium models of biodiversity

Marc Manceau, Amaury Lambert, Hélène Morlon

Contents

A Effects of parameter values on the shape of the phylogeny

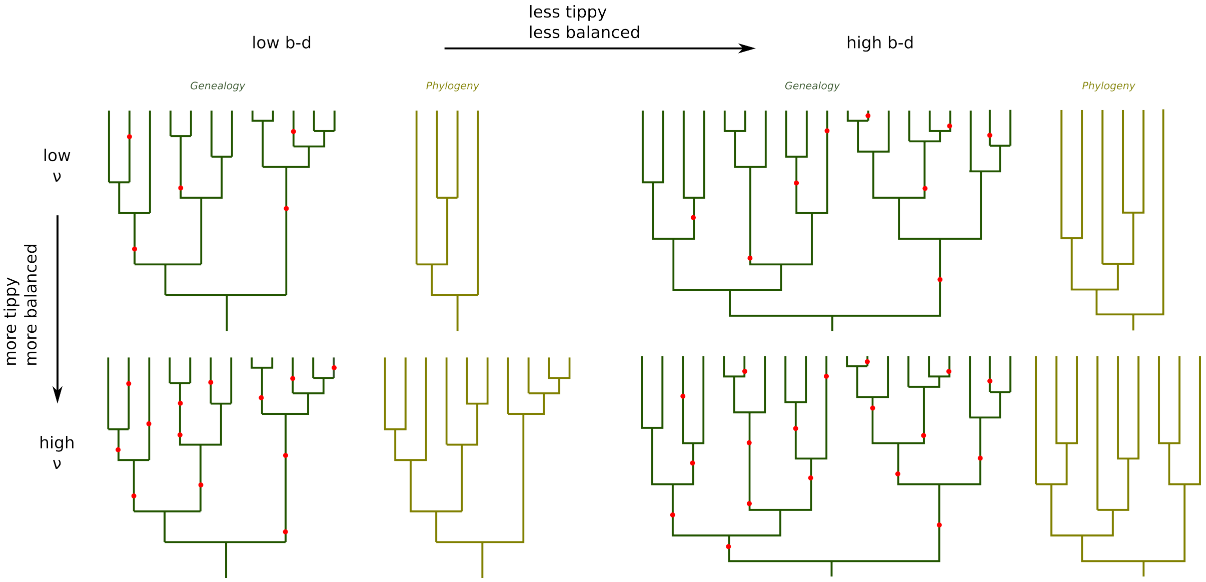

Figure S.1: Effect of the parameter values $b - d$ and ν on the shape of phylogenies.

Decreasing the growth rate at constant mutation rate (from the right column to the left column in Figure S.1) has the same effect as increasing the mutation rate at constant growth rate (from the top to the bottom row). We detail here the later effect. By our definition of speciation, all nodes from the phylogenies are nodes from the genealogy.

Note first that deep nodes from the genealogy tend to be phylogenetic.

Increasing the mutation rate at constant growth rate increases the number of mutations on the genealogy, and thus the number of phylogenetic nodes. As more nodes from the genealogy are conserved on the phylogeny, nodes that are close to the tips are increasingly conserved and phylogenies become more tippy.

The genealogies are generated by a constant birth-death process and their expected balance is thus $\beta = 0$ by definition of β. When there are few mutations, they tend to fall on long lines in the genealogy according to the Poisson process, such that nodes from stemmy subtrees (with short lines) tend to not be phylogenetic. This creates imbalance in the resulting phylogeny. When there are more mutations, short lines are also hit by mutations, and phylogenies become more balanced.

B Derivation of g(t) and m(t)

B.1 Survival probability of a population up to a time t

We denote $u_{b,d}(t)$ the extinction probability before time t of a population originally composed of one individual, following a birth-death process with inhomogeneous birth rate $b(t)$ and death rate $d(t)$. This probability is derived in Kendall (1948) :

$$u_{b,d}(t) = \frac{1 + \int_0^t b(s) e^{\int_0^s b(z)-d(z)dz} ds - e^{\int_0^t b(z)-d(z)dz}}{1 + \int_0^t b(s) e^{\int_0^s b(z)-d(z)dz} ds} \tag{1}$$

B.2 Branching rate $g(t)$ on the reconstructed genealogy

We consider the genealogy of individuals (Figure 1A) given by the linear birth-death model. We introduce the following notation to describe what happens at a birth time :

$M_t = \{$At least one descendant from an ancestral individual giving birth at time t is still alive at present.$\}$

$L_t = \{$The left descent from an ancestral individual giving birth at time t is still alive at present.$\}$

$R_t = \{$The right descent from an ancestral individual giving birth at time t is still alive at present.$\}$

A branching event on the reconstructed genealogy (Fig 1B) corresponds to a birth event that leads to two descents that are both alive at present. The instantaneous rate of such events at time t (conditioned on non extinction) is given by :

$$\begin{aligned} g(t)dt &= P(\text{ birth} \in dt,\ L_t,\ R_t |\ M_t) \\ &= b(t)\frac{P(L_t, R_t)}{P(M_t)}dt \\ &= b(t)\frac{(1-u_{b,d}(t))^2}{1-u_{b,d}(t)}dt \\ &= b(t)(1-u_{b,d}(t))dt \\ &= \frac{b(t)e^{\int_0^t b(z)-d(z)dz}}{1+\int_0^t b(s)e^{\int_0^s b(z)-d(z)dz}ds} \end{aligned} \tag{2}$$

B.3 Survival probability of a clonal population

We call clonal descent from an ancestral individual living at time t the whole descent from this individual in which no mutation occurred (see Figure 1A). We introduce the following notation :

$M_t^C = \{$The clonal descent from an ancestral individual living at time t is still alive at present$\}$

$m(t) = P(M_t^C \mid M_t)$

We get

$$m(t) = P(M_t^C \mid M_t) = \frac{P(M_t^C \cap M_t)}{P(M_t)} = \frac{P(M_t^C)}{P(M_t)}$$

Remember that the dynamics of the whole population is a birth-death process, with birth rate $b(t)$ and death rate $d(t)$, and the dynamics of the clonal population is a birth-death process, with birth rate $b(t)$ and death rate $d(t) + \nu(t)$. This gives us, $\forall t \in [0, T]$:

$$\begin{aligned} m(t) &= \frac{1 - u_{b,d+\nu}(t)}{1 - u_{b,d}(t)} \\ &= \frac{e^{\int_0^t b(z)-d(z)-\nu(z)dz}}{1 + \int_0^t b(s) e^{\int_0^s b(z)-d(z)-\nu(z)dz} ds} \frac{1 + \int_0^t b(s) e^{\int_0^s b(z)-d(z)dz}}{e^{\int_0^t b(z)-d(z)dz}} \end{aligned} \tag{3}$$

C Forward-in-time phylogeny simulation

C.1 A three-type branching process

We need to define three types of lineages in order to simulate the process at the lineage-level (see Figures 1C and S.2 for an illustration) :

a “type 0” lineage is a line from the underlying genealogy that has at least one descendant of same genetic type at present.

a “type 1” lineage is a line from the underlying genealogy that has no descendant of same genetic type at present.

a type 0 lineage "freezes" at the first node (in forward time of the underlying phylogeny) that is not divergent. In other words, it freezes at the first node that has at least two descendants at present time, one in each of the two incident descents, having the same genetic type. In this case, the whole descent of this node is collapsed into a single species, and the lineage is “frozen”, in the sense that no further splitting or extinction event can happen to this lineage up to the present.

The phylogenetic tree, considered as a time-inhomogeneous branching process with three types, is simulated forward-in-time. Figure S.2 illustrates the definition of types.

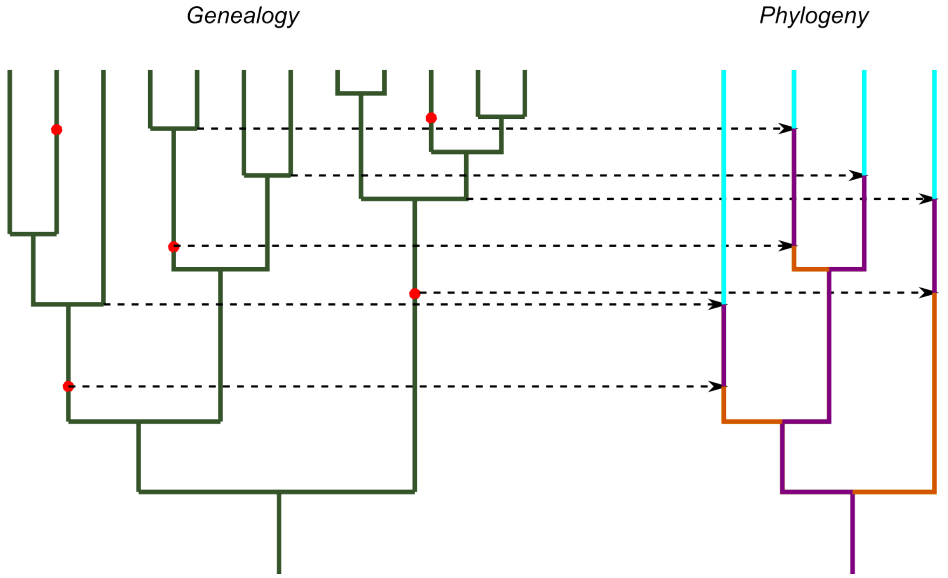

Figure S.2: Left, genealogy is in green, and red dots are mutation events. Right, the corresponding phylogeny is in purple for type 0 lineages, orange for type 1 lineages and blue for frozen lineages.

C.2 Transition rates

C.2.1 Possible events

The following events can occur on a type 1 lineage between time t and $t+dt$:

- There is a mutation on $[t,t+dt]$ and the lineage is changed into a type 0 lineage.
- There is a mutation on $[t,t+dt]$ but the lineage remains of type 1.
- A branching occurs on $[t,t+dt]$ and gives rise to two type 1 lineages.
- Nothing happens on the genealogy between $[t,t+dt]$, nor in the phylogeny.

The following events can occur on a type 0 lineage between time t and $t+dt$:

- A branching occurs on $[t,t+dt]$ and the clonal type disappears in one population. It gives rise to one type 0 lineage, and one type 1 lineage.
- A branching occurs on $[t,t+dt]$ and the clonal type survives in both populations. The lineage is frozen.
- There is no birth on $[t,t+dt]$, the lineage remains of type 0.

We represent on figure S.3 all events happening in our three-type process.

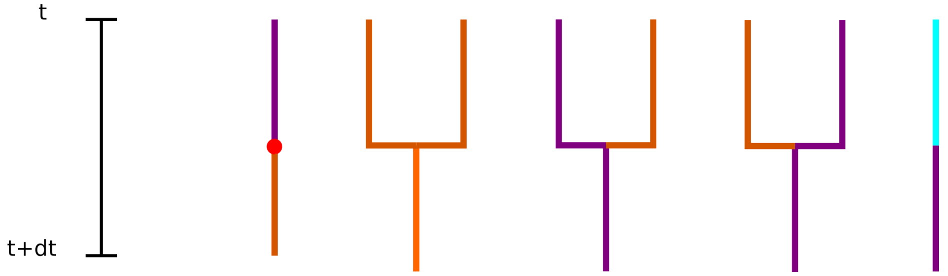

Figure S.3: Type 0 lineages are in purple, type 1 lineages are in orange, and frozen lineages are in blue. The simulation goes "forward" from $t+dt$ to t, up to the present.

C.2.2 Derivation of the rates

Recall that M_t^C denotes the survival of a clonal descent from an ancestral individual living at time t up to time 0. Let $\overline{M_t^C}$ denote the extinction before present of the clonal descent from an ancestral individual at time t.

A lineage of type 1 becomes of type 0 between t and $t+dt$ when a mutation occurs in this time interval and the clonal descent from the ancestral individual carrying the mutation does not get extinct before present. This happens with rate :

$$\begin{aligned}\rho_{1\to 0}(t)dt &= P(\text{ mutation} \in dt,\ M_t^C |\ M_{t+dt},\ \overline{M_{t+dt}^C}\ )\\ &= \frac{\nu(t)(1-u(t))m(t)}{(1-u(t))(1-m(t))}dt\\ &= \frac{\nu(t)m(t)}{(1-m(t))}dt\end{aligned}$$

A lineage of type 1 branches and gives rise to two lineages of type 1 when there is a birth event, survival of the two descents and extinction of the two clonal descents. This happens with rate :

$$\begin{aligned}\rho_{1\to+1}(t)dt &= P(\text{ birth} \in dt,\ L_t,\ R_t,\ \overline{L_t^C},\ \overline{R_t^C} |\ M_{t+dt}, \overline{M_{t+dt}^C}\ )\\ &= b(t)\frac{(1-u(t))^2(1-m(t))^2}{(1-u(t))(1-m(t))}dt\\ &= g(t)(1-m(t))dt\end{aligned}$$

A lineage of type 0 branches and gives rise to one lineage of type 0 and one lineage of type 1 when there is a birth event, survival of the two descents, and extinction of the clonal descent in one of the two descents. This happens with rate :

$$\begin{aligned}\rho_{0\to+1}(t)dt &= P(\text{ birth} \in dt,\ L_t,\ R_t,\ (L_t^C, \overline{R_t^C})\ \cup\ (\overline{L_t^C}, R_t^C) |\ M_{t+dt}^C\ )\\ &= b(t)\frac{2(1-u(t))^2 m(t)(1-m(t))}{(1-u(t))m(t)}dt\\ &= 2g(t)(1-m(t))dt\end{aligned}$$

A lineage of type 0 “freezes”, giving rise to a tip lineage in the phylogeny, when there is a birth event and survival of the two clonal descents :

$$\begin{aligned}\rho_{0\to\varnothing}(t)dt &= P(\text{ birth} \in dt,\ L_t^C,\ R_t^C |\ M_{t+dt}^C\ )\\ &= b(t)\frac{(1-u(t))^2 m(t)^2}{(1-u(t))m(t)}dt\\ &= g(t)m(t)dt\end{aligned}$$

D Likelihood of a tree

We aim to compute the likelihood of a tree arising from the SGD process from which each tip lineage has been sampled with probability f. Below, the "type of the tree" refers to the type of tree before the sampling procedure.

We define, for a given phylogenetic tree A, its likelihood under this model, up to time t, to be :

$$\begin{aligned}\mathcal{L}^i_{A,f}(t) = P(&\text{ a tree that started at one individual with stem age } t \text{ has shape } A \text{ and type } i \\ &\mid \text{survival up to time } 0, \text{the model and the parameter set } (b,d,\nu))\end{aligned}$$

We will also need the following additional notation :

$$\begin{aligned}w^i_f(t) = P(&\text{ a tree that started at one individual with stem age } t \text{ has type } i \text{ but no species sampled} \\ &\mid \text{survival up to time } t, \text{the model and the parameter set } (b,d,\nu))\end{aligned}$$

We study here how these probabilities change as t increases, and new subtrees appear. We have to take into account all events happening to the genealogy and leading to the observed phylogeny.

We slice the problem into four pieces :

- ODEs driving (w^0_f, w^1_f).
- Likelihood on a tip lineage : ODEs driving $(\mathcal{L}^0_{T,f}, \mathcal{L}^1_{T,f})$, where T stands for "Tip".
- Likelihood on an internal lineage : ODE driving $(\mathcal{L}^0_{I,f}, \mathcal{L}^1_{I,f})$, where I stands for "Internal".
- Likelihood at a node (branching time).

To ease notation, we will drop the dependence of all quantities upon t, including :

$$\nu = \nu(t) \ , \quad g = g(t) \ , \quad b = b(t) \ , \quad u = u(t)$$

We need to introduce some additional notation to describe different events. In the following, the type of a line (in the genealogy) or of a lineage (in the phylogeny) at a given time will be :

0 if its clonal descent has survived to the present.

1 if it has extant descent but no extant clonal descent at the present.

e if it has no extant descent at the present.

Additionally, a (phylogenetic) lineage (but not a genealogical line, because sampling concerns species and not individuals) can have two "sampling states" :

u unsampled (none of its descending tip species is sampled at present)

s sampled (at least one of its descending tip species is sampled at present)

We write the type of a line or lineage as a superscript, and the sampling state (for lineages only) as a subscript.

We will also need the following event names :

$$S := \{ \text{ survival of the genealogical process up to time 0 } \}$$
$$\emptyset := \{ \text{ nothing happens in } [t, t+dt]\}$$
$$M := \{ \text{ a mutation event happens on } [t, t+dt]\}$$
$$L_j^i := \{ \text{ branching event in } [t, t+dt], \text{ the left tree has type i and sampling state j}\}$$
$$R_j^i := \{ \text{ branching event in } [t, t+dt], \text{ the right tree has type i and sampling state j}\}$$
$$F_j := \{ \text{ branching event in } [t, t+dt], \text{ both incident subtrees have type 0, frozen lineage has sampling state j } \}$$

Finally, we show on figures S.4, S.5, S.6, and S.7 , the different events that could happen. Type 0 trees or lineages are in green, type 1 are in purple, extinct are in black. Dotted lines stand for trees with no species sampled, whereas solid lines stand for trees with at least one species sampled.

D.1 ODEs driving w_f^i

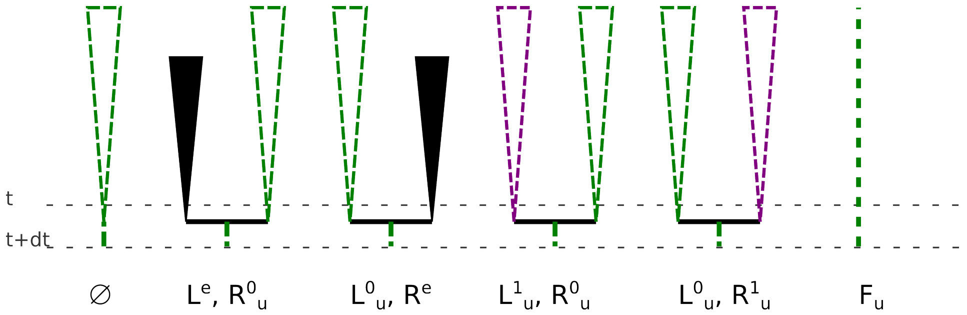

Figure S.4: All events leading to a tree of type 0 with no species sampled at $t + dt$.

We know the initial condition for $w_f^0(0) = 1 - f$, i.e., the probability of not sampling a given species. We will now derive ODEs driving w_f^0 as t increases, with corresponding events shown in figure S.4.

$$\begin{aligned}
w_f^0(t+dt) &= w_f^0(t)P(\ \emptyset \mid S\ ) \\
&+ P(\ (R_u^0 \cap L^e) \cup (R^e \cap L_u^0) \cup (R_u^1 \cap L_u^0) \cup (R_u^0 \cap L_u^1) \cup F_u \mid S\ ) \\
&= w_f^0(t)\left[1 - \nu dt - b\frac{1-u^2}{1-u}dt + 2budt + 2b(1-u)w_f^1 dt\right] + b(1-u)m^2(1-f)dt
\end{aligned}$$

This leads to the following differential equation driving w_f^0 :

$$\begin{aligned}
\frac{dw_f^0}{dt} &= w_f^0\left[-\nu - b(1+u) + 2bu + 2b(1-u)w_f^1\right] + b(1-u)m^2(1-f) \\
&= w_f^0\left[-\nu - g + 2gw_f^1\right] + gm^2(1-f) \qquad (4)
\end{aligned}$$

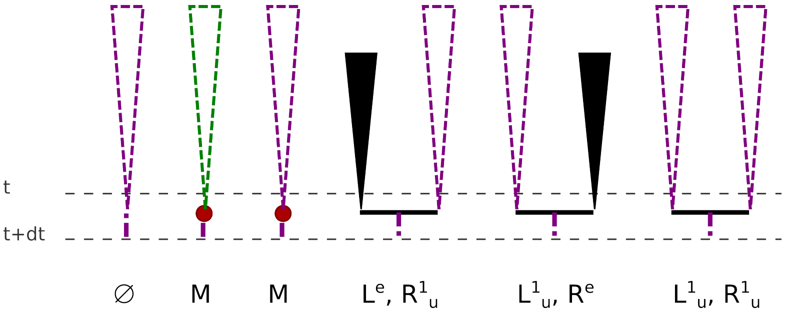

Figure S.5: All events leading to a type 1 tree unsampled at $t + dt$.

Recall that the initial condition is $w_f^1(0) = 0$, because a short tip lineage has vanishing probability of carrying a mutation. We will now derive ODEs driving w_f^1 as t increases, with corresponding events shown in figure S.5 :

$$\begin{aligned}
w_f^1(t+dt) &= w_f^1(t)P(\ \emptyset \mid S\ ) + (w_f^1(t) + w_f^0(t))P(\ M \mid S\ ) \\
&+ P(\ (R_u^1 \cap L^e) \cup (R^e \cap L_u^1) \cup (R_u^1 \cap L_u^1) \mid S\ ) \\
&= w_f^1(t)\left[1 - \nu dt - b\frac{1-u^2}{1-u}dt + 2budt + b(1-u)w_f^1 dt + \nu dt\right] + w_f^0(t)\nu dt
\end{aligned}$$

This leads to the following differential equation driving w_f^1 :

$$\begin{aligned} \frac{dw_f^1}{dt} &= w_f^1(t)\left[-b(1+u)+2bu+b(1-u)w_f^1\right]+\nu w_f^0 \\ &= -gw_f^1(1-w_f^1)+\nu w_f^0 \end{aligned} \tag{5}$$

Note that for $f = 1$, the whole tree is sampled, and we verify that $\forall t \geq 0, \quad w_f^0(t) = w_f^1(t) = 0$.

D.2 Likelihood of a tip lineage

D.2.1 Likelihood of a type 0 tip lineage

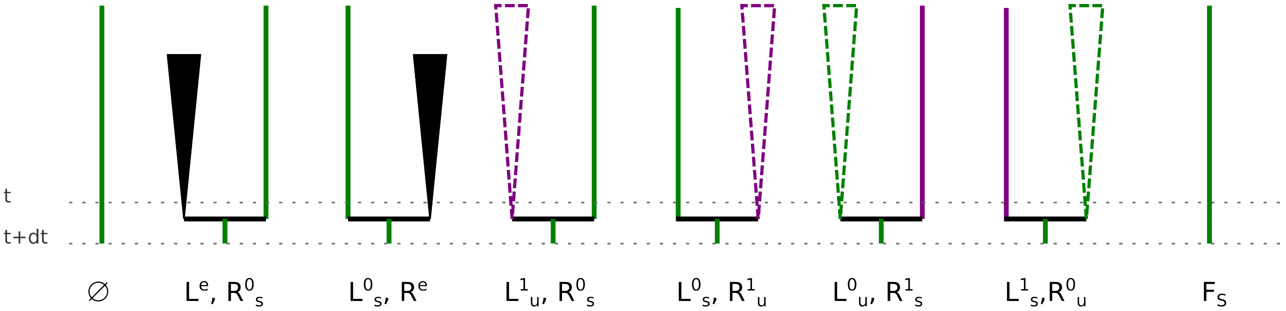

Figure S.6: All events leading to a tip lineage of type 0 at $t + dt$.

The initial condition is $\mathcal{L}^0_{T,f}(0) = f$, i.e., the probability of sampling a given species. We will now derive ODEs driving $\mathcal{L}^0_{T,f}$ as t increases, with corresponding events shown in figure S.6 :

$$\begin{aligned} \mathcal{L}^0_{T,f}(t+dt) &= \mathcal{L}^0_{T,f}(t)P(\ \emptyset \mid S\ ) \\ &+ P(\ (R^0_s \cap L^e) \cup (R^e \cap L^0_s) \cup (R^0_s \cap L^1_u) \cup (R^1_u \cap L^0_s) \cup (R^0_u \cap L^1_s) \cup (R^1_s \cap L^0_u) \cup F_s \mid S\ ) \\ &= \mathcal{L}^0_{T,f}(t)\left[1-\nu dt - b\frac{1-u^2}{1-u}dt + 2budt + 2b(1-u)w_f^1 dt\right] \\ &+ \mathcal{L}^1_{T,f}(t)2b(1-u)w_f^0 dt + b(1-u)m^2 f dt \end{aligned}$$

This leads to the following differential equation driving $\mathcal{L}^0_{T,f}$:

$$\begin{aligned} \frac{d\mathcal{L}^0_{T,f}}{dt} &= \mathcal{L}^0_{T,f}\left[-\nu - b(1+u) + 2bu + 2b(1-u)w_f^1\right] + \mathcal{L}^1_{T,f}2b(1-u)w_f^0 + b(1-u)m^2 f \\ &= \mathcal{L}^0_{T,f}\left[-\nu - g + 2gw_f^1\right] + \mathcal{L}^1_{T,f}2gw_f^0 + gm^2 f \end{aligned} \tag{6}$$

Solving the equation in the particular case $f = 1$, we derive the likelihood expression :

$$\mathcal{L}^0_{T,1}(t) = e^{-\int_0^t g(z)+\nu(z)dz} + \int_0^t g(s)m(s)^2 e^{-\int_s^t g(z)+\nu(z)dz} ds$$

D.2.2 Likelihood of a type 1 tip lineage

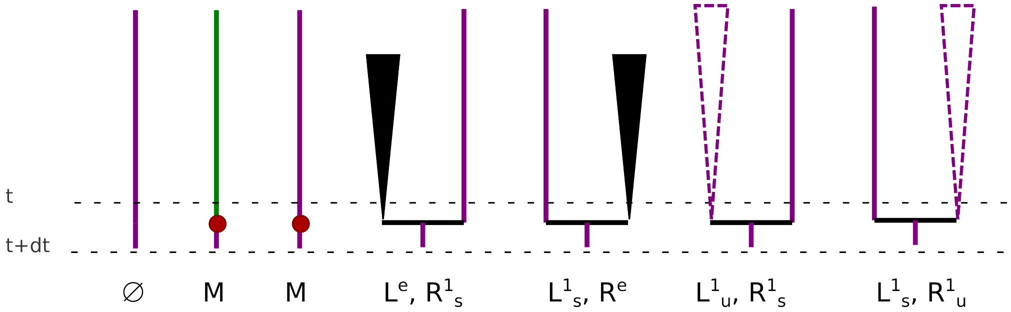

Figure S.7: All events leading to a type 1 tip lineage at $t + dt$.

The initial condition is $\mathcal{L}^1_{T,f}(0) = 0$, because no mutation can occur exactly at time 0. We will now derive ODEs driving $\mathcal{L}^1_{T,f}$ as t increases, with corresponding events shown in figure S.7 :

$$\begin{aligned}
\mathcal{L}^1_{T,f}(t+dt) &= \mathcal{L}^1_{T,f}(t)P(\ \emptyset \mid S\ ) + (\mathcal{L}^1_{T,f}(t) + \mathcal{L}^0_{T,f}(t))P(\ M \mid S\ ) \\
&+ P(\ (R^1_s \cap L^e) \cup (R^e \cap L^1_s) \cup (R^1_s \cap L^1_u) \cup (R^1_u \cap L^1_s) \mid S\ ) \\
&= \mathcal{L}^1_{T,f}(t)\left[1 - \nu dt - b\frac{1-u^2}{1-u}dt + 2budt + 2b(1-u)w^1_f dt + \nu dt\right] + \mathcal{L}^0_{T,f}(t)\nu dt
\end{aligned}$$

This leads to the following differential equation driving $\mathcal{L}^1_{T,f}$:

$$\begin{aligned}
\frac{d\mathcal{L}^1_{T,f}}{dt} &= \mathcal{L}^1_{T,f}(t)\left[-b(1+u) + 2bu + 2b(1-u)w^1_f\right] + \nu\mathcal{L}^0_{T,f} \\
&= g\mathcal{L}^1_{T,f}(2w^1_f - 1) + \nu\mathcal{L}^0_{T,f} \qquad (7)
\end{aligned}$$

Solving the equation in the particular case $f = 1$, we derive the likelihood expression :

$$\mathcal{L}^1_{T,1}(t) = e^{-\int_0^t g(z)dz}\left(1 - e^{-\int_0^t \nu(z)dz}\right) + \int_0^t g(s)m(s)^2 e^{-\int_s^t g(z)dz}\left(1 - e^{-\int_s^t \nu(z)dz}\right) ds$$

D.3 Likelihood on internal lineages

We call internal lineages all segments of the phylogenies between two nodes. Similarly as for tip lineages, we get :

$$\frac{d\mathcal{L}^0_{I,f}}{dt} = \mathcal{L}^0_{I,f}\left[-\nu - g + 2gw^1_f\right] + 2\mathcal{L}^1_{I,f}gw^0_f \tag{8}$$

And :

$$\frac{d\mathcal{L}^1_{I,f}}{dt} = g\mathcal{L}^1_{I,f}(2w^1_f - 1) + \nu\mathcal{L}^0_{I,f} \tag{9}$$

Solving the equation in the particular case $f = 1$, between two nodes with depths $t_1 \leq t_2$, we find the likelihood expressions :

$$\begin{aligned}
\mathcal{L}^0_{I,1}(t_2) &= \mathcal{L}^0_{I,1}(t_1)e^{-\int_{t_1}^{t_2}(g(z)+\nu(z))dz} \\
\mathcal{L}^1_{I,1}(t_2) &= \mathcal{L}^0_{I,1}(t_1)e^{-\int_{t_1}^{t_2}g(z)dz}\left(1 - e^{-\int_{t_1}^{t_2}\nu(z)dz}\right) + \mathcal{L}^1_{I,1}(t_1)e^{-\int_{t_1}^{t_2}g(z)dz}
\end{aligned}$$

D.4 Likelihood at a branching point

We compute here the likelihood of a tree N at a branching point between two subtrees A and B. Figure S.8 illustrates different situations leading to a phylogenetic tree N composed of two subtrees A and B.

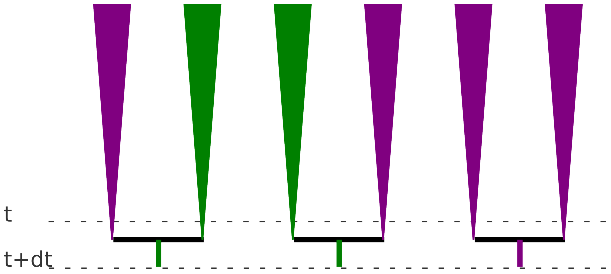

Figure S.8: All events at a node point leading to either a type 0 or a type 1 tree.

A type 0 tree is obtained if there is a branching event at time t, and either (see figure S.8) :

- (A is of type 0) and (B is of type 1) or
- (A is of type 1) and (B is of type 0)

Hence it follows :

$$\mathcal{L}^0_{N,f}(t) \quad = \quad g\left(\mathcal{L}^0_{A,f}\mathcal{L}^1_{B,f} + \mathcal{L}^1_{A,f}\mathcal{L}^0_{B,f}\right) \tag{10}$$

A type 1 tree is obtained if there is a branching event at time t, (A is of type 1) and (B is of type 1). Hence it follows :

$$\mathcal{L}^1_{N,f} \quad = \quad g\mathcal{L}^1_{A,f}\mathcal{L}^1_{B,f} \tag{11}$$

D.5 Peeling algorithm implementation

We compute the likelihood of a tree, recursively computing the likelihoods of subtrees from the root to the tips, using expressions (10-11) at node points and expressions (4-5) and (8-9) for internal lineages, and finally using expressions (4-5) and (6-7) for the tip lineages of the phylogeny.

Note that it is necessary to compute both likelihoods of type 0 and type 1 trees at the same time as they are coupled in the differential equations. The resulting likelihood of the tree X is $\mathcal{L}_{X,f} = \mathcal{L}^0_{X,f} + \mathcal{L}^1_{X,f}$.

The algorithm is written in Python and R, and is available from the authors upon request.